

Friedrich Recknagel  
William K. Michener *Editors*

# Ecological Informatics

Data Management and Knowledge  
Discovery

*Third Edition*

 Springer

# Ecological Informatics

Friedrich Recknagel • William K. Michener  
Editors

# Ecological Informatics

Data Management and Knowledge Discovery

Third Edition

 Springer

*Editors*

Friedrich Recknagel  
School of Biological Sciences  
University of Adelaide  
Adelaide, SA  
Australia

William K. Michener  
College of University Libraries  
University of New Mexico  
Albuquerque, New Mexico  
USA

ISBN 978-3-319-59926-7

ISBN 978-3-319-59928-1 (eBook)

DOI 10.1007/978-3-319-59928-1

Library of Congress Control Number: 2017950734

© Springer International Publishing AG 2003, 2006, 2018

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Printed on acid-free paper

This Springer imprint is published by Springer Nature

The registered company is Springer International Publishing AG

The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland



# Contents

## Part I Introduction

- 1 Ecological Informatics: An Introduction . . . . .** 3  
Friedrich Recknagel and William K. Michener

## Part II Managing Ecological Data

- 2 Project Data Management Planning . . . . .** 13  
William K. Michener
- 3 Scientific Databases for Environmental Research . . . . .** 27  
John H. Porter
- 4 Quality Assurance and Quality Control (QA/QC) . . . . .** 55  
William K. Michener
- 5 Creating and Managing Metadata . . . . .** 71  
William K. Michener
- 6 Preserve: Protecting Data for Long-Term Use . . . . .** 89  
Robert B. Cook, Yaxing Wei, Leslie A. Hook,  
Suresh K.S. Vannan, and John J. McNelis
- 7 Data Discovery . . . . .** 115  
William K. Michener
- 8 Data Integration: Principles and Practice . . . . .** 129  
Mark Schildhauer

### **Part III Analysis, Synthesis and Forecasting of Ecological Data**

- 9 Inferential Modelling of Population Dynamics** . . . . . 161  
Friedrich Recknagel, Dragi Kocev, Hongqing Cao,  
Christina Castelo Branco, Ricardo Minoti, and Saso Dzeroski
- 10 Process-Based Modeling of Nutrient Cycles and Food-Web  
Dynamics** . . . . . 189  
George Arhonditsis, Friedrich Recknagel, and Klaus Joehnk
- 11 Uncertainty Analysis by Bayesian Inference** . . . . . 215  
George Arhonditsis, Dong-Kyun Kim, Noreen Kelly, Alex Neumann,  
and Aisha Javed
- 12 Multivariate Data Analysis by Means of Self-Organizing Maps** . . . . . 251  
Young-Seuk Park, Tae-Soo Chon, Mi-Jung Bae, Dong-Hwan Kim,  
and Sovan Lek
- 13 GIS-Based Data Synthesis and Visualization** . . . . . 273  
Duccio Rocchini, Carol X. Garzon-Lopez, A. Marcia Barbosa,  
Luca Delucchi, Jonathan E. Olandi, Matteo Marcantonio,  
Lucy Bastin, and Martin Wegmann

### **Part IV Communicating and Informing Decisions**

- 14 Communicating and Disseminating Research Findings** . . . . . 289  
Amber E. Budden and William K. Michener
- 15 Operational Forecasting in Ecology by Inferential Models  
and Remote Sensing** . . . . . 319  
Friedrich Recknagel, Philip Orr, Annelie Swanepoel, Klaus Joehnk,  
and Janet Anstee
- 16 Strategic Forecasting in Ecology by Inferential  
and Process-Based Models** . . . . . 341  
Friedrich Recknagel, George Arhonditsis, Dong-Kyun Kim,  
and Hong Hanh Nguyen

### **Part V Case Studies**

- 17 Biodiversity Informatics** . . . . . 375  
Cynthia S. Parr and Anne E. Thessen
- 18 Lessons from Bioinvasion of Lake Champlain, U.S.A.** . . . . . 401  
Timothy B. Mihuc and Friedrich Recknagel
- 19 The Global Lake Ecological Observatory Network** . . . . . 415  
Paul C. Hanson, Kathleen C. Weathers, Hilary A. Dugan,  
and Corinna Gries

**20 Long-Term Ecological Research in the Nakdong River:  
Application of Ecological Informatics to Harmful Algal Blooms . . . 435**  
Dong-Gyun Hong, Kwang-Seuk Jeong, Dong-Kyun Kim,  
and Gea-Jae Joo

**21 From Ecological Informatics to the Generation of Ecological  
Knowledge: Long-Term Research in the English Lake District . . . 455**  
S.C. Maberly, D. Ciar, J.A. Elliott, I.D. Jones, C.S. Reynolds,  
S.J. Thackeray, and I.J. Winfield

**Part I**  
**Introduction**

# Chapter 1

## Ecological Informatics: An Introduction

Friedrich Recknagel and William K. Michener

### 1.1 Introduction

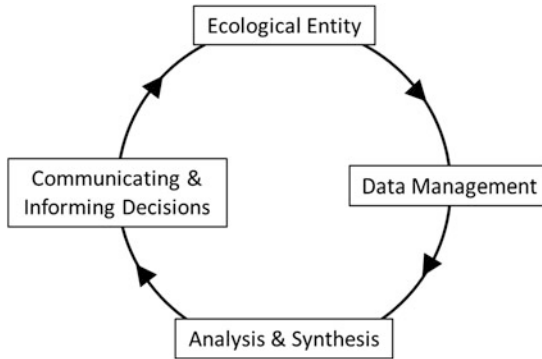
Ecological Informatics is an emerging discipline that takes into account the data-intensive nature of ecology, the valuable information content of ecological data, and the need to communicate results and inform decisions, including those related to research, conservation and resource management (Recknagel 2017). At its core, ecological informatics combines developments in information technology and ecological theory with applications that facilitate ecological research and the dissemination of results to scientists and the public. Its conceptual framework links ecological entities (genomes, organisms, populations, communities, ecosystems, landscapes) with data management, analysis and synthesis, and communicating and informing decisions by following the course of a loop (Fig. 1.1).

*Ecological Entities* range from genomes, individual organisms, populations, communities, ecosystems to landscapes and the biosphere, and are highly complex and distinctly evolving. Figure 1.2 illustrates the evolving nature of ecosystems in view of the fact that physical-chemical boundaries such as topology, temperature, pH, and substrate determine their community of organisms. Progressing shifts of physical-chemical boundaries under the influence of environmental and climate changes at seasonal and inter-annual scales restructure communities of organisms, and ecosystems adjust in due course. Over time, evolving ecosystems also alter the nature of landscapes. Ecologists are challenged by the evolving nature and

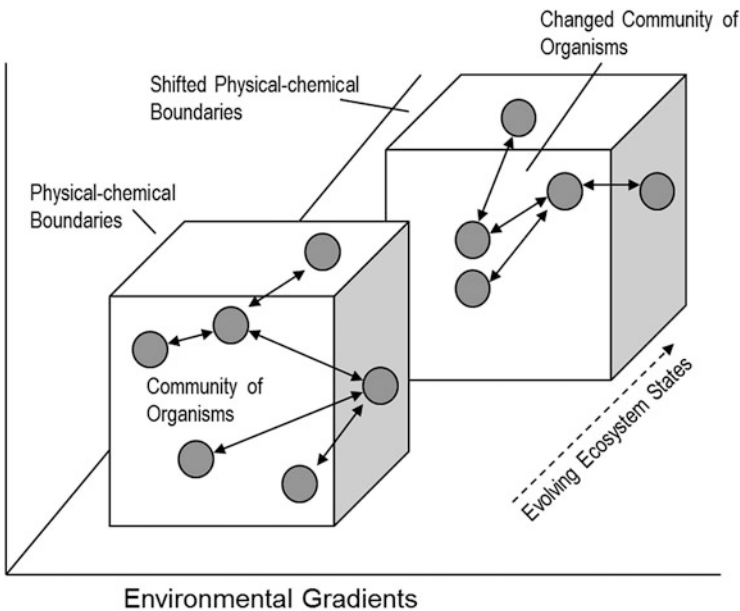
---

F. Recknagel (✉)  
University of Adelaide, Adelaide, SA, Australia  
e-mail: [friedrich.recknagel@adelaide.edu.au](mailto:friedrich.recknagel@adelaide.edu.au)

W.K. Michener  
The University of New Mexico, Albuquerque, NM, USA  
e-mail: [william.michener@gmail.com](mailto:william.michener@gmail.com)



**Fig. 1.1** Conceptual framework of ecological informatics



**Fig. 1.2** Evolving nature of ecosystems

data-intensive nature of ecology, and require suitable concepts and tools to deal appropriately with these challenges.

*Data Management* must meet requirements of many diverse sources of information, and be suitable to a wide range of spatial and temporal scales. Sources of information include paleo-ecological, eco-genomic, habitat, community and climate data. Spatial scales of ecological data range from habitat-specific to global, and time scales range from real-time to centuries-long.

*Analysis and Synthesis* utilise archived and real-time information for inventorying ecological entities, assessing sustainability of habitats and biodiversity, and



hind- and forecasting of ecological entities. Multivariate statistics are commonly applied for data analysis. Data synthesis typically applies inferential and process-based modelling techniques, and utilises remote sensing and GIS-based tools. Bayesian inference extends the predictive capacity of inferential and process-based models by quantifying model uncertainties and estimating forecasting risks.

*Communicating and Informing Decisions* supported by data analysis and synthesis is relevant for generating hypotheses for subsequent research steps as well as for identifying viable management options. While inferential models help inform short-term decisions, process-based models are more appropriate for long-term forecasts and decision-making.

## 1.2 Data Management

Ecological data management is a process that starts at the conceptualization of the project and concludes after the data have been archived and the results have informed future research as well as resource management, conservation, and other types of decision-making. Data management may be conceptualized in terms of a data life cycle (Fig. 1.3) whereby: (1) projects are conceived and data collection and analyses are planned; (2) data are collected and organized, usually into data tables (e.g., spreadsheets) or databases; (3) data are quality assured using accepted quality assurance/quality control (QA/QC) techniques; (4) data are documented through the creation of metadata that describe all aspects of the data and research; (5) data are preserved in a data repository or archive so that they may be reused and shared; (6) data are discovered or made discoverable so that they may be used in synthesis efforts or to reproduce results of a study; (7) data are integrated

**Fig. 1.3** The life cycle of data. Note the steps need not be sequential nor does research necessarily involve all steps; e.g., some synthesis efforts may involve no new data collection, thereby proceeding from data discovery through integration with other data, to analysis and visualization



with other data in order to answer specific questions such as examining the influence of climate extremes on pollination ecology; and (8) data are explored, analysed and visualized, leading to new understanding that can then be communicated to other scientists and the public.

The seven chapters in Part II discuss concepts, practices and tools that are commonly used in data management planning through data integration. In Chap. 2, Michener (2017a) provides guidance on developing effective data management plans. Chapter 3 (Porter 2017) describes different database approaches that can be used to organize and manage data, as well as key data management concepts like data modelling and data normalization. Chapter 4 (Michener 2017b) focuses on commonly used graphical and statistical QA/QC approaches to ensuring data quality. In Chap. 5, Michener (2017c) discusses the metadata standards and tools that can be used to document data so it can be easily discovered, accessed and interpreted. Cook et al. (2017) describe best practices for protecting and preserving data to support long-term acquisition and use in Chap. 6. Chapter 7 (Michener 2017d) focuses on methods that can be employed to more easily discover data as well as make data more readily discoverable by others. In Chap. 8, Schildhauer (2017) discusses the underlying principles and practices involved in integrating data from different sources—a necessary prerequisite for most data analysis and synthesis efforts.

### 1.3 Analysis and Synthesis

The five chapters in Part III discuss a subset of modern tools that can be used for analysis, synthesis and forecasting. Figure 1.4 provides an overview of basic steps and methods of data analysis and synthesis in ecology. *Conceptual Models* should be the starting point by reflecting research questions and key variables in an instructive way. Sources for *Data Acquisition* typically include field, laboratory and/or literature data. Common methods for *Data Analysis* are canonical correspondence analysis (CCA), principal component analysis (PCA) as well as self-organising maps (SOM) that reduce the data dimension and reveal nonlinear relationships by ordination and clustering of multivariate data.

In Chap. 12, Park et al. (2017) address explicitly the benefits of SOM for revealing and visualising nonlinear relationships in complex ecological data, and in Chap. 18, Mihuc and Recknagel (2017) demonstrate applications of canonical correspondence analysis for qualitative analysis of interrelationships between the native zooplankton community and invasive zebra mussel and alewife in Lake Champlain.

*Data Synthesis* can be performed by statistical, inferential and process-based modelling techniques. Statistical modelling basically utilises univariate nonlinear and multivariate linear regression analysis but fail to identify multivariate nonlinear relationships intrinsic of ecological data. By contrast, inferential models using artificial neural networks (ANN) and evolutionary algorithms (EA) are well suited

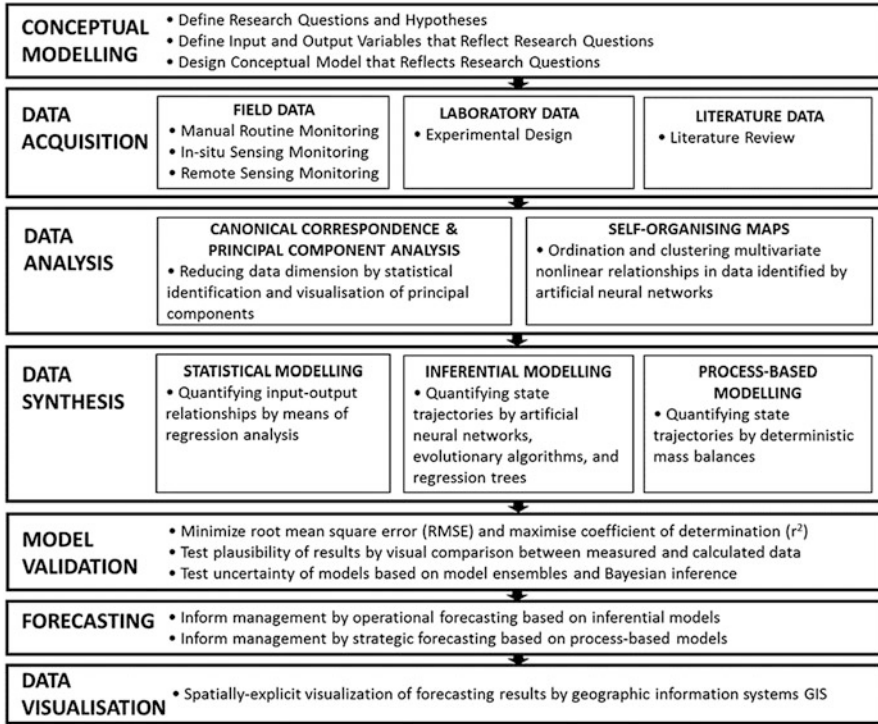


Fig. 1.4 Basic steps and methods of data analysis and synthesis

as tools to encapsulate and predict the highly complex and interrelated behaviour of ecological entities solely based on inductive reasoning. ANN do not explicitly represent models and this is viewed as a major shortcoming of this computational technique. By contrast EA represent models explicitly by IF-THEN-ELSE rules. In Chap. 9, Recknagel et al. (2017a) introduce the rationale of the hybrid evolutionary algorithm (HEA) and demonstrate applications of HEA for threshold identification, predictive modelling and meta-analysis. Inferential modelling by HEA proves also suitable for operational forecasting and early warning as discussed in Chap. 15 by Recknagel et al. (2017b). Inferential models by regression trees represent correlations between habitat properties and ecological entities by hierarchical structured IF-THEN-ELSE rules. Case studies in Chap. 15 demonstrate their capability to identify threshold conditions responsible for changing ecological entities.

Process-based models as outlined in Chap. 10 by Arhonditsis et al. (2017a) synthesize data by nonlinear differential equations that contain algebraic equations of Michaelis-Menten-type kinetics, causal and empirical relations. As demonstrated by case studies in Chap. 10, process-based modelling of specific ecosystems requires substantial data sets as well as *ad hoc* parameter optimization and calibration. If simulation results achieve reasonable validity for a specific ecosystem as indicated by a ‘low’ root mean squared error RMSE and a ‘high’ coefficient of

determination  $r^2$ , the underlying model may be applied for hypotheses testing or long-term forecasting by scenario analysis. The credibility of scenario analyses may be constrained by the scope and inherent uncertainties of models. The analysis of model uncertainty by means of Bayesian inference is explicitly addressed in Chap. 11 by Arhonditsis et al. (2017b) and demonstrated by several case studies.

*Forecasting* of ecosystem behaviour is prerequisite for preventing or mitigating events that cause rapid deterioration of ecological entities. In Chap. 16, Recknagel et al. (2017c) address forecasting by model ensembles in order to overcome single model constraints. Case studies in Chap. 16 demonstrate that ensembles of complementary models extend the scope of an individual model, which is necessary to more realistically reveal complex interrelationships between adjacent ecosystems such as catchments and lakes under the influence of global change, and that model-specific uncertainties may be compromised by Bayesian analysis of ensembles of alternative models (see also Chap. 11). As shown in Chap. 15 (Recknagel et al. 2017b), predictive inferential models and remote sensing appear capable of short-term forecasting of rapid outbreaks of population density. Two case studies demonstrate that inferential models based on HEA allow early warning of harmful algal blooms in lakes by real-time forecasts up to 30-day-ahead. The chapter also discusses the potential of remote sensing for real-time monitoring of the spatio-temporal distribution of water quality parameters and cyanobacteria blooms in water bodies. *Data Visualisation* is prerequisite to successfully communicate and disseminate findings from data analysis and synthesis. In Chap. 13, Rocchini et al. (2017) address the potential of GIS-tools to visualise spatially-explicit modelling and forecasting results.

## 1.4 Communicating and Informing Decisions

Research findings must be accessible to technical and general audiences to inform decision-making, contribute to new knowledge, and educate about complex topics. Part IV includes three chapters that illustrate how information can best be conveyed to diverse audiences. In Chap. 14, Budden and Michener (2017) discuss best practices for communicating and disseminating research outputs via publications, presentations, illustrations and social media. Various modelling approaches can be particularly useful for informing near-term and long-term decisions. In Chap. 15, Recknagel et al. (2017b) highlight the potential for inferential models and remote sensing to inform operational decisions by short-term forecasting. In Chap. 16, Recknagel et al. (2017c) present scenario analysis by complementary and alternative model ensembles that can inform strategic decision-making by long-term forecasting.

## 1.5 Case Studies

The five specific case studies included in Part V illustrate how ecological informatics has evolved to meet the needs of the various disciplines that comprise the domain of ecological science. In Chap. 17, Parr and Thessen (2017) present two user stories that highlight the latest tools and procedures that are used to manage biodiversity data, including identification tools, phylogenetic trees, ontologies, controlled vocabularies, standards, and genomics. In Chap. 18, Mihuc and Recknagel (2017) demonstrate applications of CCA and HEA to long-term limnological data of Lake Champlain (USA). In Chap. 19, Hanson et al. (2017) provide an overview of the Global Lake Ecological Observatory Network and emphasize the role of coordinated social and technical change in a successful research network. Chapter 20 (Hong et al. 2017) describes efforts to analyse and synthesize data resulting from the Nakdong River (South Korea) Long Term Ecological Research effort. Maberly et al. (2017) report research outcomes from the LTER English Lake District in Chap. 21.

## References

- Arhonditsis G, Recknagel F, Joehnk K (2017a) Process-based modeling of nutrient cycles and food-web dynamics, Chapter 10. In: Recknagel F, Michener W (eds) Ecological informatics. Data management and knowledge discovery. Springer, Heidelberg
- Arhonditsis G, Kim D-Y, Kelly N, Neumann A, Javed A (2017b) Uncertainty analysis by Bayesian inference, Chapter 11. In: Recknagel F, Michener W (eds) Ecological informatics. Data management and knowledge discovery. Springer, Heidelberg
- Budden AE, Michener WK (2017) Communicating and disseminating research findings, Chapter 14. In: Recknagel F, Michener W (eds) Ecological informatics. Data management and knowledge discovery. Springer, Heidelberg
- Cook RB, Wei Y, Hook LA, Vannan SKS, McNelis JJ (2017) Preserve: protecting data for long-term use, Chapter 6. In: Recknagel F, Michener W (eds) Ecological informatics. Data management and knowledge discovery. Springer, Heidelberg
- Hanson PC, Weathers KC, Dugan HA, Gries C (2017) The global lake ecological observatory network, Chapter 19. In: Recknagel F, Michener W (eds) Ecological informatics. Data management and knowledge discovery. Springer, Heidelberg
- Hong D-G, Jeong K-S, Kim D-K, Joo G-J (2017) Long-term ecological research in the Nakdong River: application of ecological informatics to harmful algal blooms, Chapter 20. In: Recknagel F, Michener W (eds) Ecological informatics. Data management and knowledge discovery. Springer, Heidelberg
- Maberly SC, Ciar D, Elliott JA, Jones ID, Reynolds CS, Thackeray SJ, Winfield IJ (2017) Long-term ecological research in the English Lake District: from ecological informatics to the generation of ecological knowledge, Chapter 21. In: Recknagel F, Michener W (eds) Ecological informatics. Data management and knowledge discovery. Springer, Heidelberg
- Michener WK (2017a) Project data management planning, Chapter 2. In: Recknagel F, Michener W (eds) Ecological informatics. Data management and knowledge discovery. Springer, Heidelberg

- Michener WK (2017b) Quality assurance and quality control (QA/QC), Chapter 4. In: Recknagel F, Michener W (eds) Ecological informatics. Data management and knowledge discovery. Springer, Heidelberg
- Michener WK (2017c) Creating and managing metadata, Chapter 5. In: Recknagel F, Michener W (eds) Ecological informatics. Data management and knowledge discovery. Springer, Heidelberg
- Michener WK (2017d) Data discovery, Chapter 7. In: Recknagel F, Michener W (eds) Ecological informatics. Data management and knowledge discovery. Springer, Heidelberg
- Mihuc TB, Recknagel F (2017) Lessons from bioinvasion of Lake Champlain, U.S.A., Chapter 18. In: Recknagel F, Michener W (eds) Ecological informatics. Data management and knowledge discovery. Springer, Heidelberg
- Park Y-S, Chon T-S, Bae M-J, Kim D-H, Lek S (2017) Multivariate data analysis by means of self-organizing maps, Chapter 12. In: Recknagel F, Michener W (eds) Ecological informatics. Data management and knowledge discovery. Springer, Heidelberg
- Parr CS, Thessen AE (2017) Biodiversity informatics, Chapter 17. In: Recknagel F, Michener W (eds) Ecological informatics. Data management and knowledge discovery. Springer, Heidelberg
- Porter JH (2017) Scientific databases for environmental research, Chapter 3. In: Recknagel F, Michener W (eds) Ecological informatics. Data management and knowledge discovery. Springer, Heidelberg
- Recknagel F (2017) Ecological informatics. In: Gibson D (ed) Oxford bibliographies in ecology. Oxford University Press, New York. <http://www.oxfordbibliographies.com/view/document/obo-9780199830060/obo-9780199830060-0174.xml>
- Recknagel F, Kocev D, Cao H, Branco CC, Minoti R, Dzeroski S (2017a) Inferential modelling of population dynamics, Chapter 9. In: Recknagel F, Michener W (eds) Ecological informatics. Data management and knowledge discovery. Springer, Heidelberg
- Recknagel F, Orr P, Swanepoel A, Joehnk K, Anstee J (2017b) Operational forecasting in ecology by inferential models and remote sensing, Chapter 15. In: Recknagel F, Michener W (eds) Ecological informatics. Data management and knowledge discovery. Springer, Heidelberg
- Recknagel F, Arhonditsis G, Kim D-K, Nguyen HH (2017c) Strategic forecasting in ecology by inferential and process-based models, Chapter 16. In: Recknagel F, Michener W (eds) Ecological informatics. Data management and knowledge discovery. Springer, Heidelberg
- Rocchini D, Garzon-Lopez CX, Barbosa AM, Delucchi L, Olandi JE, Marcantonio M, Bastin L, Wegmann M (2017) GIS-based data synthesis and visualization, Chapter 13. In: Recknagel F, Michener W (eds) Ecological informatics. Data management and knowledge discovery. Springer, Heidelberg
- Schildhauer M (2017) Data integration: principles and practice, Chapter 8. In: Recknagel F, Michener W (eds) Ecological informatics. Data management and knowledge discovery. Springer, Heidelberg



**Part II**  
**Managing Ecological Data**

# Chapter 2

## Project Data Management Planning

William K. Michener

**Abstract** A data management plan (DMP) describes how you will manage data during a research project and what you will do with the data after the project ends. Research sponsors may have very specific requirements for what should be included in a DMP. In lieu of or in addition to those requirements, good plans address 11 key issues: (1) research context (e.g., what questions or hypotheses will be examined); (2) how the data will be collected and acquired (e.g., human observation, *in situ* or remote sensing, surveys); (3) how the data will be organized (e.g., spreadsheets, databases); (4) quality assurance and quality control procedures; (5) how the data will be documented; (6) how the data will be stored, backed up and preserved for the long-term; (7) how the data will be integrated, analyzed, modeled and visualized; (8) policies that affect data use and redistribution; (9) how data will be communicated and disseminated; (10) roles and responsibilities of project personnel; and (11) adequacy of budget allocations to implement the DMP. Several tips are offered in preparing and using the DMP. In particular, researchers should start early in the project development process to create the DMP, seek input from others, engage all relevant project personnel, use common and widely available tools, and adopt community practices and standards. The best DMPs are those that are referred to frequently, reviewed and revised on a routine basis, and recycled for use in subsequent projects.

### 2.1 Introduction

A data management plan (DMP) describes how you will manage data throughout the life of a research project and what you will do with the data after the project ends. Many research sponsors now require that a DMP be submitted as part of a grant proposal. The plan is included in the package that is reviewed to determine whether the proposal is worthy of funding. Once a project commences, some sponsors regularly review a project's data management activities against what was included in the DMP.

---

W.K. Michener (✉)  
University of New Mexico, Albuquerque, NM, USA  
e-mail: [william.michener@gmail.com](mailto:william.michener@gmail.com)

It is good practice to prepare a DMP before a project is initiated, regardless of whether or not the research sponsor requires it. The process of creating a DMP causes one to think through many issues that will affect the costs, personnel needs, and feasibility of a project such as:

1. How much data will be collected and how will it be treated?
2. How much time is needed to manage the data and who will be responsible for doing so?
3. How long should the data be preserved and where is the best location to do so?
4. Are there any legal constraints associated with acquiring, using and sharing project data?

Understanding these issues upfront can save significant time, money, and aggravation over the long-term. For example, a project's data management activities may reasonably be expected to cost more in terms of personnel and equipment when many terabytes of data are collected as opposed to just a few megabytes of data. Similarly, extra precautions and security are normally required when human subject data are collected. In effect, a good DMP helps position your research project for success.

The remainder of this chapter offers guidance on what is needed to create a good plan as well as some tools and tips that can be employed. First, the components of a DMP are presented along with relevant examples and links to additional resources. Next, the overall process of developing and using a DMP is described. This section includes suggestions on who creates the plan, when it gets created, and how the plan is devised and used.

## 2.2 Components of a Data Management Plan

Research sponsors often have very specific requirements or guidance for the types of information to be included in a DMP. These requirements are usually listed in the request for proposals (or funding opportunity announcement) or in the sponsor's grant proposal guide. It is a good idea to consult these documents, which are normally located on the sponsor's web site. You may also identify requirements by checking the Data Management Planning Tool website (DMPTool 2016) or the DMPonline website (Digital Curation Center 2016) for US and UK research sponsors, respectively. The websites are useful resources that provide funding agency requirements for data management plans in the form of templates with annotated advice for filling in the template. The DMPTool website also includes numerous example plans that are published by DMPTool users. Many universities and other organizations support Research Data Librarians that are knowledgeable about sponsor requirements and can provide assistance in developing DMPs. As a last resort, don't hesitate to contact the relevant program officials with any questions about DMP requirements.

Regardless of the specifics, DMP requirements typically apply to all or portions of the data life cycle—e.g., data collection and organization, quality assurance and quality control, documentation (i.e., metadata), data storage and preservation, data analysis and visualization, and sharing with others (e.g., data policies and dissemination approaches). In addition, it is usually a good idea to identify the roles and responsibilities of all project participants that are engaged in data management activities, and to include a budget that covers relevant personnel, hardware, software, and services. Note that research sponsors may place page limits on the DMP (e.g., two pages). Nevertheless, a DMP should be a useful resource for your project. DMPs that exceed page limits can easily be shortened into a summary that meets sponsor requirements. The various components of a comprehensive DMP are described in the remainder of this section.

### ***2.2.1 Context***

A brief summary of the project context can be quite instructive for those involved directly in the project as well as others that may wish to use the data after they have been shared. A good summary indicates:

- Why the data are being collected (e.g., questions or hypotheses that are being addressed)
- Who will create and use the data (e.g., names and roles of project participants and collaborators)
- How the data will be used (e.g., intended uses of the data, potential limitations on data use)
- How the project is being supported (e.g., sponsors, supporting organizations such as field stations and marine laboratories)

Such information may later be expanded upon and incorporated into the metadata (see Sect. 2.5 and Michener 2017b).

### ***2.2.2 Data Collection and Acquisition***

All components of a DMP depend upon knowing sources, types and volumes of data that will be collected as part of the project. It is useful to document who is responsible for acquiring and processing the data as well as where the data are acquired. Data sources may include remote sensing platforms (e.g., aerial, satellite, balloon, drone), *in situ* environmental sensor networks (Porter et al. 2009, 2012), environmental observatories and research networks [e.g., Long-Term Ecological Research Network (Michener and Waide 2009; Michener et al. 2011), National Ecological Observatory Network (Schimel et al. 2011), Ocean Observatories Initiative (Consortium for Ocean Leadership 2010), and others (see Peters et al.

2014)], data centers and repositories (Sect. 2.6; Cook et al. 2017), surveys and interviews, and human observation in the field. Other data may be acquired by laboratory instruments or derived from models or computer simulations. It is important to note whether the acquired data involve human subjects or have any proprietary restrictions that may affect use and sharing.

It is also useful to list the types of data that will be collected as part of the project. Keep in mind that many research sponsors and journals define data broadly to include physical and biological specimens, software algorithms and code, and educational materials. Data types can include text, spreadsheets, audio recordings, movies and images, geographic information system data layers, patient records, surveys, and interviews. Each data type may have multiple options for data and file formats. It is usually a good idea to store data in unencrypted, uncompressed, non-proprietary formats that are based on open standards that are widely employed by the scientific community.

Both the volume of data and number of data files affect hardware, software, and personnel needs. For example, spreadsheets have limits to the number of cells (i.e., data values) that can be recorded and they are not designed for managing geospatial data.

### ***2.2.3 Data Organization***

Once the types and volume of data to be collected are known, it is then desirable to plan how the data will be organized and, if possible, identify the tools that will be used. A spreadsheet program like Microsoft Excel or LibreOffice Calc may be sufficient for a few relatively small data tables (tens of columns, thousands of rows), but would not be applicable for a project where many large data files are generated. In cases where many large data files are anticipated, a relational database management system (e.g., ORACLE or MySQL), a Geographic Information System (e.g., ArcGIS, GRASS, QGIS), or NoSQL database (e.g., MongoDB) may be more appropriate (see Porter 2017). For most classes of software including database programs, there are numerous commercial and free or inexpensive open source programs available (Hampton et al. 2015). That said, it is important to consider the skills and training that may be required to effectively use different types of software.

### ***2.2.4 Quality Assurance/Quality Control***

Quality assurance and quality control (QA/QC) refer to the approaches that are used to assess and improve data quality. Some research sponsors and funding programs impose specific requirements on the QA/QC procedures and standards that should be followed by researchers. In most cases, however, QA/QC is up to the individual

researcher(s). Regardless, research sponsors, reviewers, and project personnel benefit from knowing that sound QA/QC procedures will be employed prior to, during, and after data collection (see Michener 2017a). For example, many data errors can be prevented from occurring or minimized by providing project personnel with training in instrumentation and data collection, and by adopting a routine maintenance and calibration schedule. Double blind manual data entry (when human data entry is required) and automated laboratory information systems can also prevent data entry errors or, minimally, make it easy to detect and rectify such errors when they occur. Various statistical and graphical approaches can be used to detect and flag anomalous values in the data (see Michener 2017a).

### 2.2.5 Documentation

Metadata—the details about how, where, when, why and how the data were collected, processed and interpreted—should be as comprehensive as possible. Human memory is not infallible. Specific details are usually the first to be forgotten but, eventually, even the more general information about a project is lost. Seemingly minor details, such as the model and serial number of an analytical instrument, often prove crucial when one attempts to verify the quality of a data value or reproduce a result. The metadata provide a comprehensive record that can be used by you and others to discover, acquire, interpret, use, and properly cite the data products generated as part of the research (see Michener 2017b).

A good approach is to assign a responsible person to document data and project details in a shared document or electronic lab notebook that is available to all project personnel. The documentation should be routinely reviewed and revised by another team member and backed up in one or more safe locations. This documentation provides the foundation for the metadata that will be associated with project data products that will be stored, reused, and shared with others.

The DMP should minimally include a concise description of how data will be documented. This description ideally includes:

- Metadata standards that will be adopted by the project [e.g., Dublin Core (see Dublin Core ® Metadata Initiative 2016), Ecological Metadata Language (Fegeus et al. 2005)]
- Metadata tools that will be used to create and manage project metadata [e.g., Morpho (Higgins et al. 2002)]
- Identification of who is responsible for creating and managing the metadata



## 2.2.6 *Storage and Preservation*

Laptop and desktop computers and websites generally have a lifespan of just a few years. All storage media can be expected to either degrade gradually over time or experience catastrophic failure. Thus, short-term data backup and long-term data preservation are key components of a sound DMP. The plan should specifically address three issues:

- how long the data will be accessible after the project ends
- the backup procedures that are to be followed throughout the project
- where and how the data and associated code will be stored for the short- and long-term

Planned data longevity depends upon several factors. For instance, the research sponsor, the research community to which you belong, or your home institution may have specific guidelines, norms or requirements. It is also important to consider the value of the data as a resource. Long-term ecological research data and other data that cannot be easily replicated, such as observations of environmental phenomena like natural disturbances or expensive experimental data, should typically be preserved for the long-term. Easily replicated experimental data may have a much shorter period of relevance (e.g., months to a few years). Other data such as simulation data and intermediate data products may be kept for a short period of time (days to months) or may not need to be preserved at all, especially if the software code or models that generated the data are retained.

Accidents and disasters happen. Data should be protected throughout the course of the project. A good strategy is to store at least three copies of the data in two separate locations. For example, data should minimally be stored on the original desktop or laptop computer, on an external hard drive that can be stored in a safe or locked cabinet, and at one or more offsite locations such as an institutional data repository or a commercial data storage service like Amazon, Dropbox, or Google. Your backup plan should indicate the location and frequency of backup, who is responsible for backup, as well as procedures for periodically verifying that backups can be retrieved and read.

Long-term preservation (e.g., years to decades) requires that data and associated code and workflows be deposited in a trusted data center or repository. Many agencies, organizations or disciplines support specific repositories for particular types of data. Examples include GenBank for nucleotide sequence data (Benson et al. 2013; NCBI 2016), Global Biodiversity Information Facility for biodiversity data (Flemons et al. 2007; GBIF 2016) and the US National Centers for Environmental Information for climate, coastal and marine data (NCEI 2016). Other examples of discipline-specific data repositories are listed and discussed in Cook et al. (2017). Useful resources and examples of general science repositories for data, code and workflows are included in Table 2.1.

**Table 2.1** Useful registries and general repositories for data, code, workflows, and related outputs

Repository name	URL/References	Description of services
BioSharing	<a href="http://www.biosharing.org">http://www.biosharing.org</a> ; Sansone et al. (2012)	<b>Registry</b> of community-based data and metadata reporting standards, policies, and databases for the biological, natural and biomedical sciences
Dryad	<a href="http://datadryad.org/">http://datadryad.org/</a> ; Vision (2010)	<b>Repository</b> for a diverse array of data that underlie scientific publications; data are easily discovered, freely reusable, and citable
Figshare	<a href="http://figshare.com/">http://figshare.com/</a>	<b>Repository</b> where researchers can preserve and share data, figures, images, and videos
GitHub	<a href="https://github.com/">https://github.com/</a>	<b>Repository</b> for code (primarily) that supports <b>distributed revision control</b> and <b>source code management</b>
KNB or the Knowledge Network for Biocomplexity	<a href="https://knb.ecoinformatics.org/">https://knb.ecoinformatics.org/</a> ; Andelman et al. (2004)	<b>Repository</b> for ecological and environmental data from individuals and institutions world-wide
myExperiment	<a href="http://www.myexperiment.org">http://www.myexperiment.org</a> ; Goble et al. (2010)	<b>Repository</b> of scientific workflows for a variety of workflow systems (e.g., Taverna, Kepler)
REgistry of REsearch data Repositories	<a href="http://www.re3data.org/">http://www.re3data.org/</a> ; Pampel et al. (2013)	<b>Registry</b> of research data repositories on the web
Zenodo	<a href="http://zenodo.org">http://zenodo.org</a>	<b>Repository</b> where researchers can store and share data, text, spreadsheets, audio, video, and images across all fields of science

### 2.2.7 Data Integration, Analysis, Modeling and Visualization

Researchers can rarely predict all data integration, analysis, modeling and visualization procedures that will be employed during a project. It is useful, however, to identify the software and algorithms that will be used or created during the project planning. Some software products are complex, expensive and difficult to use. In such cases, budgetary resources for training and purchasing and supporting the software (see Sect. 2.11) will be essential to include. Oftentimes, new code or software tools will necessarily be generated as part of a project. Ideally, the DMP will include a description of the software, models and code that will be employed or developed during the project. It is a good idea to document procedures for managing, storing and sharing any new code, models, software and workflows that will be created.

## 2.2.8 Data Policies

It is necessary to understand any legal requirements that may affect your proposed research such as regulations associated with intellectual property rights and data pertaining to human subjects, endangered and threatened species, and other sensitive material. Furthermore, it is good practice and often required by research sponsors to initially document project policies with respect to data use, data sharing, and data citation. Three issues should be considered as you develop your DMP.

First, will your project make use of pre-existing materials such as data and code? If so, document any licensing and sharing arrangements in the DMP. Proprietary restrictions and intellectual property rights laws may prevent or limit your capacity to use and redistribute code and software.

Second, will your project access, generate or use data that deal with human subjects, live animals, endangered and threatened species, issues of national security or competitiveness, or other sensitive material? If so, the research sponsor and your home institution will generally have a set of formal procedures that must be followed to obtain permission. Usually, you must receive approval from an Institutional Review Board before the research is undertaken or before the grant proposal is submitted. Approvals may be granted with certain stipulations such as that informed consent must be granted or that data are anonymized or presented in a way that humans and specific locations cannot be identified.

Third, what are your plans for sharing, embargoing, and licensing data and code? Increasingly, research sponsors, publishers and reviewers expect or require that data be made available when findings based on the data are published. Likewise, data collected by graduate students should be shared no later than when the thesis is published or the graduate degree is awarded. Embargoes or delays in data availability associated with publications, patent applications, or other reasons should be explicitly stated in the DMP. A good practice is to adopt a license that specifies how data and other intellectual products may be subsequently used. Table 2.2 provides a brief description of relevant licenses from the Creative Commons Organization. The Dryad data repository, for instance, has adopted the CC0 (CC Zero) Waiver as the *de facto* standard for how all data deposited in the repository should be treated. Dryad also specifies how data products should be cited by others (Box 2.1).

### Box 2.1 Recommended Data Citation Guidelines from Dryad Digital Repository (2016)

#### “How do I cite data from Dryad?”

When citing data found in Dryad, **please cite both the original article as well as the Dryad data package**. It is recommended that the data package be cited **in the bibliography** of the original publication so that the link between the publication and data is indexed by third party services. Dryad provides a

(continued)

**Box 2.1** (continued)

generic citation string that includes authors, year, title, repository name and the Digital Object Identifier (DOI) of the data package, e.g.

*Westbrook JW, Kitajima K, Burleigh JG, Kress WJ, Erickson DL, Wright SJ (2011) Data from: What makes a leaf tough? Patterns of correlated evolution between leaf toughness traits and demographic rates among 197 shade-tolerant woody species in a neotropical forest. Dryad Digital Repository. <http://dx.doi.org/10.5061/dryad.8525>*

Dryad also assigns a DOI to each data file, which should only be used in contexts where the citation to the data package as a whole is already understood or would not be necessary (such as when referring to the specific file used as part of the methods section of an article)."

If you are using a large number of data sources, it may be necessary to provide a list of the relevant data packages/files rather than citing each individually in the References. The list can then be submitted to Dryad so others who read your publication can locate all of the original data.

Legal requirements and sponsor and institutional policies may be confusing or, even, difficult to discover. Whenever doubt exists, it is good practice to contact someone from your institution's sponsored research office or Institutional Review

**Table 2.2** The Creative Commons licenses (Creative Commons Corporation 2016)

License	Description
No Rights Reserved [CC0 (tool)]	"Allows licensors to waive all rights and place a work in the public domain"
Attribution (CC BY)	"Lets others distribute, remix, tweak, and build upon your work, even commercially, as long as they credit you for the original creation"
Attribution-NonCommercial (CC BY-NC)	"Lets others remix, tweak, and build upon your work non-commercially, and although their new works must also acknowledge you and be non-commercial, they don't have to license their derivative works on the same terms"
Attribution-NoDerivs (CC BY-ND)	"Allows for redistribution, commercial and non-commercial, as long as it is passed along unchanged and in whole, with credit to you"
Attribution-ShareAlike (CC BY-SA)	"Lets others remix, tweak, and build upon your work even for commercial purposes, as long as they credit you and license their new creations under the identical terms"
Attribution-NonCommercial-ShareAlike (CC BY-NC-SA)	"Lets others remix, tweak, and build upon your work non-commercially, as long as they credit you and license their new creations under the identical terms"
Attribution-NonCommercial-NoDerivs (CC BY-NC-ND)	Allows "others to download your works and share them with others as long as they credit you, but they can't change them in any way or use them commercially"

Board, a data librarian at your academic library, or the program manager(s) for the research program to which you may be applying.

### **2.2.9 *Communication and Dissemination of Research Outputs***

A good DMP describes what data products will be generated as well as when and how they will be shared with others. Passive and, generally, more ineffective approaches to data sharing include posting the data on a personal website or emailing the data upon request. Active and more effective approaches include publishing the data: (1) as a contribution to an open data repository (see Sect. 2.6 and Chap. 6); (2) as a supplement to a journal article as described above for Dryad (Sects. 2.6 and 2.8); or as a standalone data paper that includes the data, metadata, and, possibly, associated code and algorithms. Examples of journals that publish data papers include the Ecological Society of America's *Data Papers*, *Scientific Data* (a *Nature* publication), the *GeoScience Data Journal* (a Wiley publication in association with the Royal Meteorological Society), and *GigaScience* (a joint BioMed Central and Springer publication). More active approaches may require a little more work upfront in terms of generating sufficient metadata and adhering to data formatting and other requirements. However, significant time and effort may be saved in the long-term as the data originator no longer needs to respond to queries or attempt to maintain a website or individual data repository.

### **2.2.10 *Roles and Responsibilities***

It is good practice to delineate the roles and responsibilities of project personnel including time allocations if possible. Consider who will be responsible for data collection, data entry, metadata creation and management, QA/QC, data preservation, and analysis. Make note of the management support activities (e.g., systems administration, high-performance computing and data archival) that will be performed by other individuals or organizations. Identifying roles and responsibilities as part of the DMP helps ensure that the data will be appropriately managed and that the staff needs are adequate. Research sponsors and reviewers are often reassured that a DMP will be adhered to when named individuals are associated with key project tasks. Moreover, clear articulation of roles and responsibilities prevents confusion among project personnel.

### **2.2.11 Budget**

Data management is a non-trivial activity that costs money and takes time. The dollar amount and percentage of a budget devoted to data management can vary enormously from one project to another. Projects that involve collection and management of a small amount of straightforward data may suffice on less than 5% of the budget being devoted to data management. Projects involving massive amounts of data and complex analyses and modeling may require that more than 50% of the budget be devoted to data management. Most projects fall in between the two extremes (e.g., 10–25% of the project budget devoted to data management).

A good DMP ideally includes a budget or pointers to budget lines that demonstrate that financial resources are available to support the requisite hardware, software, services, and personnel allocations (Sect. 2.10). Consider real project costs as well as in-kind support that may be covered by your organization (e.g., systems administration, high-performance computing). If you plan to use commercial or other service providers for particular activities (e.g., for data backup, long-term storage and preservation), make sure that their fees are appropriately budgeted.

## **2.3 Developing and Using a Data Management Plan**

Section 2 described the various components that may be included in a comprehensive DMP. This section addresses issues such as when and how the DMP is created and by whom (Sect. 3.1), as well as how the DMP can be most effectively used during the project (Sect. 3.2).

### **2.3.1 Best Practices for Creating the Plan**

Good data management plans, like well-written research papers, require time to evolve and mature. A wide array of data and metadata standards, data management approaches, and data repositories are often available to meet the needs of a specific community. Choosing among the various options requires deliberation. An effective tactic is to start filling in a draft data management plan template as soon as key decisions are made such as those related to methods, data sharing, and choice of a data repository for long-term storage. Much of the information included in a data management plan may be excerpted directly from proposal text or possibly from other plans that you and your colleagues have previously prepared. The emerging draft can then be shared with colleagues and others who can incorporate their best ideas. In so doing, the plan becomes a living and more useful document from the onset.



Few researchers are taught data management skills. An effective strategy is to seek input from colleagues that have created and implemented data management plans—i.e., request a copy of their plan(s), review, and ask questions. Librarians at many research universities provide data management services that include guidance about data management plans, metadata standards and tools, and trusted data repositories. One may also view and take ideas from plans that have been created by others and published on the DMPTool website.

Increasingly, research is a team effort. A typical project may engage one or more senior researchers, a post-doctoral associate, and one or more graduate and undergraduate students. Each of these individuals will likely “touch” the data at some point in the research process, potentially affecting the quality of the data and the interpretations. It is good practice to actively engage the entire team in developing the data management plan. In so doing, you are seeking their best ideas as well as their buy-in to the plan. Buy-in is critical, as the entire team must implement the plan.

Many excellent tools, often open-source, exist for creating and managing metadata, performing QA/QC, and analyzing and visualizing data. It is recommended that you use the best, widely available tools whenever possible. Reviewers of your data management plan and your colleagues will appreciate the fact that you are focusing valuable time on research as opposed to creating new tools.

It is good practice to use and cite a community standard if it exists and if it is sufficient for the task at hand. All too often, inexperienced researchers create their own unique methodologies, procedures, and standards (e.g., data encoding schema, metadata formats, etc.). Adopting good community standards of practice will save you time and effort from “reinventing the wheel.” Furthermore, community standards can typically be cited and are more likely to be perceived favorably by reviewers.

### **2.3.2 *Using the Plan***

A DMP should be viewed and treated as a living document. An effective approach is to use and re-visit your plan frequently—at least on a quarterly basis. The plan represents a valuable resource for new students and staff that are brought onto the project team. Plans should be revised to reflect any new changes in protocols and policies. Laboratory and project team group meetings are ideal times for reviewing and revising plans. It is important to track and document any changes to the DMP in a revision history that lists the date that any changes were made to the plan along with the details about those changes.

## 2.4 Conclusion

A good data management plan will provide you and your colleagues with an easy-to-follow road map that will guide how data are treated throughout the life of the project and afterwards. No plan will be perfect from the start. This chapter provides guidance with respect to the components and content included in a DMP. Some research sponsors may require only a two-page synopsis of a DMP. However, by considering all of the components described in Sect. 2, your plan is likely to be more thorough, realistic, and adequately budgeted and staffed. Section 3 offers suggestions about preparing and using the DMP. In particular, best practices dictate that one: (1) starts early in the process to create the DMP; (2) seeks input and examples from others; (3) engages all relevant project personnel; (4) uses common and widely available tools for data management activities; and (5) follows and adopts community practices and standards. Lastly, the best DMPs are those that are referred to frequently, reviewed and revised on a routine basis, and recycled (i.e., the most effective and proven approaches are used again in subsequent projects).

## References

- Andelman SJ, Bowles CM, Willig MR et al (2004) Understanding environmental complexity through a distributed knowledge network. *BioSci* 54:243–249. doi:[10.1641/0006-3568\(2004\)054\[0240:UECTAD\]2.0.CO;2](https://doi.org/10.1641/0006-3568(2004)054[0240:UECTAD]2.0.CO;2)
- Benson DA, Cavanaugh M, Clark K et al (2013) GenBank. *Nucleic Acids Res* 41(Database issue): D36–D42. doi:[10.1093/nar/gks1195](https://doi.org/10.1093/nar/gks1195)
- Consortium for Ocean Leadership (2010) Ocean observatories initiative: final network design. [http://www.oceanobservatories.org/wp-content/uploads/2012/04/1101-00000\\_FND\\_OOI\\_ver\\_2-06\\_Pub.pdf](http://www.oceanobservatories.org/wp-content/uploads/2012/04/1101-00000_FND_OOI_ver_2-06_Pub.pdf). Accessed 14 Apr 2016
- Cook RB, Wei Y, Hook LA et al (2017) Preserve: protecting data for long-term use, Chapter 6. In: Recknagel F, Michener W (eds) *Ecological informatics. Data management and knowledge discovery*. Springer, Heidelberg
- Creative Commons Corporation (2016) Creative Commons. <https://creativecommons.org>. Accessed 14 Apr 2016
- Digital Curation Center (2016) About DMPonline. [https://dmponline.dcc.ac.uk/about\\_us](https://dmponline.dcc.ac.uk/about_us). Accessed 14 Apr 2016
- DMPTool (2016) Data management planning tool. <https://dmptool.org>. Accessed 14 Apr 2016
- Dryad Digital Repository (2016) Dryad. <http://datadryad.org>. Accessed 14 Apr 2016
- Dublin Core ® Metadata Initiative (2016) DCMI home: dublin core metadata initiative (DCMI). <http://dublincore.org>. Accessed 14 Apr 2016
- Fegraus EH, Andelman S, Jones MB et al (2005) Maximizing the value of ecological data with structured metadata: an introduction to Ecological Metadata Language (EML) and principles for metadata creation. *Bull Ecol Soc Am* 86:158–168
- Flemons P, Guralnick R, Krieger J et al (2007) A web-based GIS tool for exploring the world's biodiversity: The Global Biodiversity Information Facility Mapping and Analysis Portal Application (GBIF-MAPA). *Ecol Inf* 2(1):49–60
- Global Biodiversity Information Facility (GBIF) (2016) Global Biodiversity Information Facility: free and open access to biodiversity data. <http://www.gbif.org>. Accessed 14 Apr 2016

- Goble CA, Bhagat J, Aleksejevs S et al (2010) myExperiment: a repository and social network for the sharing of bioinformatics workflows. *Nucleic Acids Res* 38(suppl 2):W677–W682. doi:10.1093/nar/gkq429
- Hampton SE, Anderson SS, Bagby SC et al (2015) The Tao of open science for ecology. *Ecosphere* 6:art120. <http://dx.doi.org/10.1890/ES14-00402.1>
- Higgins D, Berkley C, Jones M (2002) Managing heterogeneous ecological data using Morpho. In: Proceedings of the 14th international conference on scientific and statistical database management, pp 69–76
- Michener WK (2017a) Quality assurance and quality control (QA/QC), Chapter 4. In: Recknagel F, Michener W (eds) *Ecological informatics. Data management and knowledge discovery*. Springer, Heidelberg
- Michener WK (2017b) Creating and managing metadata, Chapter 5. In: Recknagel F, Michener W (eds) *Ecological informatics. Data management and knowledge discovery*. Springer, Heidelberg
- Michener WK, Waide RB (2009) The evolution of collaboration in ecology: lessons from the United States Long Term Ecological Research Program. In: Olson GM, Zimmerman A, Bos N (eds) *Scientific collaboration on the Internet*. MIT Press, Boston, pp 297–310
- Michener WK, Porter J, Servilla M et al (2011) Long term ecological research and information management. *Ecol Inf* 6:13–24
- National Center for Biotechnology Information (NCBI) (2016) GenBank overview. <http://www.ncbi.nlm.nih.gov/genbank/>. Accessed 14 Apr 2016
- National Centers for Environmental Information (NCEI) (2016) NOAA National Centers for Environmental Information. <https://www.nodc.noaa.gov>. Accessed 14 Apr 2016
- Pampel H, Vierkant P, Scholze F et al (2013) Making research data repositories visible: the re3data.org registry. *PLoS One* 8:e78080. doi:10.1371/journal.pone.0078080
- Peters DPC, Loescher HW, SanClements MD et al (2014) Taking the pulse of a continent: expanding site-based research infrastructure for regional- to continental-scale ecology. *Ecosphere* 5:29. <http://dx.doi.org/10.1890/ES13-00295.1>
- Porter JH (2017) Scientific databases for environmental research, Chapter 3. In: Recknagel F, Michener W (eds) *Ecological informatics. Data management and knowledge discovery*. Springer, Heidelberg
- Porter JH, Nagy E, Kratz TK et al (2009) New eyes on the world: advanced sensors for ecology. *BioSci* 59:385–397
- Porter JH, Hanson PC, Lin C-C (2012) Staying afloat in the sensor data deluge. *Trends Ecol Evol* 27:121–129
- Sansone S-A, Rocca-Serra P, Field D et al (2012) Toward interoperable bioscience data. *Nat Genet* 44:121–126. doi:10.1038/ng.1054
- Schimel D, Keller M, Berukoff S et al (2011) NEON science strategy: enabling continental-scale ecological forecasting. NEON, Inc., Boulder, CO
- Vision TJ (2010) Open data and the social contract of scientific publishing. *BioSci* 60:330–330. doi:10.1525/bio.2010.60.5.2

# Chapter 3

## Scientific Databases for Environmental Research

John H. Porter

**Abstract** Databases are an important tool in the arsenal of environmental researchers. There are a rich variety of database types available to researchers for the management of their own data and for sharing data with others. However, using databases for research is not without challenges due to the characteristics of scientific data, which differ in terms of longevity, volume, diversity and ways they are used from many business applications. This chapter reviews some successful scientific databases, pathways for developing scientific data resources, and general classes of Database Management Systems (DBMS). It also provides an introduction to data modeling, normalization and how databases and data derived from databases can be interlinked to produce new scientific products.

### 3.1 Introduction

The development of environmental databases constitutes a new paradigm in the evolution of environmental research. The traditional model for ecological research has been investigator-based. A researcher and his or her students would collect data, analyze that data and publish the results. The data underlying the publication and analyses, more often than not, would then be placed into file cabinets, never to be seen again. This model assumes that the full utility of the data is “consumed” by the publications based on the data in a short period immediately following the collection of the data. The model is inefficient, in the sense that different researchers may duplicate data collection efforts, but it does not require resources for the development of formal databases. Where the questions being asked are localized in space and time and the number of types of data needed is limited to those a single researcher is capable of gathering, the traditional model for environmental data has been sufficient.

When the types of research needed to answer pressing environmental questions are neither localized in space or in time the traditional model fails. Questions

---

J.H. Porter (✉)

Department of Environmental Sciences, University of Virginia, Charlottesville, VA 22904, USA

e-mail: [jporter@virginia.edu](mailto:jporter@virginia.edu)

concerning the nature, extent and causes of global change, and regional analyses require data resources that extend beyond those which it is logistically possible for an individual researcher to collect. Understanding of complex ecosystem processes can require a multidisciplinary approach that transcends individually-collected data. Similarly, understanding long-term processes often requires stewardship of data on decadal scales and longer. Data not only needs to be preserved, but it needs to be made available. As an editorial in the prestigious journal *Nature* put it: “Research cannot flourish if data are not preserved and made accessible” (Campbell 2009). Many environmental journals, as in other areas of science, have instituted requirements that the data used in a paper be accessible (Whitlock et al. 2010), and science funding agencies have developed new policies dictating that the products of federally-funded science, including the data, be made available (Holdren 2013). The requirements for long-term curation of data and the need for data exchange between scientists to support integrative and synthetic analyses are facilitated by the development of databases at a number of different levels, from the investigator to the institution and even to the discipline.

There are several other advantages to developing and using scientific databases. The first is that databases lead to an overall improvement of data quality. Multiple users provide multiple opportunities for detecting and correcting problems in data. A second advantage is cost. Data costs less to save than to collect again. Often, environmental data cannot be collected again at any cost because of the complex of poorly controlled factors, such as weather, that influence population and ecosystem processes.

However, the primary reason for developing scientific databases must be the new types of scientific inquiry that they make possible (Hampton et al. 2013). Gilbert (1991) discusses the ways databases and related information infrastructure are leading to a paradigm shift in biology (Fig. 3.1). Nowhere has this been more evident than in the genomic community, where the creation of databases and associated tools have facilitated a tremendous increase in our understanding of the relationship between the genetic sequences and the actions of specific genes.

The environmental area is beginning to see a similar renaissance, brought on through improvements in databases and data sharing. Specific inquiries which require databases include long-term studies, which depend on databases to retain project history, syntheses, which combine data for a purpose other than which it was originally collected, and integrated multidisciplinary projects, which depend on



**Fig. 3.1** Archiving and sharing data enable new types of analyses. The traditional model of data sharing (*top row of boxes, open arrows*) limits data use to one or more publications. When data is archived and shared (*second row, solid arrows*) it enables new types of studies as data is combined with other data, or new theories are tested

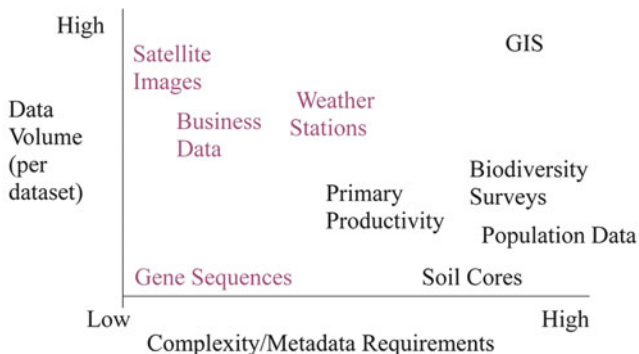
databases to facilitate sharing of data. Public decisions involving environmental policy and management frequently require data that are regional or national, but most ecological data is collected at smaller scales. Databases make it possible to integrate diverse data resources in ways that support decision-making processes.

## 3.2 Challenges for Scientific Databases

Scientific databases face challenges that are different than those experienced by most business-oriented databases (Pfaltz 1990; Robbins 1995). As noted by Robbins (1995), the technologies supporting business databases emphasize data integrity and internal consistency. It would not do to have two disparate estimates of hours worked when paychecks are being issued! However, scientific databases may well contain different observations of the same phenomena that are inconsistent, resulting from differences in methodology and measurement imprecision or even different models of the physical processes underlying the process under study. Additionally, with the exception of correction of errors, scientific data is seldom altered once it has been placed in a database. This contrasts with business data where an account balance may be altered repeatedly as funds are expended. For this reason, several authors (Cinkosky et al. 1991; Robbins 1994, 1995; Strebel et al. 1994, 1998; Meeson and Strebel 1998; Costello 2009) have proposed that a publication model, rather than a traditional database model, is the correct model for scientific data databases. This approach has wide acceptance, although Parsons and Fox (2013) discuss some limitations and alternatives to this approach.

The biggest challenge for scientific databases is dealing with diversity. Science means asking *new* questions. Scientific databases thus need to be adaptable so that they can support new kinds of queries. In most business-oriented databases, the focus is on development of standardized queries. This month's sales report is similar in form (although not content) to last month's report. Business-oriented database software has many features that aid in the production of standardized reports. In contrast, the scientific focus is on identifying new relationships within a given dataset or finding new linkages with other datasets. To this end, both graphical and statistical analyses are required that transcend the capabilities of business-oriented database products.

The volume and complexity of scientific data vary widely (Fig. 3.2). Some types of data have a high volume but are relatively homogeneous. An example of this is image data from satellites. Although each image may require hundreds of megabytes to store, the storage formats of the data are relatively standardized and hence require relatively little metadata for use. In contrast, certain types of manually collected environmental data have an extremely small volume but require extensive



**Fig. 3.2** Data vary widely in both volume and complexity. Most software is designed for moderate data volumes and relatively low complexity (e.g., Business Data), but most ecological data (*bold*) is relatively complex, requiring detailed metadata

documentation to be useful. For example, deep soil cores are very expensive to obtain, so data is usually restricted to a few cores. However, these cores may be exposed to many different analyses examining the density, mineral content, physical characteristics, biological indices, isotopic ratios, etc. Each of these analyses needs to be well documented, making the metadata required exceed the volume of numerical data itself by several orders of magnitude. Some data is both high in volume and complex. Geographic Information System (GIS) data layers can be very high in volume (depending on the resolution of the system) and require metadata which covers not only the actual data collection, but the processing steps used to produce it (FGDC 1994).

Some classes of data are referred to as “Big Data.” They are often characterized by the “3 Vs”—large Volume, high Velocity and wide Variety (Madden 2012). These characteristics make the data difficult to manage or analyze using conventional software and analyses. Typical sources of Big Data are social media sites, where millions (volume) of messages or postings appear each day (velocity) in a wide diversity of forms (variety), and automated sensors and sensor systems. Some researchers see the use of Big Data as allowing a whole new realm of analyses based on correlation and exhaustive sampling of populations (Mayer-Schönberger and Cukier 2013) in some cases rendering conventional model and theory-based analyses obsolete (Anderson 2008), whereas others caution that correlation alone is an unreliable guide to future behavior and that sampling bias can occur even in large datasets (Harford 2014). Regardless, there is no question that both massive, rapidly growing and diverse datasets generated by a growing array of automated sensors pose challenges for science (Porter et al. 2012).

Scientific data is heterogeneous and diverse. In some areas of science (e.g., genomics, water quality) there is wide agreement on particular types of measurements and the techniques for making them. However, in other areas (e.g., the measurement of primary productivity) there is relatively little agreement on standards and the types of data collected are much more diverse. Standards are most

common in “mature” areas of inquiry (Yarmey and Baker 2013). In less mature areas of science, experimentation with methodologies is a necessary part of the scientific process. Eventually, there is a convergence in methodologies, which leads to the informal adoption of emergent standards, which may subsequently be adopted as formal standards. This process is especially rapid where there are a limited number of specialized instruments for making measurements. Conversely, the standardization process is especially difficult where a methodology needs to operate across a range of physical systems. Techniques that are developed for aquatic systems may be impossible to apply in forested systems.

The challenge of diversity extends to users as well. Scientific users have different backgrounds and goals that need to be supported by the database. Moreover, the user community for a given database will be dynamic as the types of scientific questions being asked change and new generations of scientists use the database (Pfaltz 1990).

Scientific databases require a long-term perspective that is foreign to many other types of databases. In business, there is a limited need for maintaining most types of records beyond several years. Although tax records might be maintained, inventory and payroll, and contact data are of little use after 2–5 years. However, for environmental research, data retains utility for many decades. Indeed, data that are centuries old are particularly valuable as they allow us to assess changes that would otherwise be invisible to us (Magnuson 1990).

A frequently cited goal is that an environmental database should assure that data is both accessible and interpretable 20-years in the future [Justice et al. (1995) extend this goal to 100 years]. Reaching this goal depends on overcoming technological, cultural and semantic barriers (Cook et al. 2017). There is a technological requirement for persistent media that does not degrade over time or become technologically obsolete (otherwise you could end up with long-term data, but no way to read it).

All storage media tend to degrade over time, both due to physical factors, because the magnetic signals become indistinct over time and charges fade. It has been said that there is no valuable data on 20-year old magnetic tapes, simply because there is no readable information left on the tape after 20 years. Additionally, the rapid developments in technology render old media physically incompatible with newer computers. Gone are the floppy disks and reel-to-reel magnetic tape drives that once were the standard for offline storage. Although DVDs (Digital Video Disks) are still in use, they are rapidly being supplanted by flash memory drives, which are in turn being replaced by cloud resources. Also serious is the high rate of change in the formats in which information is stored. Information in proprietary formats, such as a spreadsheet file from the early 1980s, may be difficult or impossible to interpret once that software is no longer available. There has been a progression of “standard” spreadsheet programs, from DigiCalc to Lotus 1-2-3 to Microsoft Excel, each using different formats. Even within a single product line formats can change sufficiently that files can no longer be read. For example, Microsoft Access 2013 no longer can read databases that use the Access 97 file format, rendering that data inaccessible to future users.



Despite the challenge posed by technological change, this is the easiest barrier to overcome. Active information management, where information on older media are transferred to new media and new software formats, can assure that data is not lost due to media failure or technological change. There is a transition period between major storage technologies of approximately 5–10 years and during that time the translation between media is relatively easy. However, if that window of opportunity is missed, due to inattention or lack of funds, data may be irrevocably lost or impracticably expensive to recover.

Long-term interpretability of data also depends on capturing the context of data collection. Not all the information needed to understand and use the data is inherent in the data itself. Strebel et al. (1994) and Michener et al. (1997) discuss the rapid “decay” of data that is not actively managed. A critical feature of that decay is the loss of detailed information about how the data was collected and the processing steps to which it has been subjected. This loss occurs early in the process, resulting in a steep decrease in the value of data if steps are not undertaken to capture the information before it fades from the short-term memory of the researchers collecting the data. Strebel et al. (1994) note that the decline in data usability comes with increasing costs for use over time. At periods of less than a decade, the costs of data use can become prohibitive if data are not adequately documented. To create consistent, human and machine-readable metadata a number of standards have been developed. There is a huge array of discipline-based standards, including: Dublin Core, MARC, METS (McCray and Gallagher 2001; Guenther and McCallum 2003) in the library community, Darwin Core for museum collections (Wieczorek et al. 2012) and FGDC and ISO19115 for geographic information systems (Nogueras-Iso et al. 2004). In the ecological area, Ecological Metadata Language (EML) (Fegraus et al. 2005) has been formally adopted by the Long-Term Ecological Research Network and used for Ecological Society of America and Organization of Biological Field Station data registries. Storing metadata in machine-readable ways, such as eXtensible Markup Language (XML), makes it possible to automate, or at least semi-automate crosswalks and conversions among the different standards. See Chap. 5 (Michener 2017) for more details on metadata and its uses.

Finally, to assure long-term usability, terms used in documentation need to be well defined so that semantic differences do not cloud future interpretation. Terms, which have one meaning now, may have different meanings in the future or in different disciplines. For example, taxonomic identifiers associated with a given entity may vary over time as phylogenetic and taxonomic relationships are revised. This problem is especially severe when data is of value in several different scientific disciplines.

### 3.3 Examples of Scientific Databases

#### 3.3.1 *A Useful Analogy*

A useful analogy in examining cataloging and query systems for scientific data is to consider individual datasets as “volumes” in a database “library.” Articles, journals, chapters, books and volumes all have their analogs in scientific databases. Libraries may have different sizes and have different requirements for cataloging systems. For example, an individual might have a home “library” consisting of a relatively small number of books. The books would not be cataloged or organized but simply placed on a shelf. An individual book would be located by browsing all the titles on the shelf. For an office library consisting of hundreds of books, a common model is to group books on the shelf by general subject so that only a subset of the library needs to be browsed. However, when the number of books in a library enters the thousands to millions, as for a public library, formal cataloging procedures are required.

This model also applies to scientific databases. If there are relatively few different datasets, a simple listing of the titles of the datasets may be sufficient to allow a researcher to locate data of interest. This is the prevailing model in single-investigator and small project databases. The databases themselves are typically in the form of esoteric web pages that do not conform to metadata (information needed to use and interpret data) standards.

#### 3.3.2 *Examples of Databases*

Some databases specialize in a single or few types of data and implement sophisticated searching and analytical capabilities. Examples of this type of database are large databases such as Genbank which serves as a primary archive of genetic sequence data for the human genome project, with over one billion bases in approximately 1.6 million sequences (Benson et al. 2013), UniProt Knowledgebase which contains over 80 million sequences (UniProt Consortium 2014), and the Research Collaboratory for Structural Bioinformatics Protein Data Bank (RCSB PDB), a protein structure database which contains atomic coordinate entries for over 38,211 protein structures (Berman et al. 2000). These are very large databases with funding in excess of one million dollars per year. In the publication analogy, these databases are analogous to large, multi-volume reference works. They are highly “indexed” but focus on a restricted region of the data universe.

There are also a variety of specialized types of databases that operate on a smaller scale. For example, SPECIFY is specialized software for managing museum specimens (Specify Software Project 2016) and BIOTA is software for management of specimen-based biodiversity data (Colwell 1997). These systems are available for download and are used by a variety of institutions and

**Table 3.1** “Deep” vs. “wide” databases

“Deep” databases	“Wide” databases
<ul style="list-style-type: none"> <li>• Specialize on one or a few types of data</li> <li>• Large numbers of observations of one (or few) type(s) of data</li> <li>• Provide sophisticated data query and analysis tools</li> <li>• Tools operate primarily on data content</li> </ul>	<ul style="list-style-type: none"> <li>• Contain many different kinds of data</li> <li>• Many different kinds of observations, but relatively few of each type</li> <li>• May provide tools for locating data, but typically do not have tools for analysis</li> <li>• Tools operate primarily on metadata content</li> </ul>

investigators. In the publication analogy, they would be books in a series that share format elements and address the same topic, but have different content. Like the large databases (Genbank, RCSB PDB), these databases are “deep” rather than “wide” (Table 3.1), providing in-depth services for a particular type of data.

“Wide” databases are data repositories that attempt to capture all the data related to a specific field of science. For example, the National Centers for Environmental Information (NCEI) is operated by the National Oceanic and Atmospheric Administration (NOAA) and hosts over 20 petabytes of oceanic, atmospheric, and geophysical data (NOAA 2016). Such “data centers” often use standardized forms of metadata (e.g., GILS, FGDC, DIF) for maintaining formal catalogs with controlled vocabularies for subjects and keywords. Similarly, the National Aeronautic and Space Administration (NASA) operates a series of Distributed Active Archive Centers (DAACs) each of which specializes in supporting a particular area of earth or space science and have a varying number of different types of data sets. In the library analogy, these databases would be comparable to public libraries.

Additional “wide” databases are project-based databases. These are databases that support a particular multidisciplinary research project and may include a wide array of data focused on a particular site or research question. Examples of this type of database are the databases at individual Long-Term Ecological Research (LTER) sites (Michener et al. 2011). These databases contain data relating to a wide array of scientific topics (i.e., weather and climate, primary productivity, nutrient movements, organic matter, trophic structure, biodiversity and disturbance), along with information that supports management of the site (i.e., researcher directories, bibliographies and proposal texts). Management of the databases requires approximately 15% of the total site funding and they focus strongly on long-term data. Within the LTER network, there are diverse approaches to data management. These are dictated by the locations of researchers (at some LTER sites the majority of researchers are at a single university, at others they are at many different universities), and the types of data collected (studies of aquatic systems have different data needs than studies of terrestrial systems). Individual LTER sites internally may use different systems and metadata standards at individual sites, but use a common standard, Ecological Metadata Language (EML) for sharing of data (Michener et al. 2011). These databases are fairly “wide”, but not particularly “deep” in the sense that they provide access to a wide variety of data, but do not provide specialized visualization or analysis tools for most types of data. In the library analogy, these databases would be comparable to a large individual or small departmental library.

Some “databases,” such as individual web pages created by individual researchers may be neither “wide” nor “deep.” The level of development of such pages varies widely, as does the quality and quantity of the associated metadata. In the library analogy, the pages from a single researcher would be comparable to a very small personal library with little need for searching and cataloging capabilities. As an aggregate, across all researchers, these databases constitute a valuable resource, but one that is difficult to exploit because data can be hard to locate and metadata may be insufficient or difficult to translate into usable forms. Brackett (1996) describes such data as “massively disparate” which are “locally useful and globally junk.” Additionally, web pages are notoriously ephemeral, so they may be a poor choice for providing data over long time periods.

### 3.4 Evolving a Database

Scientific databases come in all sizes, from a database used by an individual researcher to manage specific data, to project databases that bring together many different kinds of data, to data repositories that serve a wide community. They have some common elements such as the need to preserve and provide access to data over long time periods, but also differ in the difficulty and expense of implementation. A database to manage sampling data may be set up in a matter of hours by a single individual, whereas creating a useful scientific data repository may require years of effort by a large team. Nonetheless, they share many commonalities, not the least of which is the need to evolve and change over time, driven by scientific and technical imperatives. Thus the development of a database is an evolutionary process. During its lifetime, a database may serve a dynamic community of users or purposes, and a database needs to change to meet those changing needs.

In creating a database, you need to ask four questions. The first is: “Why is this database *needed*?” Not all data is important enough to warrant long-term storage in a database. Pfaltz (1990) makes the point that, regardless of the rate of technological advancement, our ability to collect data will exceed our ability to maintain it in databases. If data is from a specific set of experimental manipulations linked only loosely to any particular place and time, it may have little value beyond use in a paper describing the result of the experiment. Similarly, data collected using non-standard methodologies may be difficult or impossible to utilize in syntheses. This is not to say that all experimental data or all data collected using non-standard methodologies should not be preserved in a database. There are many examples of where experimental evidence has been reinterpreted and access to the data may be critical to that process. Similarly, data collected using non-standard methodologies can be integrated with that collected by other methods if a reasonable degree of caution is exercised. However, if a clear scientific need that will be met by a given database cannot be identified, it may not be reasonable to devote resources to that database.

The second question that needs to be asked is: “Who will be the *users* of the database?” This question is important on two levels. First, if you can’t identify a community of users for your database, you may want to reexamine the need for the database! Second, defining the users of your database provides guidance on what database capabilities will be critical to its success. For example, a database designed for use by experts in a given field is likely to be too complicated for use by elementary school students. Ideally a database should provide data to users in a way that maximizes its immediate utility. Data needs to be made available in a form where users can manipulate it. A table embedded in a web page may provide an attractive way for viewing data, but it can be difficult to then extract data from that table in a spreadsheet or statistical package. The technological infrastructure needed to use and interpret data should be available (and preferably in common use) by the users or there is a risk that the database won’t be used (Star and Ruhleder 1996).

The third question is: “What types of *questions* should the database be able to answer?” The answer to this question does much to dictate how data should be structured within the database. The data should be structured in a way that maximizes the efficiency of the system for common types of queries. For example, a large bibliographic database needs to support searches based on both author and title, but probably doesn’t need to support searches based on page number. As noted above, the data needs to be made available in a form where it is usable. In some cases, it may be reasonable to provide multiple representations of data in forms that are applicable to different questions.

A final question is one that is often not asked, but the answer to which has much to do with the success of a database: “What *incentives* will be available for data providers?” Any database is dependent upon one or more sources of data. The traditional scientific environment provides few rewards for individuals who share data (Porter and Callahan 1994). However, over the past decade there have been significant advances, including making data citable through the use of Digital Object Identifiers (DOIs), inclusion of data products, along with traditional publications, in reports to funding agencies and, perhaps most importantly, linking contribution of data to the acceptance of scientific papers for publication (Duke and Porter 2013). Nonetheless, it is no accident that areas where databases have been particularly successful (e.g., genome databases) are those where contribution of data to databases are an integral part of the publication process. Leading genomic journals do not accept articles where the data has not already been submitted to one of several recognized sequence databases, an approach that has now been adopted by journals in other disciplines (Whitlock et al. 2010). In the absence of support from the larger scientific community, databases need to be innovative in giving something of value back to the data provider. This return can be in the form of improved error checking, manipulation into new, easier to use forms, and improvements in data input and display. Assuring proper attribution for data collection is a critical part of providing incentives for data contribution. Originators (authors) of datasets are more likely to make future contributions if previous contributions are acknowledged. Ideally this attribution should be made in a formalized citation,

which specifies the originator, date, dataset title and “publication” information, rather than as an informal acknowledgement.

### ***3.4.1 A Strategy for Evolving a Database***

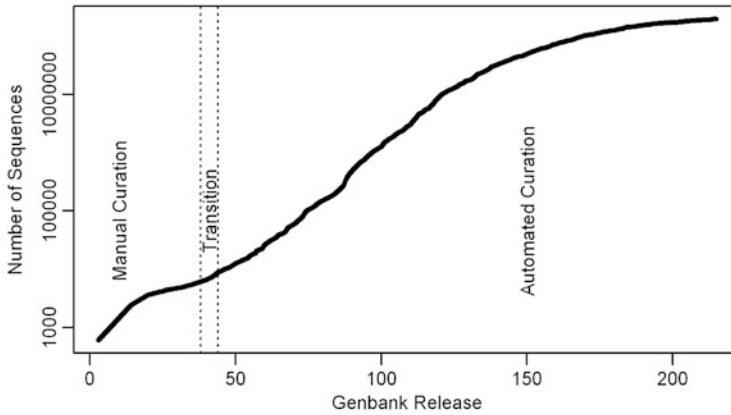
In making the myriad decisions needed to manage a database, a clear set of priorities is the developer’s most valuable friend. Every database has some things that it does well (although no part is ever perfect) and some areas that need improvement. The process of database evolution is cyclical. A part of the database may be implemented using state-of-the-art software, but several years later the state of the art has advanced to a degree that it makes sense to migrate the system to new software. Therefore, database systems should be based on current priorities, but with a clear migration path to future systems. When making decisions about the types of software to use in implementing the database and associated interfaces, it is critical to consider an “exit strategy.” Software that stores data in proprietary formats and provides no “export” capabilities are to be avoided at all costs!

The need for foresight applies to more than just software. The priorities of users may change. A keyword search capability may be a top user priority, but once it exists a spatial search capability may be perceived as increasingly important. It is not possible to implement a database system *in toto*, so the strategy adopted for development must recognize that, although some capabilities are not currently implemented, the groundwork for those capabilities in future versions must be provided for. Thus, even though an initial system may not support spatial searching, collecting and storing spatial metadata in a structured (i.e., machine-readable) form is highly desirable.

An important form of foresight is seeking scalable solutions. Scalability means that adding or accessing the 1000th piece of data should be as easy as adding the first (or easier). The genome databases faced a crisis when the flow of incoming data started to swamp the system, which depended on some level of manual curation of inputs (Fig. 3.3). The subsequent adoption of completely automated techniques for submission and quality control allows the genome databases to handle the ever-increasing flows of data. Indeed, Genbank now curates over 213 billion base pairs (NCBI 2016), a number that could never have been achieved using manual curation. Every system has some bottlenecks and their identification and elimination before they become critical is the hallmark of good planning and management.

### ***3.4.2 Choosing Software***

The choice of software for implementation of a database, be it a personal database for a specific type of data, or a large repository of diverse data, must be based on an



**Fig. 3.3** Growth of Genbank showing the type of curation used. Manual curation was becoming saturated and it was only by adopting automated curation was Genbank able to grow by several additional orders of magnitude

understanding of the tasks you want the software to accomplish (e.g., input, query, sorting, analysis) and the characteristics of the data (e.g., size, diversity). Simplicity is the watchword. The software marketplace provides an abundance of sophisticated software that is expensive and difficult to operate, but that may provide little real improvement over simpler and less expensive software. Sophistication and complexity do not always translate to utility. The factors to be considered in choosing software extend beyond the operation of the software itself. For example, is the software in the public domain (free) or commercial? Source code for public domain software is frequently available, allowing on-site customization and debugging. An additional advantage is that file formats are usually well specified (or at least decipherable using the source code). A downside is that when something is free, sometimes it is worth every cent! Difficulty of installation, insufficient documentation, bugs in the code or lack of needed features are common complaints. In contrast, commercial software comes with technical support (often for an additional charge), is frequently well documented and is relatively easy to install. However, as for public domain software, bugs in the software are not unknown! An additional problem with commercial software is that, to some degree, you are at the mercy of the developer. Source code is almost always proprietary, and file formats frequently are proprietary as well. This can create some real problems for long term archival storage if a commercial product is discontinued.

One consideration that applies to both public domain and commercial software is market share. Software that has a large number of users has a number of advantages over less frequently used software, regardless of specific features. A large user base provides more opportunities for the testing of software. Rare or unusual bugs are more likely to be uncovered if the software is widely used. Additionally, successful software tends to generate its own momentum—spawning tools that improve the utility of the software. Widely-used software also generate a host of web-accessible

forums and other information resources that can help answer even the most obscure question.

### 3.4.3 Database Management System (DBMS) Types

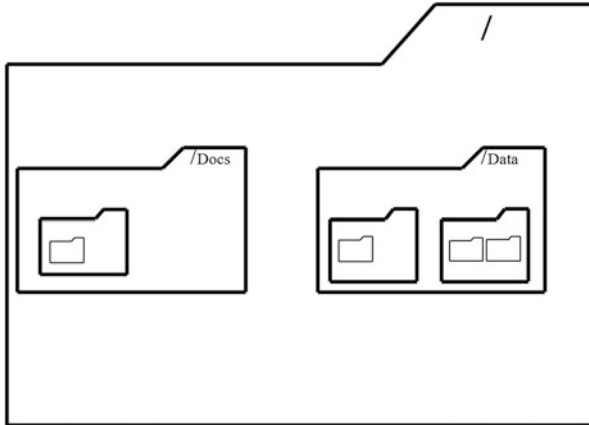
There are a wide variety of database software products available. Table 3.2 lists available options that extend beyond what might be traditionally considered as DBMS.

A file system-based database would typically not be considered a DBMS because there is no “buffer” between the physical representation of the data

**Table 3.2** DBMS types and characteristics

Type	Characteristics
<ul style="list-style-type: none"> <li>• <b>File System Based</b>—use files and directories to organize information. Examples: Gopher information servers (not typically considered a DBMS)</li> </ul>	<ul style="list-style-type: none"> <li>• Simple—can use generalized software (word processors, file managers)</li> <li>• Inefficient—as number of files increase within a directory, search speed is impacted</li> <li>• Few capabilities—no sorting or query capabilities aside from sorting file names</li> </ul>
<ul style="list-style-type: none"> <li>• <b>Hierarchical</b>—store data in a hierarchical system. Examples: IBM IMS database software, phylogenetic trees, satellite images in Hierarchical Data Format (HDF)</li> </ul>	<ul style="list-style-type: none"> <li>• Efficient storage for data that has a clear hierarchy</li> <li>• Tools that store data in hierarchically organized files are commonly used for image data</li> <li>• Relatively rigid, requires a detailed planning process</li> </ul>
<ul style="list-style-type: none"> <li>• <b>Network</b>—store data in interconnected units with few constraints on the type and number of connections. Example: Cullinet IDMS/R software, airline reservation databases</li> </ul>	<ul style="list-style-type: none"> <li>• Fewer constraints than hierarchical databases</li> <li>• Links defined as part of the database structure</li> <li>• Networks can become chaotic unless planned carefully</li> </ul>
<ul style="list-style-type: none"> <li>• <b>Relational</b>—store data in tables that can be linked by key fields. Examples: Structured Query Language (SQL) databases such as Access, Oracle, MySQL, Sybase, and SQLserver</li> </ul>	<ul style="list-style-type: none"> <li>• Widely-used, mature technology</li> <li>• Efficient query engines</li> <li>• Standardized interfaces (i.e., SQL)</li> <li>• Restricted range of data structures, may not handle image or expansive text well (although some databases allow extensions)</li> </ul>
<ul style="list-style-type: none"> <li>• <b>Object-oriented</b>—store data in objects each of which contains a defined set of methods for accessing and manipulating the data. Examples: POSTGRES database</li> </ul>	<ul style="list-style-type: none"> <li>• New, developing technology</li> <li>• Wide range of structures is extensible to handle many different types of objects</li> <li>• Not as efficient as relational for query</li> </ul>
<ul style="list-style-type: none"> <li>• <b>NoSQL</b>—“Not only SQL” databases represent a wide array of approaches, typically distributed, for dealing with “Big Data.” Examples: HAADOOP, Cassandra, MongoDB</li> </ul>	<ul style="list-style-type: none"> <li>• New, rapidly evolving</li> <li>• Individual tools prioritize different objectives (e.g., speed vs. reliability) and types of data</li> <li>• Often utilize distributed computational resources</li> </ul>





**Fig. 3.4** Directories and subdirectories can provide a way to structure data within the context of a computer file system

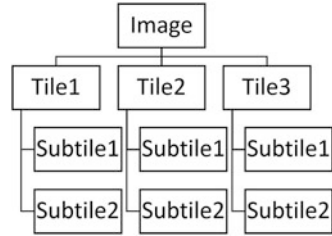
(in files and directories) and applications using the data. It lacks most of the functions commonly associated with DBMS, such as query capabilities, support for complex relationships among data types/files, enforcement of security and integrity, and error recovery (Fig. 3.4). File system-based databases typically have a heavy reliance on operating system capabilities and independent software tools to provide at least some DBMS features. However, they also have the advantage of relative simplicity and can be quite useful for data that do not encompass complex interrelationships.

Hierarchical databases, such as the IBM IMS database software, have a higher, albeit restricted range of structures (Hogan 1990). Here data are arranged in a hierarchy that makes for efficient searching and physical access (Fig. 3.5). Each entity is linked into the hierarchy so that it is linked to one, and only one, higher-level (parent) entity, although it may be linked to multiple lower-level entities (children). Note that these relationships are defined in the design of the database and are not a function of specific data stored in the database. In Fig. 3.5 each image can have multiple tile segments, but each tile is linked to only a single image.

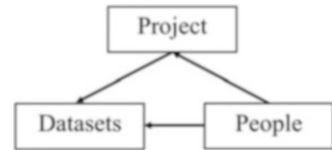
Network databases permit a wider array of relationships than hierarchical databases. Entities no longer need to be hierarchical in form (although they can be). Thus in Fig. 3.6, both projects and datasets may have links to specific people. Like hierarchical databases, the relationships are defined using pointers, not by the contents of the data. Thus modification of those relationships demands that physical changes be made to the database to update pointers.

By far the most widely used DBMS in business are relational databases, which are also widely used for scientific databases. A relational database can take on structures similar to those used in hierarchical and network databases, but with an important difference. The relational model allows interrelationships to be specified

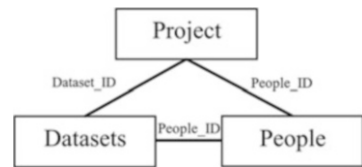
**Fig. 3.5** Example of a hierarchical database. Images are successively broken down into geographical subsets (tiles), where each tile and subtitle fall within the geographic extent of their parent element



**Fig 3.6** In a network database links between data in different tables are made explicitly



**Fig. 3.7** In a relational database, key fields such as People\_ID and Dataset\_ID are used to link tables based on the data themselves



based on key values of the data themselves. This makes it much easier to revise the structure of relational databases to add new relationships and does much to explain their popularity. In addition, relational databases benefit from a rigorous basis in mathematical theory (Bobak 1997). In Fig. 3.7, the field Dataset\_ID is shared by both the Projects and Datasets tables and is the key field for linking those tables. Similarly, People\_ID is used to link datasets and projects to people. Unlike the network database (Fig. 3.6), the links in the relational database are based on the values of key fields, not explicit pointers that are external to the records themselves.

Object-oriented models are becoming increasingly common, although most frequently those models are implemented using existing relational databases to create the structure for storing object information. Query languages for object-oriented databases are still being developed and are not standard across database vendors, unlike relational databases where the variations on the SQL standard are widely used (Keller et al. 1998). A major feature of most object-oriented databases is the ability to extend the range of data types that can be used for fields to include complex structures (e.g., image data types). They are most frequently used with object-oriented languages such as C++ and JAVA to provide persistence to program objects (Bobak 1997). This is an area of rapid innovation (Loomis and Chaudhri 1998).

NoSQL “Not only Structured Query Language” databases are a loose and rapidly growing collection of tools that may, or may not, incorporate elements of relational databases. They were often developed to meet specific challenges

associated with large, rapidly changing, or diverse data, often referred to collectively as “Big Data.” The four major types of NoSQL databases are “key-value,” “BigTable (or column family),” “document” and “graph” databases (McCreary and Kelly 2014; Sullivan 2015). Each type of database emphasizes a particular mix of characteristics that optimize it for particular tasks. For example, “document” databases facilitate queries on large quantities of unstructured or semi-structured data. In contrast, “key-value” databases emphasize rapid retrieval of data from distributed systems. “Graph” or “graph store” databases can encapsulate complex rules and relationships in data, whereas “BigTable” or “column family” databases store and access data in very large, sparse, tables using row and column identifiers.

The relational database has dominated traditional data handling, primarily in business contexts. Many relational databases incorporate ACID (Atomicity, Consistency, Isolation and Durability) principles (Haerder and Reuter 1983). Atomicity refers to the principle that each transaction should either be fully completed, or not at all. Thus, if the “city” and “state” in an address is being updated, and the computer system fails after “city” has been updated but before “state” has been updated, the entire transaction will fail, and “city” will be reset to its prior value. Consistency dictates that any validation rules must be met. Thus if a database contains a percent, and an attempt is made to set it to 101, the transaction should fail. Isolation dictates that if two transactions try to change the same piece of data at the same time, one of them will wait until the other has completed. Finally, durability dictates that the results of a transaction must be fully completed before the transaction terminates. This protects against failures where data may have been written to a disk cache, but not the disk itself when a power failure occurs.

Adherence to ACID principles means that a properly designed relational database is a very reliable place to store data. All sorts of potential errors are prevented. However, this comes with a cost in terms of flexibility and performance. If a sensor is generating thousands of measurements each second, a relational database may be unable to keep up, with each “insert” transaction taking longer to complete than the time between measurements. Similarly, if a large database is spread across many servers, and one of them fails or experiences network delays, transactions may be unable to be completed. Also, for science uses, validation rules may not be immutable. Our expectation that a temperature would be less than 30° may prove to be incorrect, so an ability to override validation limits may be needed.

NoSQL databases typically give up on one or more elements of ACID in order to increase performance and flexibility (McCreary and Kelly 2014; Leavitt 2010). For example, document stores don’t require the pre-definition of tables and database schema needed for relational databases, making them extremely flexible. But document stores also forgo most validation checks, violating the “consistency” constraint. Similarly, to increase speed, key-value databases often cache key values in memory, making them susceptible to losses of data in the event of a computer crash, violating the “durability” constraint. NoSQL databases often provide greatly improved speed, but may also allow temporary inconsistencies (McCreary and Kelly 2014). Thus, there is often the need to balance the needs for reliability against the needs for flexibility and performance when selecting the type of database to use.

Regardless of the type of database, a key feature is the ability of DBMS to interact with web servers. This makes possible dynamic web pages that immediately reflect changes to the database. Web pages can be used for both display and input, allowing users on the Internet to contribute data and metadata. Data and metadata entry systems using a web browser or app as the front-end are increasingly common. When the form is submitted, changes can be made in the database immediately, so users will have immediate access to the updated information.

### 3.4.4 Data Models and Normalization

In the creation of relational databases, the DBMS constitutes the canvas, but the data model is the painting. The purpose of a data model is to explicitly spell out the relationships between the different entities about which data is being stored (Bobak 1997). Ultimately the data model will be used as the road map for the definition of tables, objects, and relations. However, it is typically at a level of abstraction that lets us get past a mass of detail to look at the “big picture.” In a data model, the entities that will be represented in a database are defined and their attributes specified.

Normalization is a process wherein a data model is reduced to its essential elements (Hogan 1990). The aim of normalization is to eliminate redundancies and potential sources of inconsistency. During the normalization process, it is not unusual to define new entities and attributes or to eliminate old ones from a data model. Note that data modeling is in many ways, a creative process. Although there are rules for normalization, the data model inevitably reflects the purpose of the database and the thought process of its creator.

The data modeling process is best described through an example. Let’s look at a simple example of a database of species observations. The simplest model would be to store all the data in a single table (Table 3.3).

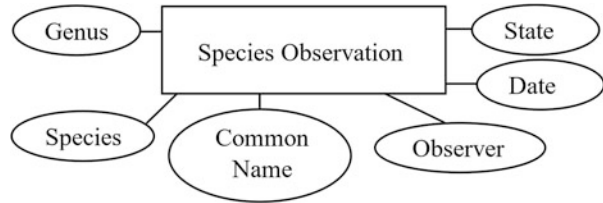
Figure 3.8 shows how this table would be represented in an entity-relationship diagram (E-R diagram).

There are several deficiencies in this model. First, the table is full of redundancies. The species *Quercus alba* is represented numerous times within the table, as is the common name “White Oak.” This gives many opportunities for errors to enter into the table. For example, in the third line, “White Oak” is misspelled “White Oat.”

**Table 3.3** “Flat file” species observation database

Genus	Species	Common name	Observer	Date	State
Quercus	alba	White Oak	Jones, D.	15-Jun-1998	NC
Quercus	alba	White Oak	Smith, D.	12-Jul-1935	VA
Quercus	alba	<i>White Oat</i>	Doe, J.	15-Sep-1920	PA
Quercus	rubra	Red Oak	Fisher, K.	15-Jun-1998	VA
Quercus	rubra	Red Oak	James, J.	15-Sep-1920	NC

**Fig. 3.8** In an entity-relationship (E-R) diagram tables are represented by boxes and columns in the table (fields, attributes) by ovals



**Table 3.4** Table for species entity

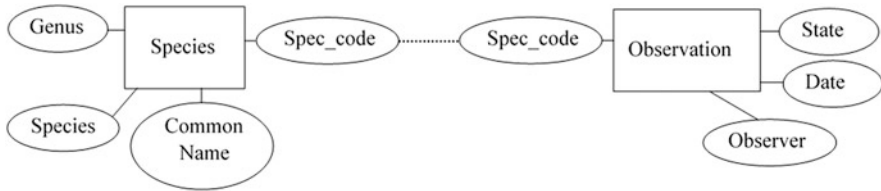
Spec_code	Genus	Species	Common name
1	Quercus	alba	White Oak
2	Quercus	rubra	Red Oak

**Table 3.5** Data table for observations entity

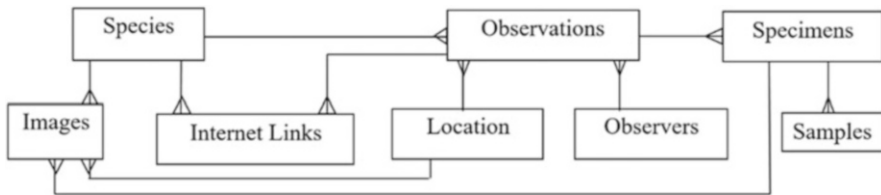
Spec_code	Observer	Date	State
1	Jones, D.	15-Jun-1998	NC
1	Smith, D.	12-Jul-1935	VA
1	Doe, J.	15-Sep-1920	PA
2	Fisher, K.	15-Jun-1998	VA
2	James, J.	15-Sep-1920	NC

A second option is to split our table into two entities, one representing the species-level data and another entity for the observations (Tables 3.4 and 3.5). We need to add an attribute (here called “Spec\_code”) that can act as a key to link the two entities together. This key attribute takes on any unique code, including text, however here we have chosen a numerical code. In Fig. 3.9 a dotted line is used to show that link. With this data model it is not possible to have the inconsistency (“White Oat”) in Table 3.3 because we have eliminated redundant copies of the Common Name.

In a real application, the species entity would incorporate all of the data relevant to the species that is independent of any specific observation. Additional attributes of the species entity might include additional taxonomic information (e.g., family, order), physical characteristics (e.g., mean size, branching pattern, leaf type) and natural history information (e.g., reproductive characteristics, habitat). We might even want to include images or Internet links. The observation entity might be expanded to include additional information on the observation (e.g., method of reporting, citation, voucher specimens), information on the location of the observation and additional details on the observer (e.g., contact information, such as address and email). As this process proceeds it may become evident that additional entities are required. For example, if a single observer makes multiple observations, it may make sense to establish an observer entity with attributes such as address, phone number, email, along with a new key attribute—Observer\_code. Similarly,



**Fig. 3.9** In the revised data model, the `Spec_code` field provides a link between observations in the two tables



**Fig. 3.10** An expanded data model for a database of species observations. One-to-many relationships are shown using a *split line* on the end connected to the table containing multiple rows of data linked to a single row of data in the associated table

we might want to add an additional entity that describes locations in detail, including coordinates, habitats etc. Linking images to specimens and locations as well as species could further extend the model. A more comprehensive (but by no means exhaustive) data model showing the entities (but not attributes) for a species observation database is shown in Fig. 3.10. Note that multiple lines connecting the entities depict a one-to-many relationship. Thus a species may have multiple observations, Internet links and images, but each observation, Internet link or image may be linked to only one species.

The degree to which formal normalization methods can be applied to NoSQL databases varies, although the primary principle, that each piece of data should be represented in a single place in the database, does not.

### 3.4.5 Advantages and Disadvantages of Using a DBMS

There are numerous advantages to using a DBMS. The first is that a DBMS has many useful built-in capabilities such as sorting, indexing and query functions (Maroses and Weiss 1982; Hogan 1990). Additionally, most relational databases include integrity and redundancy checks and support transaction processing with ACID characteristics. There has been substantial research into making relational DBMS as efficient as possible and many DBMS can operate either independently, or as part of a distributed network. This aids in scalability because if one computer starts to become overloaded, another can be added without having to substantially

restructure the underlying system. Finally, most relational DBMS include interfaces that allow linkage to user-written programs or other software, such as statistical packages. This is useful because it allows you to change the underlying structure of the data without having to alter programs that use the data.

Despite these advantages, most DBMS are designed to meet the needs of business applications and these may be quite different from the needs of scientists (Maroses and Weiss 1982; Pfaltz 1990). For example, most commercial DBMS have few graphical or statistical capabilities. DBMS are typically designed to create standardized reports. These may be of little use to researchers asking new questions. Additionally, DBMS are typically designed to deal with large volumes of data of a few specific types. They are less useful when dealing with relatively small volumes of data of many different types (the “variety” side of Big Data). If addition of each new type of data to a database requires creation of new data tables, the data model for the database can soon become incomprehensible. Similarly, DBMS can be relatively inefficient in dealing with sequential data (e.g., data ordered by time of collection). Some functions, such as highly optimized updating capabilities, are not frequently used for scientific data because, barring detection of an error, data is seldom changed once it is in the database, making some of the NoSQL alternatives to relational databases more attractive. Additionally, not all analysis tools can be easily interfaced with a DBMS. Proprietary data formats used by any DBMS may limit archival quality of data. A final disadvantage of DBMS is that they require expertise and resources to administer. This applies to all types of DBMS, including NoSQL databases. For large projects, the costs of administration may be easily absorbed, but for smaller projects or individuals, the resources required may exceed the benefits accrued by using a DBMS for managing data. However, even if a relational DBMS is not used for data, you may want to consider using a DBMS for metadata (documentation). The structure of metadata is frequently more complex than that of data and conforms better to the model of business data (relatively few types of data, standard reports are useful). Most data is located based on searching metadata rather than the data so the query capabilities of a DBMS are useful. Similarly, metadata is changed more often than data, so that the updating capabilities of a DBMS are more useful for metadata.

### 3.5 Interlinking Information Resources

Maximizing utility of database resources requires that we go beyond the simple creation of individual databases. Synthetic and integrative research approaches require the combination of data, often from diverse sources (Carpenter et al. 2009; Reichman et al. 2011). Users benefit from being able to search multiple databases via a single query. Similarly, the value of the data that is contained in an individual database is elevated when users are able to easily locate ancillary and related data found elsewhere. A frequent phenomenon that accompanies development of successful databases is that they spawn a series of “value added” databases which tailor the raw information contained in one or more “basic” databases to meet

the needs of a specific community. In our library and publishing analogy, we would not expect useful physical constants and formula only to be found in a single book. Instead we find them in a number of different reference texts aimed at different audiences. Similarly, we should not expect only one source or format for a specific kind of scientific data.

### ***3.5.1 A Database Related to the Human Genome Project***

The Human Genome Project provides an excellent case study of the opportunities (and pitfalls) inherent in linking databases together. The data from GenBank, EMBL and the Genome Data Base serve as “grist for the mill” of other databases. For example, the Ribosomal Database Project (RDP; Cole et al. 2014) harvests data on RNA data sequences from releases of GenBank. RDP then performs additional analyses to align sequences from different sources and develop phylogenetic diagrams. It also provides specialized tools for locating “probes” which may be used to distinguish classes of sequences. The RDP is then used by communities of ecological and health researchers to identify microbes, or in the case of unknown microbes to estimate the probable characteristics of such microbes based on their similarity to known microbes. Although it contains no raw data that is not available in sequence databases, RDP and similar databases reduce the duplication inherent in having each individual researcher analyze that raw data.

### ***3.5.2 Environmental Databases for Sharing Data***

One approach to sharing data would be for each researcher to post their own on their own web or social media page. This would certainly make data available, but at the same time poses many problems. It would be difficult to formulate a coherent strategy to search for similar data, the data would be highly diverse and the level of metadata or documentation would vary widely. Most serious is that many Internet resources are ephemeral—with web pages disappearing when new systems are introduced or researchers move or retire. For this reason, there has been growth in the development of repositories and clearinghouses for environmental data.

Government is a major generator of publicly-accessible environmental data collected for regulatory and informational purposes. Within the U.S. government there has been a trend towards consolidation of data resources. For example, the National Oceanic and Atmospheric Administration merged several individual data centers focusing on specific aspects of weather, climate, fisheries, oceans and geophysical data into the National Centers for Environmental Information (NOAA 2016). However, different agencies, with different missions and data needs, have led to a wide array of federal data systems.

Non-governmental data is increasingly being made available via a growing number of repositories (Michener et al. 2012). Some focus on a specialized type



of data. For example, Vegbank (Ecological Society of America 2016) is an archive primarily aimed at vegetation plot data (Peet et al. 2012). In contrast figshare (2016) allows contributors to upload a variety of files including figures, tables and data, regardless of the topic. Some restrict the source of data. The Dryad Data Repository (Dryad 2016; White et al. 2008) provides a repository for data associated with publications in selected journals.

Repositories vary widely in the types of data they allow and the amount of metadata they require. Some, such as figshare, have minimal requirements, with little or no metadata required and almost all types of files accepted, including those in proprietary formats. In contrast, the Long Term Ecological Research Network Data Portal (LTER 2016) uses EML metadata to describe tabular data, most of which are in generic, rather than proprietary formats. For recently collected data the difference between proprietary and generic formats may be unimportant—the software is available to read both of them. However, as the data ages, data stored in proprietary formats or with minimal metadata may become increasingly difficult to analyze or interpret.

In addition to data repositories that actually contain the data files, data clearing-houses or data registries, which provide links to data provided elsewhere, have also been popular. For example, at the level of the U.S. federal government [Data.gov](#) (2016) attempts to provide links to all the types of data collected by the federal government. The Ecological Society of America Data Registry contains links to data provided by society members, but it is up to the individual members to assure that the registered data becomes accessible. However, such registries and clearing-houses are difficult to keep current. Web sites frequently change, or disappear altogether, leaving the clearinghouse with non-functional links.

DataONE provides a middle ground (DataONE 2016; Michener et al. 2012). It provides searches that span a large number of environmental data providers, including Dryad, the Long-Term Ecological Research network, the Knowledge Network for Biocomplexity, the National Phenology Network and dozens of other data repositories. Individual data repositories or “member nodes” are responsible for providing and maintaining the content, but the search interface and links to data are standardized, providing a simplified user experience and facilitating comprehensive data searches. DataONE also provides Digital Object Identifiers (DOIs) similar to those used for conventional digital publications, and supports versioning, so that data used in an analysis can be accurately retrieved.

Digital Object Identifiers and similar identification systems, provide one solution to the problem of changes in the locations or ownership of information resources. They link to a database that can give the current location of a resource and if the location of a resource changes, the database can be edited to reflect that change. However, they remain dependent on providers updating the database to reflect new locations and keeping the material available on the network. Resources that disappear from the network or whose change in location are not noted can still be lost. Additionally, DOIs were designed with immutable documents in mind. However, many data are dynamic, with frequent additions. So although a DOI points to a dataset, the contents of the dataset itself may be different. For this reason, DOIs are

often coupled with other versioning mechanisms or DOIs are issued for specific versions of a dataset.

### 3.5.3 *Tools for Interlinking Information*

Systems for searching data, such as DataONE, provide only the first step in interlinking information. Once data is located it still needs to be acquired and integrated. The integration process is often complex, involving adjustment of sampling intervals (resampling, interpolation), aggregation, merging, unit conversions and extensive quality assurance checks. The process can be simplified for data that conform to standards (Kolb et al. 2013), but ultimately the process is often complex, iterative and esoteric. Such complex analyses are best supported by software that provide an auditable record of the steps used to synthesize multiple datasets in an analysis (Borer et al. 2009).

There are a large number of analytical tools used by ecologists ranging from spreadsheets, to database management systems, to statistical packages and finally to specialized user-written software. Spreadsheets are probably the most widely used, but are also problematic, with poor or marginal features for merging, aggregating and transforming data. Moreover, unless each click is recorded, it may be impossible to reconstruct what actually was done with data.

Better are statistical packages such as R, Matlab, SAS and SPSS, which allow text copies of analytical steps to be preserved, and re-run or modified as needed and support a wide array of functions for integrating data from different sources. The choice of a specific package can be driven by cost, user background, and specific analytical needs. However, there is such a wide range of overlap in capabilities, that any of the statistical packages can be usefully applied to most data integration challenges.

Database systems are designed to address integration of standardized data, and can be used, albeit with more difficulty, to integrate the diverse, often non-standard data generated by ecologists (Kolb et al. 2013). As with statistical packages, there is extensive overlap in the capabilities of different databases, and also with the data integration capabilities of spreadsheets. Ultimately it is the expertise of the user, rather than the characteristics of the statistical package or database system that dictate how successful a data integration effort will be.

## 3.6 Conclusions

Database management tools can be used by individual researchers and research projects to improve the quality of data, the speed and accuracy of retrieval and the ease of manipulation. However, these advantages do not come without a cost, in terms of the resources spent on database creation and administration. At the larger

scale, shared scientific databases are increasingly setting the boundaries for science itself. Taking ecological and environmental science to the next step will require taking ecological and environmental databases to a new level. A key to the success of scientific databases lays in developing incentives for individuals who collect data to make that data available in databases (Duke and Porter 2013; Porter and Callahan 1994; Roche et al. 2014). Additionally, mechanisms for funding databases need to be developed. For-profit databases have been successful in some areas (e.g., Chemical Abstracts), but are an unlikely candidate for success where the number of potential users is small (regardless of the importance of the data to our understanding of nature). Direct funding of databases has had some successes, but in many ways a successful database is a funding agency's worst nightmare: a project that grows year after year and never, never goes away! Technological innovations in computer systems and software can reduce the cost of operating data centers, but maintaining data in the face of entropy that attempts to disorganize that data is no simple task.

Despite these challenges, there are an increasing number of environmental databases of shared environmental data, such as Dryad and DataONE. These databases provide access to large numbers of individual datasets and are an essential first step. Researchers can't use data they can't access! However, it still falls on the researcher, once data has been acquired, to perform the many manipulations required to successfully integrate data from diverse sources. New approaches will be required to simplify the data integration process and to make it possible for researchers to easily access the data they want, in the form they want. Such approaches may involve tool development, wherein on-the-fly manipulations are performed, or "value added" databases ingest raw datasets but produce standardized and well-documented data products ready for researchers to use. Ideally, researchers should not be restricted to using tens of datasets (a practical maximum when each dataset requires substantial, individual, manipulation), but rather hundreds or even thousands of datasets.

Scientific databases evolve, but they don't spontaneously generate. We are at an exciting time in the development of scientific databases. Scientific questions and technological advances are coming together to make a revolution in the availability and usability of scientific data possible. However, the ultimate success of scientific databases will depend on the intelligence and commitment of individuals creating and operating databases.

**Acknowledgments** This chapter benefited immensely from conversations with information managers and scientists in the LTER network and DataONE, Robert Robbins, William K. Michener, Susan Stafford, Bruce P. Hayden, Dick Olson and John Pfaltz. It was supported by NSF grant DEB-1237733 to the University of Virginia.

## References

- Anderson C (2008) The end of theory: the data deluge makes the scientific method obsolete. *Wired* 16.07. <http://www.wired.com/2008/06/pb-theory/>. Accessed 15 Aug 2016
- Benson DA, Cavanaugh M, Clark K et al (2013) GenBank. *Nucleic Acids Res* 41:D36–D42
- Berman HM, Westbrook J, Feng Z et al (2000) The protein data bank. *Nucleic Acids Res* 28 (1):235–242
- Bobak AR (1997) *Data modeling and design for today's architectures*. Artech House, Norwood, MA
- Borer ET, Seabloom EW, Jones MB et al (2009) Some simple guidelines for effective data management. *Bull Ecol Soc Am* 90(2):205–214
- Brackett MH (1996) *The data warehouse challenge: taming data chaos*. Wiley, New York
- Campbell P (2009) Data's shameful neglect. *Nature* 461:145–145
- Carpenter SR, Armbrust EV, Arzberger PW et al (2009) Accelerate synthesis in ecology and environmental sciences. *Bioscience* 59(8):699–701
- Cinkosky MJ, Fickett JW, Gilna P et al (1991) Electronic data publishing and GenBank. *Science* 252(5010):1273–1277
- Cole JR, Wang Q, Fish JA et al (2014) Ribosomal Database Project: data and tools for high throughput rRNA analysis. *Nucleic Acids Res* 42:D633–D642
- Colwell RK (1997) Biota: the biodiversity database manager. <http://viceroi.eeb.uconn.edu/Biota/>. Accessed 15 Aug 2016
- Cook RB, Wei Y, Hook LA et al (2017) Preserve: protecting data for long-term use, Chapter 6. In: Recknagel F, Michener W (eds) *Ecological informatics. Data management and knowledge discovery*. Springer, Heidelberg
- Costello MJ (2009) Motivating online publication of data. *Bioscience* 59(5):418–427
- Data.gov (2016) The home of the U.S. government's open data. <https://www.data.gov>. Accessed 15 Aug 2016
- DataONE (2016) DataONE: data observation network for earth. <https://dataone.org>. Accessed 15 Aug 2016
- Dryad (2016) Dryad. <http://datadryad.org>. Accessed 15 Aug 2016
- Duke CS, Porter JH (2013) The ethics of data sharing and reuse in biology. *Bioscience* 63 (6):483–489
- Ecological Society of America (2016) VegBank. <http://vegbank.org/vegbank/index.jsp>. Accessed 15 Aug 2016
- Federal Geographic Data Committee (FGDC) (1994) Content standards for digital spatial metadata (June 8 draft). Federal Geographic Data Committee, Washington, DC. <http://geology.usgs.gov/tools/metadata/standard/940608.txt>. Accessed 15 Aug 2016
- Fegraus EH, Andelman S, Jones MB et al (2005) Maximizing the value of ecological data with structured metadata: an introduction to Ecological Metadata Language (EML) and principles for metadata creation. *Bull Ecol Soc Am* 86(3):158–168
- figshare (2016) figshare. <https://figshare.com>. Accessed 15 Aug 2016
- Gilbert W (1991) Towards a paradigm shift in biology. *Nature* 349:99
- Guenther R, McCallum S (2003) New metadata standards for digital resources: MODS and METS. *Bull Am Soc Inf Sci Technol* 29(2):12–15
- Haerder T, Reuter A (1983) Principles of transaction-oriented database recovery. *ACM Comput Surv* 15(4):287–317. doi:10.1145/289.291
- Hampton SE, Strasser CA, Tewksbury JJ et al (2013) Big data and the future of ecology. *Frontiers Ecol Env* 11(3):156–162
- Harford T (2014) Big data: a big mistake? *Significance* 11(5):14–19
- Hogan R (1990) *A practical guide to data base design*. Prentice Hall, Englewood Cliffs, NJ
- Holdren JP (2013) Increasing access to the results of federally funded scientific research. Memorandum, Office of Science and Technology Policy, Washington, DC. [https://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp\\_public\\_access\\_memo\\_2013.pdf](https://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf). Accessed 15 Aug 2016

- Justice C, Bailey G, Maiden M et al (1995) Recent data and information system initiatives for remotely sensed measurements of the land surface. *Remote Sens Environ* 51(1):235–244
- Keller W, Mitterbauer C, Wagner K (1998) Object-oriented data integration: running several generations of database technology in parallel. In: Chaudhri AB, Loomis M (eds) *Object databases in practice*. Prentice-Hall, New Jersey
- Kolb TL, Blukacz-Richards EA, Muir AM et al (2013) How to manage data to enhance their potential for synthesis, preservation, sharing, and reuse—a Great Lakes case study. *Fisheries* 38(2):52–64
- Leavitt N (2010) Will NoSQL databases live up to their promise? *Computer* 43(2):12–14
- Loomis MES, Chaudhri AB (1998) *Object databases in practice*. Prentice Hall, Upper Saddle River, NJ
- LTER (2016) LTER Network Data Portal. <https://portal.lternet.edu>. Accessed 15 Aug 2016
- Madden S (2012) From databases to big data. *IEEE Internet Comput* 16(3):4–6
- Magnuson JJ (1990) Long-term ecological research and the invisible present. *Bioscience* 40(7):495–501
- Maroses M, Weiss S (1982) Computer and software systems. In: Lauff G, Gorentz J (eds) *Data management at biological field stations*. WK Kellogg Biological Field Station, Hickory Corners, MI, pp 23–30
- Mayer-Schönberger V, Cukier K (2013) *Big data: a revolution that will transform how we live, work, and think*. Houghton Mifflin Harcourt, New York
- McCray AT, Gallagher ME (2001) Principles for digital library development. *Commun ACM* 44(5):48–54
- McCreary D, Kelly A (2014) *Making sense of NoSQL: a guide for managers and the rest of us*. Manning, Shelter Island, NY
- Meeson BW, Strebel DE (1998) The publication analogy: a conceptual framework for scientific information systems. *Remote Sens Rev* 16(4):255–292
- Michener WK (2017) Creating and managing metadata, Chapter 5. In: Recknagel F, Michener W (eds) *Ecological informatics. Data management and knowledge discovery*. Springer, Heidelberg
- Michener WK, Brunt JW, Helly JJ et al (1997) Nongeospatial metadata for the ecological sciences. *Ecol Appl* 7:330–342
- Michener WK, Porter J, Servilla M et al (2011) Long term ecological research and information management. *Ecol Inform* 6(1):13–24
- Michener WK, Allard S, Budden A et al (2012) Participatory design of DataONE—enabling cyberinfrastructure for the biological and environmental sciences. *Ecol Inform* 11:5–15
- NCBI (2016) GenBank and WGS statistics. <http://www.ncbi.nlm.nih.gov/genbank/statistics>. Accessed 15 Aug 2016
- NOAA (2016) National Oceanic and Atmospheric Administration National Centers for Environmental Information. <https://www.ncei.noaa.gov/>. Accessed 15 Aug 2016
- Nogueras-Iso J, Zarazaga-Soria FJ, Lacasta J et al (2004) Metadata standard interoperability: application in the geographic information domain. *Comput Environ Urban Syst* 28(6):611–634
- Parsons MA, Fox PA (2013) Is data publication the right metaphor? *Data Sci J* 12:WDS32–WDS46
- Peet RK, Lee MT, Jennings MD et al (2012) VegBank: a permanent, open-access archive for vegetation plot data. *Biodiv Ecol* 4:233–241
- Pfaltz J (1990) Differences between commercial and scientific data. In: French JC, Jones AK, Pfaltz JL (eds) *Report of the first invitational NSF workshop on scientific database management, technical report 90-21*. Department of Computer Science, University of Virginia
- Porter JH, Callahan JT (1994) Circumventing a dilemma: historical approaches to data sharing in ecological research. In: Michener WK, Stafford S, Brunt JW (eds) *Environmental information management and analysis: ecosystem to global scales*. Taylor and Francis, London, pp 193–203

- Porter JH, Hanson PC, Lin CC (2012) Staying afloat in the sensor data deluge. *Trends Ecol Evol* 27 (2):121–129
- Reichman OJ, Jones MB, Schildhauer MP (2011) Challenges and opportunities of open data in ecology. *Science* 331(6018):703–705. doi:10.1126/science.1197962
- Robbins RJ (1994) Biological databases: a new scientific literature. *Publ Res Q* 10:3–27
- Robbins RJ (1995) Information infrastructure. *IEEE Eng Med Biol Mag* 14(6):746–759
- Roche DG, Lanfear R, Binning SA et al (2014) Troubleshooting public data archiving: suggestions to increase participation. *PLoS Biol* 12(1):e1001779
- Specify Software Project (2016) Specify Collections Management Software (Version 6.6.04). <http://specifyx.specifysoftware.org>. Accessed 16 Aug 2016
- Star SL, Ruhleder K (1996) Steps toward an ecology of infrastructure: design and access for large information spaces. *Inf Syst Res* 7(1):111–134
- Strebel DE, Meeson BW, Nelson AK (1994) Scientific information systems: a conceptual framework. In: Michener WK, Stafford S, Brunt JW (eds) *Environmental information management*. Taylor and Francis, London, pp 59–85
- Strebel DE, Landis DR, Huemrich KF et al (1998) The FIFE data publication experiment. *J Atmos Sci* 55(7):1277–1283
- Sullivan D (2015) *NoSQL for mere mortals*. Addison-Wesley, Hoboken, NJ
- UniProt Consortium (2014) UniProt: a hub for protein information. *Nucleic Acids Res* 43:D204–D212
- White HC, Carrier S, Thompson A et al (2008) The Dryad data repository. In: *International conference on Dublin Core and metadata applications-metadata for semantic and social applications*, 22–26 September 2008, Berlin (DC-2008). Humboldt-Universität zu Berlin, Berlin
- Whitlock MC, McPeck MA, Rausher MD et al (2010) Data archiving. *Am Nat* 175:145–146
- Wieczorek J, Bloom D, Guralnick R et al (2012) Darwin Core: an evolving community-developed biodiversity data standard. *PLoS One* 7(1):e29715
- Yarmey L, Baker KS (2013) Towards standardization: a participatory framework for scientific standard-making. *Int J Digital Curation* 8(1):157–172

# Chapter 4

## Quality Assurance and Quality Control (QA/QC)

William K. Michener

**Abstract** This chapter introduces quality assurance processes and procedures that are employed to prevent data contamination from occurring and, secondly, quality control processes and procedures that are used to identify and deal with errors after they have been introduced. In addition, QA/QC activities are described that can be implemented throughout the entire data life cycle from data acquisition through analysis and preservation and general rules of thumb for promoting data quality are presented.

### 4.1 Introduction

Quality assurance and quality control refer to the procedures that are used to prevent errors from occurring and identifying and flagging those errors when they do occur. In the context of data, quality assurance (QA) is a set of processes and procedures that are employed to prevent or minimize data contamination (i.e. the introduction of errors into a data set or database). Quality control (QC) focuses on the data set or database after it has been created and includes processes that are designed to identify, flag and, sometimes, correct errors that have been introduced in the data product.

### 4.2 Quality Assurance

Data errors can be introduced by numerous sources. Humans can introduce errors unknowingly (e.g., insufficient training and experience), accidentally (e.g., spilled coffee) or purposely (e.g., mischief). In addition, instruments, sensors, computer and data networks, and other field and laboratory equipment can malfunction or cease to function for a variety of reasons. Such sources of data errors include natural

---

W.K. Michener (✉)  
University of New Mexico, Albuquerque, NM, USA  
e-mail: [william.michener@gmail.com](mailto:william.michener@gmail.com)

phenomenon (e.g., fire, wind, floods, treefall, animal activities, biofouling), electromagnetic interference and power problems (e.g., spikes, loss of power, brown-outs, battery failure), environmental factors (e.g., sand and dust, temperature spikes, moisture, freezing), as well as fatigue, corrosion, photodegradation, and the normal wear and tear and breakage that can affect sensors, instruments and their components, and computer and data networks (see Ganesan et al. 2004; Suri et al. 2006; Campbell et al. 2013).

Quality assurance is aimed at minimizing or preventing the introduction of errors in data. QA includes a wide array of proactive and preventative administrative and procedural processes. Several steps can be taken to minimize the introduction of errors by humans. First, hire or otherwise engage qualified individuals and provide them with adequate training. Many companies offer training in use of the instruments they sell and many citizen science programs provide online or in-the-field training and testing. It is always useful to perform trial runs in the field and laboratory before official data collection begins to verify that staff and students are comfortable with the data collection, processing and QA/QC procedures. Second, adopt community-accepted standardized data collection and analytical methods whenever possible as these methods are often well documented and are more frequently associated with online or in-person training programs (e.g., Patrick Center for Environmental Research 2002; American Public Health Association 2005). Third, accidents can be minimized through training (e.g., annual safety training) and following laboratory best practices; universities, corporations and laboratories often have developed standard laboratory operating procedures and may also employ a safety officer that can provide guidance. Fourth, mischievous acts can often be prevented or reduced by either adding security (e.g., fencing, security cameras, locks) or by camouflaging instruments and sensors in the field.

Manual data entry continues to be used to record many field observations as well as some laboratory measurements. For critical data, it is beneficial to have data entered by two independent data entry personnel and, then, compare the two data sets after they have been created; any differences can be compared with the field notes, audio recordings, or other original sources. In addition, many spreadsheet, data entry and database programs allow one to check the validity of data as it is being entered. For example, entering an invalid date, a value that exceeds a particular range, an invalid categorical value such as “D” for sex when only “F” and “M” are allowed, or other invalid or questionable data may generate a comment or pop-up window (e.g., “Data Exceed Instrument Range” or “Invalid Date Entered”). Use of such data filters is strongly encouraged since it is much easier to identify and correct an error at the source (e.g., data acquisition, data entry) than after-the-fact (e.g., during data analysis).

Sensor and instrument errors can be minimized by establishing a program for routine checks, maintenance, calibration and replacement of components susceptible to degradation and failure. The maintenance program may include stocking replacement parts on site and tracking and recording maintenance activities. Proper grounding, shielded cables and backup power supplies may also help reduce equipment and sensor downtime. For critical data, it may be necessary to install



three or more replicate sensors so that data collection continues when one of the sensors ceases to function satisfactorily (e.g., breakage, sensor drift). Likewise, it may also be useful to routinely send replicate samples to another laboratory for processing to validate that laboratory instrumentation and procedures are working properly and collecting high quality data. Last, some programs employ automated alerts that are sent to key personnel when problems are noticed (e.g., Shafer et al. 2000).

### 4.3 Quality Control

Data quality control measures are employed to identify and flag suspect values in data products after they have been generated. Values can be suspect for many reasons. First, recorded data values may not represent the actual measurements or observations that were made in the field or laboratory. Such errors may be due to writing down erroneous values in lab or field notebooks and mistyping data in spreadsheets or data entry forms (i.e., transcription errors), power loss or spikes, and malfunctioning, miscalibrated or improperly maintained instruments and sensors. Related errors include the insertion of duplicate records and the failure to record or properly code or account for missing values. Second, recorded values may be outliers in that they lie outside the distribution of most of the other values in the record. Values that are smaller than the 5th percentile or larger than the 95th percentile are often considered outliers or extreme values. Outliers may or may not be errors, but often warrant further scrutiny as they may signal the occurrence of extreme events, unusual variation in responses, or non-normal data distributions.

Quality control activities include data filtering and a variety of graphical and statistical approaches that are discussed below.

#### 4.3.1 *Data Filters*

Data filters were discussed in Sect. 4.2 because they can be used to prevent invalid data from being entered in the first place (i.e., quality assurance). Data filters can also be employed after the data set or database has been initially created to scan for suspect data and possible errors. Quality control data filters may, for example, check for duplicate records, missing values, repetition of identical values, range exceedance, date and time chronology, erratic changes in slope (e.g., data spikes), internal consistency (e.g., minimal water temperature is lower than maximum water temperature), and spatial consistency (e.g., recorded values at one sensor do not depart dramatically from values recorded at other sensors nearby) (Collins et al. 2006; Durre et al. 2010; Campbell et al. 2013).

Most common spreadsheet and database programs provide easy-to-use tools that allow for checks of duplicate records, range checks, valid dates and times, and

**Table 4.1** Quality assurance and quality control activities associated with different components of the research and data life cycle (plan/study design, collect/acquire data, assess data quality, describe (add metadata), preserve and backup data, integrate and analyze data) (also see Michener et al. 1997; Brunt 2000; Edwards 2000; Campbell et al. 2013; Michener 2017a, b; Porter 2017; Cook et al. 2017; Schildhauer 2017).

Research component and associated QA/QC activities	Description
Plan/Design study	
Design experiment or field study	Fully define the experimental design to be followed and include description in the metadata
Describe field and lab methods	Select and describe the methods and instruments used to acquire and process data (ideally, using well documented community standards)
Create/adopt lab notebook	Adopt an electronic (or paper) lab notebook where standard operating procedures and all field, laboratory and analytical activities are fully documented
Collect/Acquire data	
Create data dictionary	Define variables and measurement units; date-time formats; site and other variable codes; acceptable values for categorical variables (e.g., sex, color) and expected ranges for continuous variables
Establish data entry processes	Develop file naming convention; design and implement data entry forms/screens, Laboratory Information Management System, etc.
Train personnel	Train individuals that will be involved in data collection and data processing
Maintain field and lab instruments	Establish a maintenance schedule that includes testing all field and lab instruments, performing routine maintenance and calibration, and tracking and recording maintenance activities
Assess data quality	
Verify files and file names	Check that file names are appropriate (e.g., descriptive, consistent, non-proprietary) and properly versioned, and that file names and sizes (e.g., checksum values, size in bytes) are consistent with metadata
Verify data completeness	Check data tables, files, databases, etc. for inclusion of date, time, location, collector(s), data processor(s), measurement units, relational key indicators and other critical values as appropriate
Verify data are reasonable	Check that measured/recorded values fall within expected ranges and geographic coordinates, and that values of categorical values are appropriate (e.g., “M” and “F” okay, but not “Z” for sex)
Verify internal consistency	Check that relationships among variable values are appropriate (e.g., minimum air temperature is less than maximum air temperature)
Verify presence or absence of anomalies	Check repeated measurements for anomalous behavior (e.g., erratic spikes, continually decreasing slope) that may indicate sensor or instrument malfunctions

(continued)

**Table 4.1** (continued)

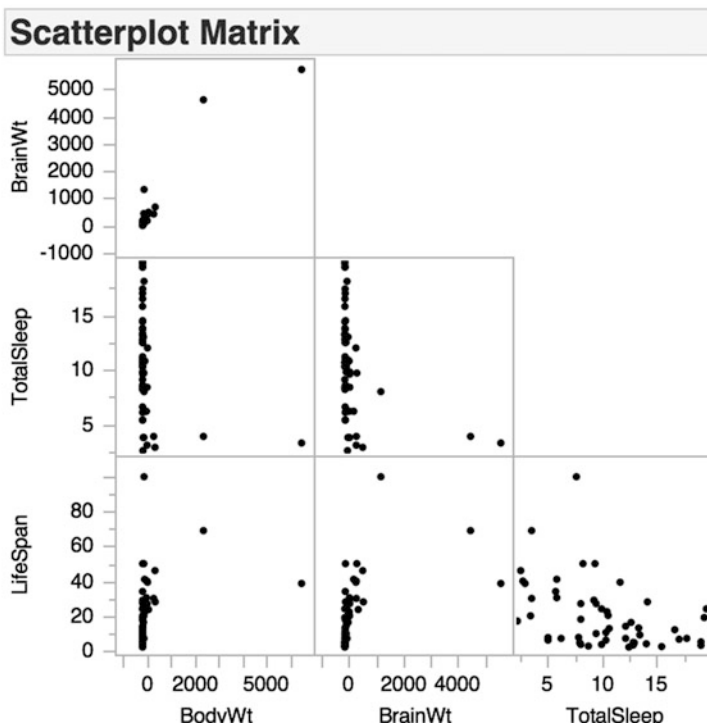
Research component and associated QA/QC activities	Description
Describe (add metadata)	
Verify data and metadata match	Check match between metadata and data to ensure that data are fully described and up-to-date with the metadata and vice versa
Verify files and file and data formats	Check that files and file and data formats are clearly defined and consistent
Verify methods and provenance	Check that all data collection and data processing steps are fully documented and that the provenance of data and derived products is clear and complete
Verify variables and units	Check that all variables and units of measurement are defined
Verify quality descriptors	Check that all data flags, issues and limitations associated with the data are fully described
Verify contextual information	Check that all information necessary to understand the study and the data is included (e.g., study objectives, study area description, lab and analytical standards)
Preserve and backup data	
Create and implement back up plan	Check that data are stored on at least three media (e.g., desktop computer, cloud storage, tape, auxillary hard drive) in at least two different locations (e.g., lab/office, off-site location such as a data center); routinely verify that backups have been performed and that data can be recovered from the backups
Create and implement preservation plan	Identify a repository where data will be preserved beyond the life of the project; follow the repository guidelines and recommendations for preparation of data and metadata products that will be preserved
Integrate and analyze data	
Verify data can be integrated and analyzed	Check that files, data tables and databases are complete, consistent with respect to units of measurement and inclusion of relational keys, and include sufficient metadata
Exploratory data analysis	Use graphical and statistical approaches to examine data distribution, highlight potential outliers, and identify other anomalies
Verify statistical assumptions	Check that data conform to assumptions underlying the statistical tests employed (e.g., data conform to normal distribution)

validity of categorical values. Otherwise, it is quite easy to add quality checks during the data analysis phase in most statistical packages, usually as a series of “if-then statements” (e.g., Edwards 2000, Table 4.1, p. 72; Gotelli and Ellison 2013, p. 216). Once suspect data are identified, it is important to have an established system for consistently flagging those values. The Oklahoma Mesonet, for instance, flags values as “good” (0), “suspect” (1), “warning” (2), or “failure” (3) (Shafer et al. 2000).

### 4.3.2 Graphical QC

There are numerous graphical approaches that can be used for quality control (e.g., identifying potential outliers) as well as exploring the data. Stem-and-leaf plots have long been used to visualize data and identify extreme data points in relation to the median and upper and lower quartiles (Gotelli and Ellison 2013); such plots can be easily constructed in R using the “stem” command (e.g., Horton and Kleinman 2011, p. 169; Gardener 2012, p. 63). These plots work well for small to medium-size data sets, but are often not as easy to interpret for very large data sets.

Scatterplot matrices are frequently used to initially explore data distributions and to identify potentially interesting relationships in the data (Gotelli and Ellison 2013), and are suitable for data sets of any size. Figure 4.1 illustrates the pairwise relationships among a subset of variables included in the classic data set on sleeping



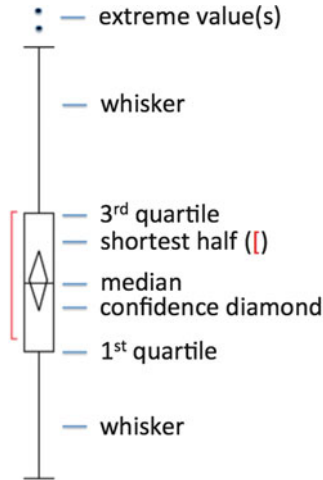
**Fig. 4.1** Scatterplot matrix illustrating the relationship between body weight (kg), brain weight (g), total sleep (h/day), and maximum life span (years) of mammals. Data are from Allison and Cicchetti (1976) and are also included as sample data sets in the R and JMP statistical programs. Note: the original data set also includes slow wave (“non-dreaming”) sleep, paradoxical (“dreaming”) sleep, gestation time, species, and three additional categorical variables (i.e., predation, exposure and danger indices). This scatterplot matrix was created using the “Graph—Scatterplot Matrix” command in JMP® Pro 11.0.0 (Copyright © 2013 SAS Institute Inc.; see SAS Institute Inc. 2013a)

mammals (Allison and Cicchetti 1976). This data set was used to examine the relationships among sleep habits and other characteristics of mammal species adapted to a variety of ecological niches. The scatterplots indicate that most of the data values for the variables (especially for Body Weight, Brain Weight, and Life Span) appear clustered near zero and that each variable has data values that lie at the extremes (i.e., African and Asian elephants have body weights of 6654 kg and 2547 kg and brain weights of 5712 g and 4603 g, respectively; and humans and Asian elephants have life spans of 100 years and 69 years, respectively). Such data distributions often indicate that the data are not normally distributed and that they should be transformed prior to formal statistical analysis (Edwards 2000). Gotelli and Ellison (2013) describe a diverse array of data transformations (i.e., logarithmic, square-root, cube-root, reciprocal, arcsine, and Box-Cox) that can be used so that the transformed data meet the assumptions of statistical tests (e.g., homoscedasticity and normality of residuals). Figure 4.1 was created using the statistical package JMP (Copyright © 2013 SAS Institute Inc.); similar matrices can be created using R (e.g., see page 167 (“pairs” function) in Horton and Kleinman 2011).

The box plot or box-and-whisker plot is particularly effective for visualizing data distributions and identifying possible outliers. Figure 4.2 illustrates the components of a box plot constructed using the statistical package JMP (Copyright © 2013 SAS Institute Inc.); similar box plots can be created in R (e.g., see page 169 (“boxplot” function) in Horton and Kleinman 2011). The box plot is notable for the information content that is condensed into a relatively simple diagram and the ease with which it can be used to rapidly assess the data distributions of many variables.

Box plots are often combined with histograms to aid in identifying extreme values (e.g., see page 213 in Gotelli and Ellison 2013); histograms can be created in R (e.g., see page 168 (“hist” function) in Horton and Kleinman 2011). Combined box plots and histograms are shown in Fig. 4.3 for the Body Weight, Brain Weight, and Life Span data from Allison and Cicchetti (1976) that were previously shown in the scatterplot matrix in Fig. 4.1. Figure 4.3 further illustrates the clustering of data values near zero and the presence of apparent outliers. In fact, the clustering is so significant that the box plots of body weight and brain weight are collapsed to the extent that they are un-interpretable. Conversely,  $\log_{10}$ -transformed body weight, brain weight and life span data approximate a normal distribution in the combined box plots and histograms illustrated in Fig. 4.4. For all three variables, the median falls within the confidence diamond for the mean, both the mean and median are included within the densest 50% of the observations (i.e., “shortest half” as defined by Rousseeuw and Leroy 1987), and no extreme values are indicated. A normal quantile plot can be used to further examine the extent to which a variable is normally distributed. In Fig. 4.5, the  $\log_{10}$ -transformed brain weight data fall approximately along a straight line in the normal quantile plot and all data values fall within the Lilliefors confidence bounds—conditions that are indicative of a normal distribution (Conover 1980).

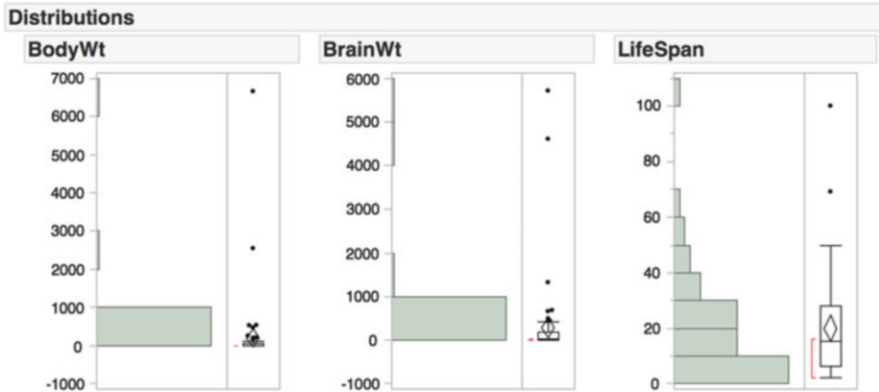
If we now plot the  $\log_{10}$ -transformed values of body weight versus  $\log_{10}$ -transformed values of brain weight, the relationship appears linear (Fig. 4.6),



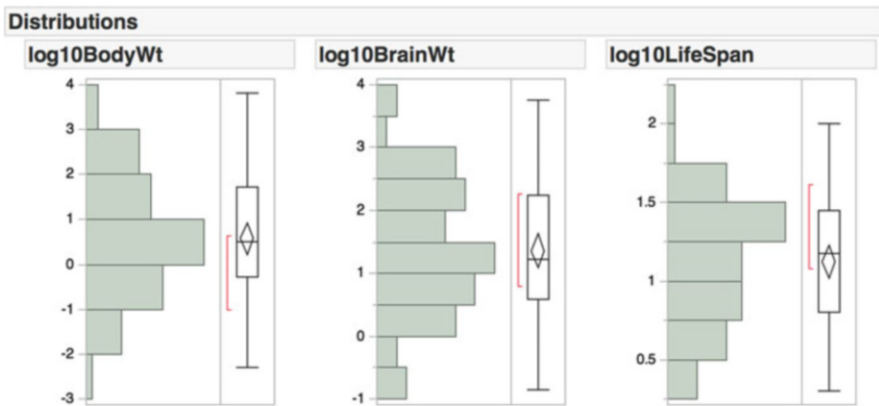
**Fig. 4.2** A typical box plot includes a *horizontal line* within the box (indicates the median data value), a box (encompasses half of the data values ranging from the 1st quartile through the 3rd quartile; note the difference between the 1st and 3rd quartiles is referred to as the interquartile range or IQR), whiskers (extending from the box to upper and lower values (calculated as: 3rd quartile + 1.5\*IQR and 1st quartile - 1.5\*IQR, or the upper and lower data point values (excluding extreme values) if the data points do not extend to the computed ranges), and *dots* or *asterisks* (indicating extreme values and potential outliers). This box plot also includes a confidence diamond that contains the mean (i.e. an imaginary line drawn through the center of the diamond—near where the median is located in this example—would show the mean) and the upper and lower 95% of the mean (i.e., top and bottom points of the diamond). Also, a bracket encompasses the shortest half, which is defined as the densest 50% of the observations (Rousseeu and Leroy 1987). The illustration is based on an Outlier Box Plot that was created using the “Analyze—Distribution” command in JMP ® Pro 11.0.0 (Copyright © 2013 SAS Institute Inc.; see SAS Institute Inc. 2013b)

which is in sharp contrast to the relationship exhibited in Fig. 4.1. Moreover, the two largest mammals by body weight (the African and Asian elephants) no longer appear to be outliers (refer to Fig. 4.1); i.e., they follow the overall linear trend and are located in the right uppermost corner of the figure. Likewise, if we plot the  $\log_{10}$ -transformed values of brain weight versus  $\log_{10}$ -transformed values of maximum life span, the relationship appears approximately linear (Fig. 4.7). As in Fig. 4.6, the two largest mammals by body weight (the African and Asian elephants) no longer appear to be outliers (refer to Fig. 4.1); i.e., they follow the overall linear trend and are located in the far right of the figure just below humans which have the longest life span. Interestingly, the Little Brown and Big Brown Bats now stand out because of their relatively long life spans in relation to their brain weights as compared to 56 other mammals for which data was available (Allison and Cicchetti 1976).

Parallel coordinate plots can also be used for identifying outliers, discontinuities and trends in the data (e.g., positive or negative correlations; Inselberg 1985, 1997;

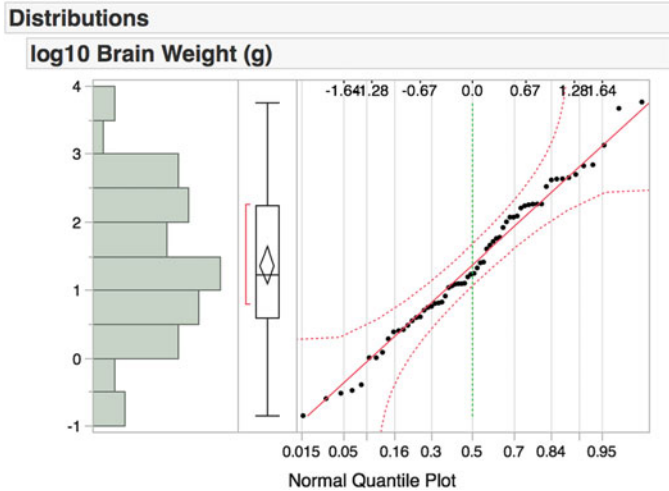


**Fig. 4.3** Combined histograms and box plots illustrating the distribution of body weight (kg), brain weight (g), and maximum life span (years) data values (data from Allison and Cicchetti 1976; see Fig. 4.1 for more information). The figure was created using the “Analyze—Distribution” command in JMP® Pro 11.0.0 (Copyright © 2013 SAS Institute Inc.; see SAS Institute Inc. 2013b)

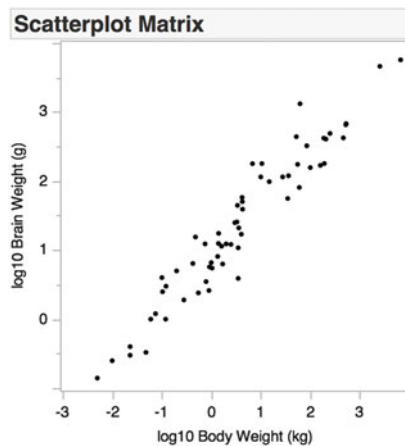


**Fig. 4.4** Combined histograms and box plots illustrating the  $\log_{10}$ -transformed distributions of body weight (kg), brain weight (g), and maximum life span (years) data values (data from Allison and Cicchetti 1976; see Fig. 4.1 for more information). The figure was created using the “Analyze—Distribution” command in JMP® Pro 11.0.0 (Copyright © 2013 SAS Institute Inc.; see SAS Institute Inc. 2013b)

Wegman 1990). A parallel coordinate plot allows one to visualize each cell in a data table, and a single line segment represents a row in a data table. In Fig. 4.8, for example, the horizontal line in the upper left corner reflects a discontinuity in the data as “nondreaming” and “dreaming” sleep were not measured for the African elephant although “total sleep,” “life span,” and “gestation” were measured. In contrast, the red line immediately below the African elephant represents the Asian elephant for which all variables were measured. Humans also stand out in the figure for their long life span (blue point at top of chart for LifeSpan variable). Overall, it

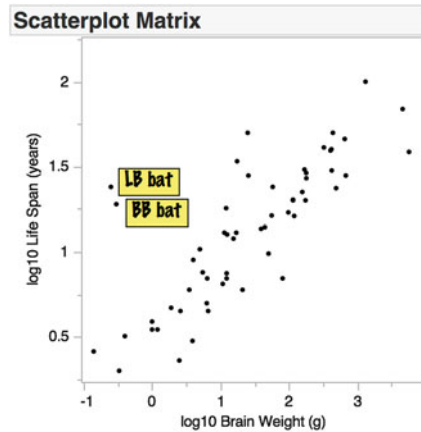


**Fig. 4.5** Combined histogram, box plot and normal quartile plot illustrating the  $\log_{10}$ -transformed distributions of brain weight data values (data from Allison and Cicchetti 1976; see Fig. 4.1 for more information). The normal quartile plot (on the *right*) includes the normal quartile scale (*top*), probability scale (*bottom*), and the Lilliefors confidence bounds (see Conover 1980 and SAS Institute Inc. 2013b) in red dots and a diagonal straight line through the data to aid in interpretation. The figure was created using the “Analyze—Distribution” command and the “Normal Quartile Plot” option in JMP <sup>®</sup> Pro 11.0.0 (Copyright © 2013 SAS Institute Inc.; see SAS Institute Inc. 2013b)

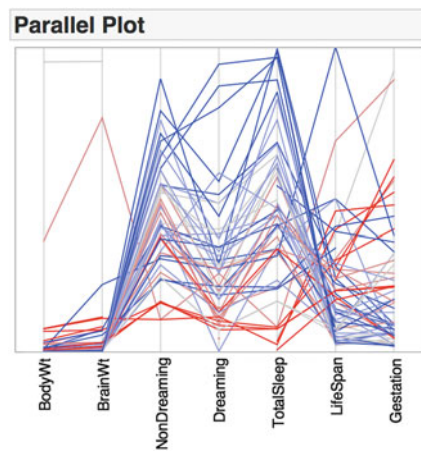


**Fig. 4.6** Scatterplot matrix illustrating the relationship between  $\log_{10}$ -transformed body weight (kg) and  $\log_{10}$ -transformed brain weight (Data from Allison and Cicchetti 1976). This scatterplot matrix was created using the “Graph—Scatterplot Matrix” command in JMP <sup>®</sup> Pro 11.0.0 (Copyright © 2013 SAS Institute Inc.; see SAS Institute Inc. 2013a)





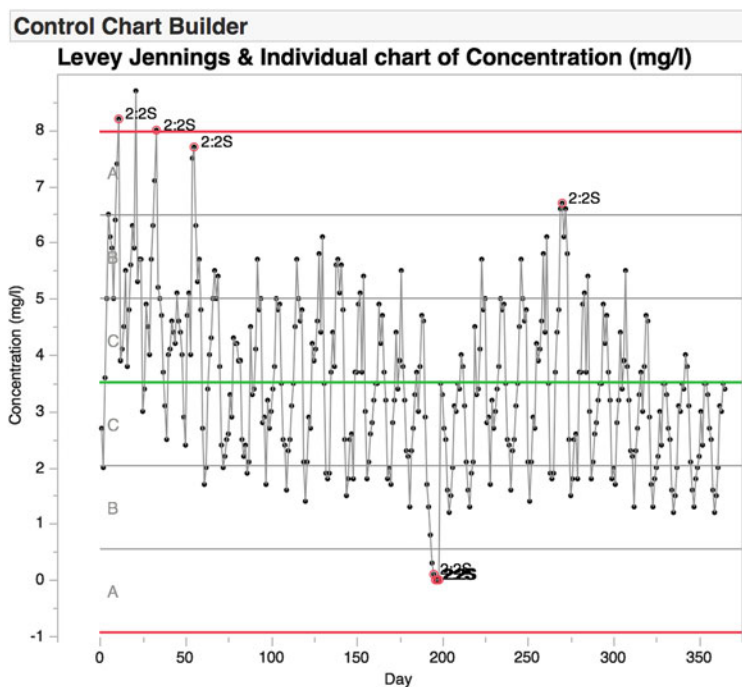
**Fig. 4.7** Scatterplot matrix illustrating the relationship between  $\log_{10}$ -transformed brain weight (g), and  $\log_{10}$ -transformed maximum life span (years) (data from Allison and Cicchetti 1976). This scatterplot matrix was created using the “Graph—Scatterplot Matrix” command in JMP® Pro 11.0.0 (Copyright © 2013 SAS Institute Inc.; see SAS Institute Inc. 2013a)



**Fig. 4.8** Parallel coordinate plot showing cells (raw data) from the mammalian sleep data table created by Allison and Cicchetti (1976). A line represents a row from the data table and is continuous from left to right, except where data are missing. Blue lines represent those mammals that are least exposed and that are least susceptible to predation (e.g., Giant armadillo, which scores “1” on each of the predation, exposure and danger indices) whereas red lines represent the converse (e.g., rabbit, which scores “5” on each of the three indices). This parallel plot was created using the “Graph—Parallel Plot” command in JMP® Pro 11.0.0 (Copyright © 2013 SAS Institute Inc.; see SAS Institute Inc. 2013a)

also appears that those mammals that are categorized as most susceptible to predation (i.e., danger category 5; represented by the bright red lines) tend to spend less time in nondreaming, dreaming and total sleep.

Statistical process control charts (or control charts or Shewhart charts, named after Walter A. Shewhart) have long been used to assess and improve the quality of manufacturing and industrial processes by visualizing how a process changes over time (Montgomery 2008). Control charts include a central line that represents the average or mean of the variable and upper and lower lines for the control limits (e.g., 3 standard deviations of the mean). Control limits can be based on long-term historical data or upon the variability exhibited in a group of data (e.g., a month or year or sampling period). A continuous line connects the data values or, in some cases, the moving average of the process measurements. In industry, control charts are used for identifying and correcting problems as they occur, assessing process stability and identifying non-routine variation. In addition to charting measured values over time, “special cases rules” can be invoked to highlight non-routine variation (e.g., Nelson Rules and Westgard Rules; Nelson 1984, 1985; SAS Institute Inc. 2013c). Figure 4.9 illustrates how a control chart can aid in visualizing environmental data patterns and processes over time. In this case, the hypothetical concentration of some constituent, element or pollutant is measured daily for one



**Fig. 4.9** This quality control plot was created using the “Analyze—Quality and Process—Control Chart Builder” command (with options set to “Sigma: Levey Jennings” and “Warnings: Westgard Rules: Rule 2 2S”) in JMP® Pro 11.0.0 (Copyright © 2013 SAS Institute Inc.; see SAS Institute Inc. 2013c). Red lines indicate 3 standard deviations from the mean (green line); Zones C, B and A include points within 1, 2 and 3 standard deviations of the mean, respectively. Westgard Rule 2 2S is triggered (red highlighted points) when two consecutive measurements are greater than 2 standard deviations of the mean (i.e., at days 11, 33, 55, 195–198, 270)

year. On three occasions (days 11, 21 and 33) the measured concentration exceeded the upper limit (i.e., three standard deviations of the mean); the lower limit was never exceeded. In this example Westgard Rule 2: 2S was triggered four times (days 11, 33, 55, 270) when two consecutive measurements positively exceeded two standard deviations of the mean, as well as four times (days 195–198) when two consecutive measurements fell two standard deviations below the mean. All extreme values may warrant further examination to determine if some unusual event has occurred. In this case, the measurements clustered around and preceding days 195–198 may be indicative of instrument or sensor failure since this was the only period during the entire year when measured concentrations approached zero.

A variety of similar charts can be created to aid in interpreting trends and identifying extreme values. For example, instead of or in addition to including upper and limits based on the historic or observed variability in the data, one could include known limits based on illegal values such as wind speeds that are below zero or that exceed the limits of the instrument or sensor. When long-term data are available as is often the case with multi-decadal meteorological data, one approach is to plot the variable's daily minima and maxima for the year based on the historic data; thus, any value(s) that approached or extended outside this range might be flagged for further examination.

### ***4.3.3 Statistical QC***

In addition to the graphical methods presented in Sect. 4.3.2, it may also be useful to test statistical assumptions using formal statistical methods. Some univariate normality tests include: (1) Anderson-Darling test (Anderson and Darling 1954; Horton and Kleinman 2011; Razali and Wah 2011); (2) Kolomogorov-Smirnov test (Edwards 2000; Razali and Wah 2011; Gotelli and Ellison 2013); (3) Lillifors test (Conover 1999; Horton and Kleinman 2011; Razali and Wah 2011); and Shapiro-Wilk test (Shapiro and Wilk 1965; Razali and Wah 2011). The Grubbs test is a formal test for identifying outliers (Grubbs 1969; Grubbs and Beck 1972; Edwards 2000). Comprehensive statistical QC and data analysis are outside the purview of this chapter, but the importance of testing the assumptions of the statistical methods that are employed cannot be over-emphasized (e.g., see Gotelli and Ellison 2013).

### ***4.3.4 Treatment of Errors and Outliers***

A data value is only an error when it is definitively known that the value is incorrect. Outliers and extreme values are not necessarily errors and should never be deleted. Both errors and outliers can be extremely informative. An error provides valuable information that can potentially be used to improve quality control processes. Gotelli and Ellison (2013, p. 213) argue that

There are only two reasons for justifiably discarding data: (1) the data are in error (e.g., they were entered incorrectly in the field notebook); or (2) the data no longer represent valid observations from your original sample space (e.g., one of your dune plots was overrun by all-terrain vehicles). Deleting observations simply because they are “messy” is laundering or doctoring the results and legitimately could be considered scientific fraud.

When data are determined to be in error, approaches may sometimes be employed to correct for those errors (e.g., sensor drift; Horsburgh et al. 2010).

Outliers, on the other hand, may represent normal responses to extreme events or extreme responses to normal conditions; in either case, outliers can lead to new knowledge about the patterns or processes under study. Outliers should never be deleted unless they can be shown to represent contaminated data. Edwards (2000) suggests that

If no explanations for a severe outlier can be found, one approach is to formally analyse the data both with and without the outlier(s) and see if conclusions are qualitatively different.

## 4.4 Implementing QA/QC

Decisions and actions taken throughout the entire data and research life cycles (from study and data design through collection, adding metadata, preservation and backup, and integration and analysis) can affect data quality (Table 4.1). Some general rules of thumb for promoting quality data include: (1) plan for QA/QC from the start of the project (i.e., don’t wait until you first detect contaminated data); (2) all project personnel have a stake in data quality so engage them early on in establishing standard operating procedures in the field and lab, training, and quality assessment; and (3) routinely review the data and the QA/QC protocols as part of an ongoing process by all project personnel to improve data quality (Edwards 2000). All processes followed in data acquisition, QA/QC, and analyzing project data should ideally be maintained as part of an audit trail that is included in the metadata (Gotelli and Ellison 2013). Audit trails may be consulted in order to replicate project findings or as part of legal proceedings. Scientific workflow programs such as Kepler can be used to simplify and automate the process of capturing all data processing and analysis steps (e.g., Barseghian et al. 2010).

## 4.5 Conclusion

QA/QC procedures encompass all activities aimed at ideally preventing data contamination from occurring in the first place and, failing that, identifying and dealing appropriately with errors that are introduced in a data set. QA activities such as instituting comprehensive training, adopting community standards and following best practices, employing data filters to prevent the entry of illegal

values, establishing a rigorous maintenance program, and supporting independent, manual double-entry of data can significantly reduce or prevent data contamination.

Perfect data sets are rare. QC steps such as data filters and graphical and statistical procedures can be very effective at identifying and flagging outliers and errors. Virtually every research project can benefit by visually examining the data using tools such as scatterplot matrices, box plots, parallel plots and statistical process control charts. These graphical QC procedures allow one to explore the data and identify trends, search for outliers and unusual observations, and flag potential errors for further examination. A good understanding of the data including underlying trends, distribution of values, and unusual observations contributes to sound analysis and interpretation.

Data quality is affected throughout every step of the research and data life cycles from planning and data collection through analysis and dissemination of results. Quality data requires the attention of every individual that is engaged in the project from the researcher(s) that plan the project, to the student who collects samples in the field, to the post doctoral associate or technician who processes the samples in the laboratory. Good QA/QC programs address all phases of the research and data life cycles, engage all relevant personnel, and are routinely reviewed and improved upon.

## References

- Allison T, Cicchetti DV (1976) Sleep in mammals: ecological and constitutional correlates. *Science* 194:732–734
- American Public Health Association (2005) Standard methods for the examination of water and wastewater, 21st edn. American Public Health Association, Washington, DC
- Anderson TW, Darling DA (1954) A test of goodness-of-fit. *J Am Stat Assoc* 49:765–769
- Barseghian D, Altintas I, Jones MB et al (2010) Workflows and extensions to the Kepler scientific workflow system to support environmental sensor data access and analysis. *Ecol Inf* 5:42–50
- Brunt JW (2000) Data management principles, implementation and administration. In: Michener WK, Brunt JW (eds) *Ecological data*. Blackwell Science, Oxford, pp 25–47
- Campbell JL, Rustad LE, Porter JH et al (2013) Quantity is nothing without quality: automated QA/QC for streaming environmental sensor data. *BioSci* 63:574–585
- Collins SL, Bettencourt LMA, Hagberg A et al (2006) New opportunities in ecological sensing using wireless sensor networks. *Front Ecol Environ* 4:402–407. [http://dx.doi.org/10.1890/1540-9295\(2006\)4\[402:NOIESU\]2.0.CO;2](http://dx.doi.org/10.1890/1540-9295(2006)4[402:NOIESU]2.0.CO;2)
- Conover WJ (1980) *Practical nonparametric statistics*. Wiley, New York
- Conover WJ (1999) *Practical nonparametric statistics*, 3rd edn. Wiley, New York
- Cook RB, Wei Y, Hook LA et al (2017) Preserve: protecting data for long-term use, Chapter 6. In: Recknagel F, Michener W (eds) *Ecological informatics. Data management and knowledge discovery*. Springer, Heidelberg
- Durre I, Menne MJ, Gleason BE et al (2010) Comprehensive automated quality assurance of daily surface observations. *J Appl Meteorol Climatol* 49:1615–1633
- Edwards D (2000) Data quality assurance. In: Michener WK, Brunt JW (eds) *Ecological data*. Blackwell Science, Oxford, pp 70–91
- Ganesan D, Cerpa A, Ye W et al (2004) Networking issues in wireless sensor networks. *J Parallel Distrib Comp* 64:799–814

- Gardener M (2012) *Statistics for ecologists using R and Excel*. Pelagic Publishing, Exeter
- Gotelli NJ, Ellison AM (2013) *A primer of ecological statistics*. Sinauer Associates, Sunderland, MA
- Grubbs F (1969) Procedures for detecting outlying observations in samples. *Technometrics* 11:1–21
- Grubbs FE, Beck G (1972) Extension of sample sizes and percentage points for significance tests of outlying observations. *Technometrics* 14:847–854
- Horsburgh JS, Jones AS, Stevens DK, Tarboton DG, Mesner NO (2010) A sensor network for high frequency estimation of water quality constituent fluxes using surrogates. *Env Mod Software* 25:1031–1044
- Horton NJ, Kleinman K (2011) *Using R for data management, statistical analysis, and graphics*. CRC Press, Boca Raton, FL
- Inselberg A (1985) The plane with parallel coordinates. *Vis Comput* 1:69–91
- Inselberg A (1997) Multidimensional detective. In: *Proceedings of the 1997 IEEE symposium on information visualization*, pp 100–107
- Michener WK (2017a) Project data management planning, Chapter 2. In: Recknagel F, Michener W (eds) *Ecological informatics. Data management and knowledge discovery*. Springer, Heidelberg
- Michener WK (2017b) Creating and managing metadata, Chapter 5. In: Recknagel F, Michener W (eds) *Ecological informatics. Data management and knowledge discovery*. Springer, Heidelberg
- Michener WK, Brunt JW, Helly J et al (1997) Nongeospatial metadata for the ecological sciences. *Ecol Appl* 7:330–342
- Montgomery DC (2008) *Introduction to statistical quality control*, 6th edn. Wiley, New York
- Nelson L (1984) The Shewhart control chart – tests for special causes. *J Qual Technol* 15:237–239
- Nelson L (1985) Interpreting Shewhart x control charts. *J Qual Technol* 17:114–116
- Patrick Center for Environmental Research (2002) *Protocols for the analysis of algal samples collected as part of the US Geological Survey National Water-Quality Assessment Program*. Academy of Natural Sciences, Philadelphia, PA
- Porter JH (2017) Scientific databases for environmental research, Chapter 3. In: Recknagel F, Michener W (eds) *Ecological informatics. Data management and knowledge discovery*. Springer, Heidelberg
- Razali NM, Wah YB (2011) Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. *J Stat Model Analytics* 2:21–33
- Rousseeuw PJ, Leroy AM (1987) *Robust regression and outlier detection*. Wiley, New York
- SAS Institute Inc. (2013a) *JMP® 11 Essential graphing*. SAS Institute Inc., Cary, NC
- SAS Institute Inc. (2013b) *JMP® 11 Basic analysis*. SAS Institute Inc., Cary, NC
- SAS Institute Inc. (2013c) *JMP® 11 Quality and process methods*. SAS Institute Inc., Cary, NC
- Schildhauer M (2017) Data integration: principles and practice, Chapter 8. In: Recknagel F, Michener W (eds) *Ecological informatics. Data management and knowledge discovery*. Springer, Heidelberg
- Shafer MA, Fiebrich CA, Arndt DS et al (2000) Quality assurance procedures in the Oklahoma Mesonet. *J Atmos Ocean Technol* 17:474–494. doi:[10.1175/1520-0426\(2000\)017<0474:QAPITO>2.0.CO;2](https://doi.org/10.1175/1520-0426(2000)017<0474:QAPITO>2.0.CO;2)
- Shapiro SS, Wilk MB (1965) An analysis of variance test for normality (complete samples). *Biometrika* 52:591–611
- Suri A, Iyengar S, Cho E (2006) Ecoinformatics using wireless sensor networks: an overview. *Ecol Inf* 1:287–293
- Wegman EJ (1990) Hyperdimensional data analysis using parallel coordinates. *J Am Stat Assoc* 85:664–675

# Chapter 5

## Creating and Managing Metadata

William K. Michener

**Abstract** This chapter introduces the reader to metadata as well as the standards and tools that can be used to generate and manage standardized metadata during a research project. First, metadata is defined and many of the benefits that accrue from creating comprehensive metadata are listed. Second, the different types of metadata that may be necessary to understand and use (e.g., analyze, visualize) a data set are described along with some relevant examples. Third, the content, characteristics, similarities and differences among many of the relevant ecological metadata standards are presented. Fourth, the various software tools that enable one to create metadata are described and best practices for creating and managing metadata are recommended.

### 5.1 Introduction

Metadata refers to the information that is used to describe a dataset (i.e., data about data). The term “metadata” is synonymous with “documentation” and has only relatively recently become part of the science vocabulary. Metadata typically includes the information that is necessary to understand the origin, organization and characteristics of a data set. Simply put, metadata describes the: who, what, when, where, why and how of the dataset (Table 5.1). Such information is necessary for one to discover, acquire, understand, and use the data.

Despite the obvious importance of many of the questions included in Table 5.1, many researchers fail to sufficiently document their data. This inattention to metadata may be due to: (1) perceived lack of time, financial resources and personnel to effectively manage the data; (2) lack of awareness of metadata standards, metadata management tools, and best practices; (3) a belief that the effort is wasted because no one else would want to use the data; or (4) an assumption that the data originator(s) will remember the pertinent metadata.

The first challenge (i.e., lack of resources) can be overcome if the researcher devotes modest effort to initially creating a reasonable data management plan that

---

W.K. Michener (✉)  
University of New Mexico, Albuquerque, NM, USA  
e-mail: [william.michener@gmail.com](mailto:william.michener@gmail.com)

**Table 5.1** Categories of questions that can be asked about a dataset with examples

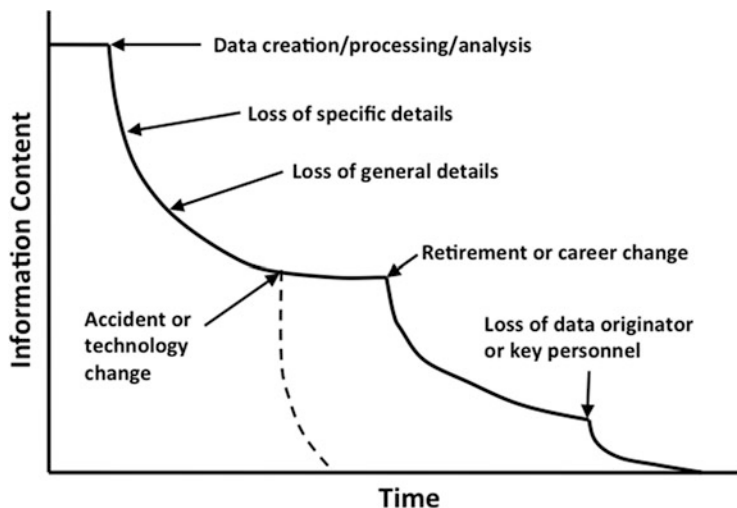
Questions	Examples
Who	Who collected the data? Who processed the data? Who funded the research that led to the data? Who is the primary contact for the data? Who is permitted to use the data?
What	What data were collected? What questions or hypotheses were being addressed? What QA/QC procedures were used? What data gaps exist? What do the variable labels mean? What do the codes mean? What limitations do the data have? What software do I need to read the data?
When	When were the data collected? When were the data processed and analyzed? When were or will the data be made available?
Where	Where were the data collected? Where were the data processed and analyzed? Where are the data stored?
Why	Why were the data collected? Why were specific collection, processing and analytical methods employed?
How	How were the data collected and processed? How precisely and accurately were the data values measured? How can the data be accessed? How are the data organized and structured? How should the data be cited?

includes addressing staffing and other budgetary needs necessary to support meta-data generation and management (see Michener 2017). The second challenge (lack of awareness of enabling tools and approaches) can be tackled through educational workshops and seminars or by reading the remainder of this chapter. Third, one might question whether the data should have been collected in the first place if it is virtually certain that no one else would ever want to see or use the data. Fourth, virtually everyone overestimates their ability to retain facts, figures and details.

There is a natural tendency for the information content of data sets to decrease over time. This loss of information content has been referred to as data entropy (Michener et al. 1997) and is illustrated in Fig. 5.1. The figure highlights the reality that researchers are generally most familiar with the data at the time of data collection and processing. Shortly after the data are analyzed, though, we often begin to forget many of the details related to data collection and processing such as the precise location where samples were collected and the underlying details of the laboratory and analytical methods. Later, more general details may be forgotten. Data and documentation can be lost through accidents such as crashed disks, electrical fires, and water leaks that can occur at any time. Finally, many of the details that are carried around in our heads are lost when personnel retire, leave a project or institution, or die.

The principal recourse for data entropy is to document your data as comprehensively as possible and to preserve the data and metadata in multiple locations (see Cook et al. 2017). There are many benefits that accrue from comprehensively documenting a project's data. First, you and your colleagues can easily interpret, use and re-use the data over time by having a record of all those details that are easily lost or forgotten. Second, the metadata can be useful for training or as a resource for new personnel or students that are brought on to the project. Third, your mind and memory can be used for more important tasks than remembering





**Fig. 5.1** Data and information entropy: factors contributing to the loss of the information content in a dataset over time (from Michener et al. 1997). Without proper curation, the investigator and the project members forget the specific and general details. Over time as investigators move to new projects and careers, the information about data products is lost. Furthermore, without proper backups, the data files will be lost

those project and data details that should be included in the metadata. Fourth, the metadata that you create and maintain can be easily incorporated into literature publications (i.e., materials and methods) and data packages (i.e., data and metadata) that are submitted to data repositories for long-term storage and, potentially, data sharing.

## 5.2 Metadata Descriptors

Metadata descriptors or elements are those attributes of a data set that are necessary for an individual to understand in order to discover, determine fitness-for-use, acquire, and use a data set. Metadata descriptors may be grouped into three broad categories: (1) high-level resource descriptors that provide basic information about the dataset; (2) contextual descriptors that describe the research context and procedures employed to collect, process, and manage the data; and (3) physical descriptors that detail the structure of the files and variables. Table 5.2 provides a comprehensive list of metadata descriptors that are grouped into these three categories. The descriptors represent a more structured way of describing the: who, what, when, where, why and how of the data as outlined in Table 5.1. Some of the descriptors are essential in order to acquire and understand the data; other descriptors may be optional but, nevertheless, provide additional information that may facilitate determination of fitness-for-use.

**Table 5.2** Metadata descriptors or elements that provide the basis for comprehensively describing ecological data by major category (derived from DCMI 2016, GBIF 2011a, ISO 2016a, Michener et al. 1997)

Metadata descriptor	Definition
<b>I. High-level Resource</b>	
A. Dataset title	Name assigned to the dataset
B. Identifier	Unique code (e.g., Digital Object Identifier) assigned by data originator or data repository to specifically identify a dataset
C. Responsible party (for data generation)	
1. <i>Creator</i>	Name(s) and address(es) of the individuals responsible for generating the dataset
2. <i>Contributor</i>	Name(s) and address(es) of the individuals responsible for contributing to the dataset (e.g., metadata creation)
D. Publication date	Date the dataset was published or made accessible (e.g., via a data repository)
E. Dataset description	
1. <i>Abstract</i>	Brief summary of dataset contents including general spatial, temporal, and taxonomic coverage
2. <i>Subject</i>	Theme of the dataset
3. <i>Keywords</i>	Subset of words and phrases that describe the dataset
F. Language	Language that the data and metadata are written in
G. Accessibility	
1. <i>Publisher or data repository</i>	Location where the data are stored
2. <i>Contact information</i>	Name(s) and address(es) of the data repository or other location(s) where the data are stored (i.e., where the data can be accessed)
3. <i>Copyright restrictions</i>	Any copyright restrictions that apply to all or portions of the dataset
4. <i>Other restrictions</i>	Any other restrictions or constraints that prevent use of all or portions of the dataset (e.g., human subjects data)
H. Citation	Recommended format for citing the dataset
<b>II. Research context</b>	
A. Project	
1. <i>Spatial coverage</i>	Latitude and longitude of sampling points or bounding box, geographic place names
2. <i>Temporal coverage</i>	Start and end date of study, sampling frequency
3. <i>Taxonomic coverage</i>	Taxonomic groups or species that are included in study
4. <i>Hypotheses/questions</i>	Questions or hypotheses that are being addressed
5. <i>Funding source(s)</i>	Research sponsor(s) including grant and contract numbers
6. <i>Project personnel</i>	Researchers, technicians and students involved in project
B. Study site	
1. <i>Location(s)</i>	Specific locations (latitude and longitude) of sampling points, how sites are marked and located in the field (e.g., in relation to reference points or landmarks), height and depth (as appropriate)

(continued)

**Table 5.2** (continued)

Metadata descriptor	Definition
2. <i>Site characteristics</i>	Description of site including (as appropriate) climate, vegetation, landform and topography, watershed, geology, lithology, soils, history of land use
C. Experimental/sampling design	Description of experimental/statistical design including details related to sampling, subsampling, replication, experimental treatments and controls
D. Research methods	
1. <i>Field and laboratory</i>	Description and/or references to field and laboratory methods and protocols
2. <i>Instrumentation</i>	Description of instruments including manufacturer, model/serial number, calibration methods
3. <i>Taxonomy and systematics</i>	References for taxonomic identification and voucher specimens
E. Data management methods	
1. <i>Data acquisition</i>	Description of instrumentation, data loggers, and manual data entry approaches that are used to acquire and record data; data verification approaches
2. <i>QA/QC</i>	Quality assurance and quality control procedures used to control data quality, identify and flag outliers, etc.
3. <i>Data processing</i>	Description of algorithms, procedures and software used in processing, deriving, integrating and transforming data
4. <i>Analyses</i>	Description of algorithms and software code used in analyzing and visualizing data
5. <i>Storage and preservation</i>	Description of how data, samples and specimens are stored and preserved including maps, field and lab notebooks, photographs and data products
F. History of dataset usage	Description of how and when data were used including references to the literature, data set revisions and updates
III. Physical structure	
A. File(s)	
1. <i>Identifier</i>	Unique name(s) assigned to the data file(s) comprising a dataset
2. <i>Size</i>	Total size in bytes of the file(s), number of rows and columns, number of records and record length
3. <i>File format and storage mode</i>	File type (e.g., ASCII, binary), data encoding and file compression schemes employed
4. <i>Authentication procedure(s)</i>	Checksum and other mechanisms for ensuring correct data transmission to others
B. Variables	
1. <i>Identifier</i>	Unique name(s) assigned to variable(s)
2. <i>Definition</i>	Definition of variable meaning
3. <i>Type</i>	Data type (integer, string, etc.)
4. <i>Units of measurement</i>	SI units of measurement
5. <i>Precision</i>	Number of significant digits

(continued)

**Table 5.2** (continued)

Metadata descriptor	Definition
<i>6. List and definition of codes</i>	List and definition of all codes encountered in the data including missing value codes and data quality flags
<i>7. Data format</i>	Fixed or variable length, start and end columns, etc.
<i>8. Data anomalies</i>	Description of any errors, anomalies, and missing periods in the data

### 5.3 Metadata Standards

Different communities of practice develop different vocabularies. “Location,” for example, can have a very different meaning for a geographer who is interested in points and areas on the earth’s surface as opposed to a librarian who is attempting to acquire an electronic record or document. Likewise, different fields may vary with respect to the detail needed in the metadata. An art historian may be satisfied with a place name, whereas a geographer may need to know the precise geographic coordinates and the georeferencing system employed.

Metadata standards have been developed to standardize the vocabularies and promote clarity in definition of terms and consistent use of those terms within or across applications and disciplines. Dozens of metadata standards exist. They have been created for the environmental sciences, social sciences, arts, financial markets, libraries, education, and many other domains and disciplines. Standards differ with respect to the number of descriptors (i.e., elements included) and the amount of structure enforced; more structured and finely detailed metadata may be “read” and interpreted by machines.

Some of the standards that are more commonly used in the ecological and environmental sciences are described below. Most researchers do not necessarily access the standards directly. Instead, the standards provide the basis for the metadata descriptors that are employed by different metadata tools and GIS and database programs (see Sect. 5.4). Nevertheless, it is usually considered good practice to know that the documentation that you are creating adheres to a community-accepted metadata standard as opposed to no standard or an ad hoc “standard” that has been created for use within a single laboratory or organization.

#### 5.3.1 *Dublin Core Metadata Initiative*

The Dublin Core Metadata Initiative (DCMI) is an open organization that is managed as a project of the Association for Information Science and Technology. DCMI supports shared innovation and design of metadata and best practices across a broad range of purposes and business models. DCMI contributes and maintains various community resources such as user guidelines, model-related specifications, and controlled vocabularies. One of the key DCMI resources is the Dublin Core

**Table 5.3** Description of the 15 elements of the Dublin Core Metadata Element Set (DCMI 2016)

Element	Description
Contributor	Name of the person, organization, or service responsible for contributing to the resource
Coverage	The spatial (e.g., named place or geographic coordinates) or temporal (e.g., named period, date, or date range) extent to which the resource applies
Creator	Name of the person, organization, or service responsible for creating the resource
Date	A single date or period associated with an event (e.g., creation, revision) in the life of the resource
Description	A description of the resource (e.g., abstract, a table of contents, free-text explanation of the resource)
Format	File format or type of physical medium for the resource
Identifier	An unambiguous reference to the resource using a formal identifier such as a DOI, URL, URN
Language	Language of the resource
Publisher	Name of the person, organization, or service responsible for making the resource available
Relation	A related resource usually identified by a formal identifier
Rights	A statement about the various rights (including intellectual property rights) associated with a resource
Source	Identification of any sources (using a formal identifier) from which the resource is partially or wholly derived
Subject	The topic of the resource as identified using key words, classification codes, etc.
Title	The name assigned to the resource
Type	The nature of the resource

Metadata Element Set—a vocabulary of fifteen properties for use in resource description. Table 5.3 describes the elements of the “Dublin Core,” which has been standardized as ISO Standard 15836:2009 (ISO 2016a) and ANSI/NISO Standard Z39.85-2012 (NISO 2016). The name “Dublin” is due to the fact that a 1995 invitational workshop was held in Dublin, Ohio; “Core” refers to generic elements that can be used to describe a wide variety of resources. Any changes to the Dublin Core Metadata Element Set are reviewed by a DCMI committee.

### 5.3.2 *Darwin Core*

The Darwin Core is a set of standards that includes a vocabulary or range of terms that are primarily used to document taxa and their occurrence (Wieczorek et al. 2012; TDWG 2016). The Darwin Core is based on the Dublin Core Metadata Initiative standards (see Sect. 5.3.1). Darwin Core documents describe how terms can be used, how terms are managed, and how terms can be extended to meet new purposes. The “Simple Darwin Core” is often used to refer to a particular specification that allows one to share taxa (or biodiversity) data in a structured flat-file

format. Several categories of terms are included in the Simple Darwin Core including: (1) *record-level* terms (e.g., language, license and access rights); (2) *occurrence* (e.g., catalog number, sex, and life stage); (3) *organism* (e.g., name, ID); (4) *material sample, living specimen, preserved specimen, or fossil specimen*; (5) *event* (e.g., year, month, day, time, field notes); (6) *location* (e.g., continent, country, state or province, latitude, longitude); (7) *geological context* (e.g., epoch); (8) *identification* (e.g., date identified); and (9) *taxon* (TDWG 2016). “Generic Darwin Core” extends the Simple Darwin Core for relational data and includes two additional terms: resource relationship and measurement or fact. Biodiversity data are typically shared as Darwin Core Archives (a popular mechanism for packaging and sharing data files).

### 5.3.3 *Ecological Metadata Language*

Ecological Metadata Language (EML) was developed specifically for the ecological and environmental sciences. EML has several key features. First, EML is modular in design and consists of numerous modules that allow one to document data resources, literature, software and protocols (Table 5.4). Its modular structure enables EML to be easily extended to support different discipline-specific data descriptors via the addition of new modules. Second, EML is comprehensive and structured in a way that key metadata elements can be machine-processed. This structured approach enables the development of advanced data discovery and processing services. Third, EML syntax is generally compatible with other metadata standards such as the Dublin Core Metadata Initiative and the International Standards Organization’s Geographic Information Standard ISO 19115 (ISO 2016b). Fourth, EML supports strong data typing, which means that the contents of an element can be validated against what is allowed for that field (e.g., is “17/01/15” an acceptable entry for date?). Importantly, EML is designed so that metadata can be generated as a standalone resource so that both metadata and data (or references to the data) can be combined in a “data package.”

### 5.3.4 *GBIF Metadata Profile*

The Global Biodiversity Information Facility (GBIF) Metadata Profile was created to standardize how data sets are described in the GBIF data portal (GBIF 2011a). The GBIF Metadata Profile is based on EML (see Sect. 5.3.3) and includes several elements (or resource types from Table 5.4): (1) dataset; (2) project; (3) people and organizations; (4) keywords; (5) coverage (i.e., taxonomic, spatial, temporal); (6) methods; (7) intellectual property rights; and (8) additional metadata and natural collections descriptions data such as collection identifiers and preservation methods (GBIF 2011a). The GBIF Integrated Publishing Toolkit (IPT) metadata editor is

**Table 5.4** Description of Ecological Metadata Language modules (KNB 2015a)

EML module	Description
<i>Top-level resources</i>	<i>Modules used to describe four different types of resources:</i>
Dataset	This module contains general information that describes a dataset such as the title, abstract, keywords, contacts, and purpose. The dataset module may import other modules that describe the dataset in greater details (e.g., methods, protocols, project, access).
Literature	The module contains information (e.g., literature citation, including title, abstract, keywords, and contacts) that describes literature resources (e.g., journal articles, books, chapters, conference proceedings, maps, and presentations).
Software	The module contains general information that describes software resources that were used to create, process and analyze a dataset.
Protocol	The module contains specific information that defines the standardized methods used to generate and process a dataset.
<i>Supporting modules</i>	<i>Modules for adding detail to top-level resources:</i>
Access	This module describes the level of access controls associated with a dataset and/or the associated metadata (e.g., individuals and groups that have been granted permission to access the resources).
Physical	This module describes the physical characteristics of a data object (e.g., filename, size, data format) as well as how to access the resource (e.g., offline and/or online locations such as the URL).
Party	The module describes the people and organizations responsible for creating, managing, and maintaining datasets and metadata.
Coverage	The module contains fields for describing the spatial, temporal and taxonomic coverage of a resource (e.g., N-S-E-W bounding coordinates, single or range of dates and times, taxon names and/or common names).
Project	This module contains information that describes the research context for the project such as hypotheses and questions being addressed, research sponsors, experimental or study design, and the study area.
Methods	This module describes the methods that were employed to generate the dataset, including field and laboratory methods, QA/QC procedures, and analytical steps.
<i>Data organization</i>	<i>Modules used to describe dataset structures:</i>
Entity	This module contains the information that characterizes each entity in the dataset. Entities are usually tables of data (e.g., ASCII text files, spreadsheets, relational database tables), but datasets may also contain relational database management system views as well as raster, vector and image data that may be further described using modules from the next section (e.g., dataTable, spatialRaster).
Attribute	This module contains the information that describes each attribute (e.g., variable, column) in a dataset entity, including the name and definition of the attribute, definitions of coded values and flags, and other information.
Constraint	This module defines the relationships among and within dataset entities, including primary and foreign key constraints, and others.
<i>Entity types</i>	<i>Modules that provide detailed information for discipline-specific entities:</i>

(continued)

**Table 5.4** (continued)

EML module	Description
dataTable	This module defines the characteristics of the data table, including columns/variables (using the EML-attribute module), coverage, methodology used to create the data table, and other information.
spatialRaster	This module supports the description of georeferenced rectangular grids of data values, including how the raster cells are organized, characteristics of the image and spectral bands, etc.
spatialVector	This module supports description of points and vectors and the relationships among them.
spatialReference	This module defines coordinate systems for referencing the spatial coordinates of a dataset employing either a library of pre-defined coordinate systems or support for customized projections.
storedProcedure	This module supports definition of the complex DBMS queries and transactions that produce a data table.
View	This module describes a DBMS view—i.e., a query statement that is stored as a database object and is executed each time the view is called.
<i>Utility modules</i>	<i>Modules that enhance metadata documentation:</i>
Text	Supports addition of enhanced text to other EML modules (e.g., sections, paragraphs, lists, subscript, superscript, emphasis, etc.).

normally used to generate metadata that conform to the GBIF Metadata Profile (see Sect. 5.4.1).

### 5.3.5 FGDC CSDGM

The US Federal Geographic Data Committee (FGDC) adopted the Content Standard for Digital Geospatial Metadata (CSDGM) in 1994 as the standard to employ for documenting geospatial data and became the *de facto* geospatial metadata standard for many years, especially by US federal, State and local governments (FGDC 2016). The Standard was revised in 1998 to allow different communities to add new elements to the CSDGM (i.e., extensions) as well as customized adaptations for specific domains (i.e., profiles). Additions to the second version (FGDC-STD-001-1998) of the FGDC Standard included: (1) CSDGM: Extensions for Remote Sensing Metadata (for remote sensing platforms and sensors); (2) the Biological Data Profile of the CSDGM (to support data types that are not explicitly geospatial such as specimen collections, laboratory results and field notes); and (3) the Metadata Profile for Shoreline Data (for defining and mapping shoreline data) (FGDC 2016). The FGDC CSDGM has largely been supplanted by ISO 19115 (ISO 2016b; see Sect. 5.3.6).



### 5.3.6 *ISO 19115*

ISO 19115 is a metadata standard for geospatial data that was created by the International Organization for Standardization (ISO 2016b). The standard was initially developed for dealing with geospatial data (e.g., map data and Geographic Information System data, much of it in point and vector forms), but was also revised to support imagery and gridded (i.e., raster) data. The standard defines how to describe geographic data, geographic services, and geographic features and feature properties. It includes a comprehensive array of metadata elements that enable users to access, interpret and, potentially, use the data such as the data content, spatial and temporal coverage, spatial referencing scheme, quality, and other properties of data. The most recent version (ISO 19115–1:2014) defines mandatory and optional metadata elements as well as the minimum metadata required to support data discovery, determination of fitness-for-use, and transfer and use of data and services (ISO 2016b). ISO 19115 can also be extended to meet specialized needs or to describe other resources such as imagery and gridded data (ISO 2016c).

## 5.4 Metadata Management

Metadata generation and management need not be an onerous burden. In this section, some of the existing metadata tools that can be used to create and manage metadata are described. In addition, several best practices that can expedite the creation of good, comprehensive metadata are presented.

### 5.4.1 *Metadata Tools*

Metadata tools can be used to greatly simplify the process of generating and managing metadata in the ecological and biodiversity sciences. That said, relatively few standalone metadata management tools exist. Table 5.5 lists some of the more commonly used tools that are currently available. The tools differ by the languages, metadata standards and operating systems that they support as well as cost, complexity and ease of use. Morpho, for example, is a free, downloadable standalone package that can be used to generate EML-compliant metadata for ecological data and other types of data (Higgins et al. 2002; Fegraus et al. 2005; Jones et al. 2007). Morpho can also be used to incorporate data tables and the associated metadata into data packages that can then be stored in archives such as KNB (Table 5.5).

In contrast to Morpho, ArcGIS is a comprehensive and expensive geographic information system that is used to manage geospatial data and metadata from many different scientific domains. Metadata are typically generated by the user (i.e., data originator) and managed in ArcGIS using a variety of styles that are compliant with

**Table 5.5** Metadata tools that can be used to generate and manage ecological and biodiversity metadata, including the metadata standards supported by the tools and relevant references

Tool name	Description	Metadata standard (s) supported	Reference(s)
ArcGIS	ESRI's ArcGIS provides metadata styles that support viewing and editing of metadata that are compliant with community standards	FGDC, ISO 19115	ESRI (2016)
CatMDEdit	A metadata editor tool that facilitates the documentation of resources, especially geographic information resources on multiple platforms and in different languages	Dublin Core, FGDC, ISO 19115	Nogueras-Iso et al. (2012)
GeoNetwork OpenSource	A comprehensive free and open source catalog application that supports metadata editing and search functions on multiple platforms	Dublin Core, FGDC, ISO19115	GeoNetwork Opensource (2016)
IPT (GBIF)	The Integrated Publishing Toolkit (IPT) is a GBIF metadata editor for publishing occurrence and taxonomic data, and general metadata about data sources as Darwin Core Archives	GBIF Meta-data Profile	GBIF (2011b, 2016a), Robertson et al. (2014)
Morpho	A free and easy to use package that works on multiple platforms that can be used to create and edit metadata, create or view and download data packages, share and publish data, and specify access controls	Ecological Metadata Language	Higgins et al. (2002), Fegraus et al. (2005), Jones et al. (2007), KNB (2015b, c)
Tkme	A tool for creating and editing FGDC-compliant metadata that runs on Windows and Unix systems	FGDC	USGS (2016a)
Xtme	A Unix-based version of tkme	FGDC	USGS (2016b)

community geospatial metadata standards such as FGDC and ISO 19115. Several standalone tools for generating geospatial metadata are listed in Table 5.5.

The situation with respect to biodiversity metadata tools is a bit more complicated. Biodiversity data are aggregations of species occurrence records and checklists that are stored and managed on numerous hardware and software platforms including: collections management software systems like Arctos (2016), EMu (2016), Symbiota (Gries et al. 2014; Symbiota 2016), and Specify Software Project (2016); more general database management systems like Microsoft Access and FileMaker; and spreadsheet programs like Microsoft Excel. The platforms may

differ significantly in how the data are stored and described (i.e., the terms used to describe the data), but usually terms can be mapped to the Darwin Core standard. The metadata are normally added when the data are published as Darwin Core Archives. Some platforms like Specify provide a Schema Mapper and Data Exporter that enables biodiversity data and metadata to be published as Darwin Core Archives. In other cases, the GBIF Integrated Publishing Toolkit (IPT) is used to encode data in the Darwin Core standard and publish data and metadata as Darwin Core Archives (Robertson et al. 2014). Darwin Core Archives are comprised of one or more comma delimited text files (i.e., CSV files), an XML document that describes the structure of the files and the relationships among the files, and a metadata file (in either Dublin Core or Ecological Metadata Language) that describes the dataset (Wieczorek et al. 2012). Darwin Core Archives can then be registered with the GBIF registry so the data are readily discoverable and can be accessed by others (Robertson et al. 2014).

### ***5.4.2 Best Practices for Creating and Managing Metadata***

There are three primary keys to success for creating good metadata. First, *start early*. Metadata should be created at the inception of your project and updated as the project evolves. In doing so, you are less likely to forget or fail to include key details and the job will be much easier if small tasks are tackled on a routine basis. Second, *engage all relevant parties* in generating and managing the metadata. For example, a typical research project may involve a lead investigator, a technician or staff member, and one or more students. All such individuals are probably “touching the data” and should, therefore, be involved in documenting protocols and procedures, data quality issues, and all other relevant aspects of the data. Third, *treat the metadata as a living document*. Initial metadata, for example, can be created and maintained in a metadata management tool or, even, in an online laboratory notebook or shared document that relevant personnel can update and review on a frequent basis. The important point is to review and revise the metadata frequently throughout the project.

Several additional best practices can simplify the process of developing metadata and facilitate the creation of metadata that will enable the data to be interpreted and used for the long term. First, *use metadata standards and tools*. Formal metadata standards and tools normally provide a comprehensive list of metadata elements or descriptors that should be captured as well as an easy-to-use interface for capturing, editing and publishing metadata. Second, *create metadata that can be understood by someone that is unfamiliar with the project*. In particular, do not use jargon, ensure that all acronyms and terms are defined, and include references (i.e., citations) to methods, protocols, and other documents that will assist others in understanding the data. It is especially important to use standardized terms for locations, subject keywords and taxa to promote consistency and facilitate data discovery. Table 5.6 lists several such resources including indexes, ontologies, thesauri, and databases.

**Table 5.6** Resources for standardizing keywords (e.g., subject terms), place names, and taxonomic names (e.g., plants, animals, fungi) in metadata

Resource	Description	Reference
Biocomplexity thesaurus	Thesaurus for the biodiversity, ecological and environmental sciences that is maintained by the USGS Core Science Analytics & Synthesis (CSAS) Program. Searches of terms identify: broader, related, and narrower terms, subject categories, scope notes, and equivalence relationships.	USGS (2016c)
Catalogue of life	A comprehensive and authoritative global index of species of animals, plants, fungi and micro-organisms	Catalogue of Life (2016)
Encyclopedia of life (EOL)	EOL contains a variety of information about species including names, text, images, video, sounds, maps and data	EOL (2016)
Environment Ontology (EnvO)	The Environment Ontology (EnvO) contains standard terms for biomes, environmental features, and environmental material	EnvO (2016)
GeoNames	Geographical database of names that covers all countries and contains over eight million place names (e.g., populated places, administrative boundaries, lakes, mountains, islands)	GeoNames (2016)
Global Biodiversity Information Facility (GBIF)	GBIF contains databases of species, species occurrences, and datasets that can be searched, viewed and downloaded	GBIF (2016b)
Global Change Master Directory	A public metadata inventory that includes >34,000 Earth science dataset and service descriptions; searches may be performed using an extensive (hierarchically structured) controlled vocabulary, free-text, and locations and dates	NASA (2016)
Global Names Index	An indexed collection of character strings that have been used as organism names based on numerous scientific names repositories	Patterson et al. (2010), GNI (2016)
Index Fungorum	The Index Fungorum database contains names of fungi at all ranks	Index Fungorum (2016)
Index to Organism Names (ION)	ION contains organism names related data (e.g., animals, plants, bacteria and viruses) gathered from the scientific literature for Thomson Reuters' databases	Thomson Reuters (2016)

(continued)

**Table 5.6** (continued)

Resource	Description	Reference
International Plant Names Index (IPNI)	A database of the names and basic bibliographical information about seed plants, ferns and lycophytes. IPNI is supported by The Royal Botanic Gardens, Kew, The Harvard University Herbaria, and the Australian National Herbarium	IPNI (2016)
<i>iPlant Taxonomic Name Resolution Service</i>	A utility for correcting and standardizing plant names against specific taxonomies	iPlant Collaborative (2016)
Integrated Taxonomic Information System (ITIS)	A searchable database of authoritative taxonomic information on plants, animals, fungi, and microbes from North America and the world	ITIS (2016)
Taxonomy Database	A curated database that contains classification and nomenclature for organisms in the public sequence databases	National Center for Biotechnology Information (2016)
<i>(Universal Biological Indexer and Organizer)</i> uBio	uBio is a comprehensive catalog of known names of living and extinct organisms	uBio (2016)
World Registry of Marine Species (WoRMS)	WoRMS contains a large authoritative and comprehensive list of names of marine organisms	WoRMS (2016)
ZooBank	ZooBank is the official registry of published scientific names for animals, according to the International Commission on Zoological Nomenclature	International Commission on Zoological Nomenclature (ICZN) (2016)

## 5.5 Conclusion

Data are neither useful nor usable unless accompanied by sufficient metadata that allows one to understand the context, format, structure, and fitness-for-use of the data. Generally, those data sets that are well documented have greater utility and longevity than data that are only minimally described. Using community-accepted metadata standards and tools can greatly simplify the process of developing comprehensive metadata that can be interpreted and used for the long term. Following three best practices will help insure that your data products persist and can be used for years to decades. First, begin creating and capturing metadata at the beginning of your project to avoid the loss of important details. Second, engage all people who touch the data (e.g., database design, data collection, data QA/QC and analysis, data preservation) in generating and reviewing the metadata. Third, use the metadata—i.e., routinely review and revise the metadata and, importantly, share the metadata with others that may wish to use and understand the data. The real test is if other

researchers that are unfamiliar with your data collection and organization can access, interpret and use data that you have produced well into the future.

**Acknowledgments** Special thanks to Laura Russell and Dave Vieglais of the University of Kansas Biodiversity Institute for explaining some of the details related to biodiversity data and metadata publication.

## References

- Arctos (2016) Arctos collaborative collection management solution. <http://arctos.database.museum>. Accessed 17 Apr 2016
- Catalogue of Life (2016) Catalogue of life. <http://www.catalogueoflife.org/>. Accessed 17 Apr 2016
- Cook RB, Wei Y, Hook LA et al (2017) Preserve: protecting data for long-term use, Chapter 6. In: Recknagel F, Michener W (eds) Ecological informatics. Data management and knowledge discovery. Springer, Heidelberg
- DCMI (Dublin Core Metadata Initiative) (2016) Dublin Core Metadata Element Set, Version 1.1. <http://dublincore.org/documents/dces/>. Accessed 17 Apr 2016
- EMu (2016) EMu Museum Management System. <https://emu.kesoftware.com/>. Accessed 17 Apr 2016
- EnvO (Environment Ontology) (2016) EnvO. <http://www.environmentontology.org/Browse-EnvO>. Accessed 17 Apr 2016
- EOL (2016) Encyclopedia of life. <http://eol.org/>. Accessed 17 Apr 2016
- ESRI (2016) ArcGIS resources. <http://resources.arcgis.com/>. Accessed 17 Apr 2016
- Fegraus EH, Andelman S, Jones MB et al (2005) Maximizing the value of ecological data with structured metadata: an introduction to Ecological Metadata Language (EML) and principles for metadata creation. *Bull Ecol Soc Am* 86(3):158–168. doi:10.1890/0012-9623(2005)86[158:MTVOED]2.0.CO;2
- FGDC (Federal Geographic Data Committee) (2016) Geospatial metadata standards. <http://www.fgdc.gov/metadata/geospatial-metadata-standards#csdgm>. Accessed 17 Apr 2016
- GBIF (2011a) GBIF metadata profile, reference guide, Feb 2011 (contributed by Ó Tuama E, Braak K), Global Biodiversity Information Facility, Copenhagen, p.19. [http://links.gbif.org/gbif\\_metadata\\_profile\\_how-to\\_en\\_v1](http://links.gbif.org/gbif_metadata_profile_how-to_en_v1). Accessed 17 Apr 2016
- GBIF (2011b) GBIF metadata profile – how-to guide, (contributed by Ó Tuama E, Braak K, Remsen D), Global Biodiversity Information Facility, Copenhagen, 11 p. <http://www.gbif.org/resource/80641>. Accessed 17 Apr 2016
- GBIF (2016a) GBIF tools overview. <http://www.gbif.org/infrastructure/tools>. Accessed 17 Apr 2016
- GBIF (Global Biodiversity Information Facility) (2016b) GBIF.org. <http://www.gbif.org/species>. Accessed 17 Apr 2016
- GeoNames (2016) GeoNames. <http://www.geonames.org/>. Accessed 17 Apr 2016
- GeoNetwork Opensource (2016) GeoNetwork Opensource. <http://geonetwork-opensource.org/>. Accessed 18 Apr 2016
- GNI (2016) GNI: Global Names Index. <http://gni.globalnames.org/>. Accessed 18 Apr 2016
- Gries C, Gilbert EE, Franz NM (2014) Symbiota – a virtual platform for creating voucher-based biodiversity information communities. *Biodiv Data J* 2:e1114. doi:10.3897/BDJ.2.e1114
- Higgins D, Berkley C, Jones MB (2002) Managing heterogeneous ecological data using Morpho. In: Proceedings of the 14th international conference on scientific and statistical database management (SSDBM'02), IEEE Computer Society, pp 69–76
- Index Fungorum (2016) Index Fungorum. <http://www.indexfungorum.org/>. Accessed 18 Apr 2016

- International Commission on Zoological Nomenclature (ICZN) (2016) International commission on zoological nomenclature: about ZooBank. <http://iczn.org/content/about-zoobank>. Accessed 18 Apr 2016
- iPlant Collaborative (2016) Taxonomic name resolution service v4.0. <http://tnrs.iplantcollaborative.org/>. Accessed 18 Apr 2016
- IPNI (2016) The International Plant Names Index (IPNI) <http://www.ipni.org/>. Accessed 18 Apr 2016
- ISO (2016a) ISO 15836:2009 Information and documentation -- The Dublin Core metadata element set. [http://www.iso.org/iso/home/store/catalogue\\_tc/catalogue\\_detail.htm?csnumber=52142](http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=52142). Accessed 18 Apr 2016
- ISO (2016b) ISO 19115-1:2014 Geographic information – metadata – Part 1: Fundamentals. [http://www.iso.org/iso/home/store/catalogue\\_tc/catalogue\\_detail.htm?csnumber=53798](http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=53798). Accessed 18 Apr 2016
- ISO (2016c) ISO 19115-2:2009 Geographic information – Metadata – Part 2: Extensions for imagery and gridded data. [http://www.iso.org/iso/home/store/catalogue\\_tc/catalogue\\_detail.htm?csnumber=39229](http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=39229). Accessed 18 Apr 2016
- ITIS (2016) Integrated Taxonomic Information System (ITIS). <http://www.itis.gov/>. Accessed 18 Apr 2016
- Jones C, Blanchette C, Brooke M et al (2007) A metadata-driven framework for generating field data entry interfaces in ecology. *Ecol Inf* 2:270–278
- KNB (Knowledge Network for Biocomplexity) (2015a) Ecological Metadata Language (EML). <https://knb.ecoinformatics.org/#external/emlparser/docs/index.html>. Accessed 30 Mar 2015
- KNB (Knowledge Network for Biocomplexity) (2015b) Morpho. <https://knb.ecoinformatics.org/#tools/morpho>. Accessed 30 Mar 2015
- KNB (Knowledge Network for Biocomplexity) (2015c) Morpho user guide. <https://knb.ecoinformatics.org/software/dist/MorphoUserGuide.pdf>. Accessed 30 Mar 2015
- Michener WK (2017) Project data management planning, Chapter 2. In: Recknagel F, Michener W (eds) *Ecological informatics. Data management and knowledge discovery*. Springer, Heidelberg
- Michener WK, Brunt JW, Helly J et al (1997) Non-geospatial metadata for the ecological sciences. *Ecol Appl* 7:330–342
- NASA (2016) NASA global change master directory. <http://gcmd.nasa.gov/>. Accessed 18 Apr 2016
- National Center for Biotechnology Information, U.S. National Library of Medicine (2016) Taxonomy. <http://www.ncbi.nlm.nih.gov/taxonomy/>. Accessed 18 Apr 2016
- NISO (2016) The Dublin Core metadata element set. [http://www.niso.org/apps/group\\_public/download.php/10256/Z39-85-2012\\_dublin\\_core](http://www.niso.org/apps/group_public/download.php/10256/Z39-85-2012_dublin_core). Accessed 18 Apr 2016
- Nogueras-Iso J, Latre MA, Bejar R et al (2012) A model driven approach for the development of metadata editors, applicability to the annotation of geographic information resources. *Data Knowl Eng*. doi:10.1016/j.datak.2012.09.001
- Patterson DJ, Cooper J, Kirk PM et al (2010) Names are key to the big new biology. *TREE* 25:686–691
- Robertson T, Döring M, Guralnick R et al (2014) The GBIF Integrated Publishing Toolkit: facilitating the efficient publishing of biodiversity data on the internet. *PLoS One* 9(8): e102623. doi:10.1371/journal.pone.0102623
- Specify Software Project (2016) Specify Software Project. <http://specifyx.specifysoftware.org/>. Accessed 18 Apr 2016
- Symbiota (2016) Symbiota. <http://symbiota.org/docs/>. Accessed 18 Apr 2016
- TDWG (Taxonomic Databases Working Group) (2016) Darwin Core. <http://rs.tdwg.org/dwc/>. Accessed 18 Apr 2016
- Thomson Reuters (2016) Index to Organism Names (ION). <http://www.organismnames.com/>. Accessed 18 Apr 2016
- uBio (2016) uBio. <http://www.ubio.org/>. Accessed 18 Apr 2016

- USGS (2016a) Tkme: another editor for formal metadata. <http://geology.usgs.gov/tools/metadata/tools/doc/tkme.html>. Accessed 18 Apr 2016
- USGS (2016b) Xtme: another editor for formal metadata. <http://geology.usgs.gov/tools/metadata/tools/doc/xtme.html>. Accessed 18 Apr 2016
- USGS (2016c) USGS core science analytics, synthesis, and libraries - biocomplexity thesaurus. [http://www.usgs.gov/core\\_science\\_systems/csas/biocomplexity\\_thesaurus/index.html](http://www.usgs.gov/core_science_systems/csas/biocomplexity_thesaurus/index.html). Accessed 18 Apr 2016
- Wieczorek J, Bloom D, Guralnick R et al (2012) Darwin Core: an evolving community-developed biodiversity data standard. PLoS One 7(1):e29715. doi:[10.1371/journal.pone.0029715](https://doi.org/10.1371/journal.pone.0029715)
- WoRMS (2016) WoRMS: world register of marine species. <http://www.marinespecies.org/> Accessed 18 Apr 2016



# Chapter 6

## Preserve: Protecting Data for Long-Term Use

Robert B. Cook, Yaxing Wei, Leslie A. Hook, Suresh K.S. Vannan,  
and John J. McNelis

**Abstract** This chapter provides guidance on fundamental data management practices that investigators should perform during the course of data collection to improve both the preservation and usability of their data sets over the long term. Topics covered include fundamental best practices on how to choose the best format for your data, how to better structure data within files, how to define parameters and units, and how to develop data documentation so that others can find, understand, and use your data easily. We also showcase advanced best practices on how to properly specify spatial and temporal characteristics of your data in standard ways so your data are ready and easy to visualize in both 2-D and 3-D viewers. By following this guidance, data will be less prone to error, more efficiently structured for analysis, and more readily understandable for any future questions that the data products might help address.

### 6.1 Introduction

Preservation certainly encompasses the idea that there should be no loss of bits associated with a data product. In this chapter, we will expand this definition of preservation, to include all of the data management practices that will preserve the data at a high-enough level of quality so that it is usable well into the future. Well-curated and -preserved data will be easily discovered and accessed, understood by future users, and serve to enable others to reproduce the results of the original study. Preservation, in this broad sense, starts when the seed-ideas for a project are first pulled together, and continues until the data have been successfully finalized, curated, archived, and released for others to use (Whitlock 2011).

Proper preservation of the data files is an important part of a research project, as important as the sample design, collection, and analysis protocols in ensuring the overall success of a project. Often researchers do not spend enough effort ensuring that the data are properly managed, described, and preserved. Without well-

---

R.B. Cook • Y. Wei • L.A. Hook • S.K.S. Vannan (✉) • J.J. McNelis  
Oak Ridge National Laboratory, Oak Ridge, TN, USA  
e-mail: [rbcook7@gmail.com](mailto:rbcook7@gmail.com); [weiy@ornl.gov](mailto:weiy@ornl.gov); [hookla@ornl.gov](mailto:hookla@ornl.gov); [santhanavans@ornl.gov](mailto:santhanavans@ornl.gov);  
[mcnelisjj@ornl.gov](mailto:mcnelisjj@ornl.gov)

prepared data—no matter how carefully the sample design, collection, and analysis were done for a project—the research team may not be able to effectively use the data to test their hypotheses. And the data will not be useful for any potential future users.

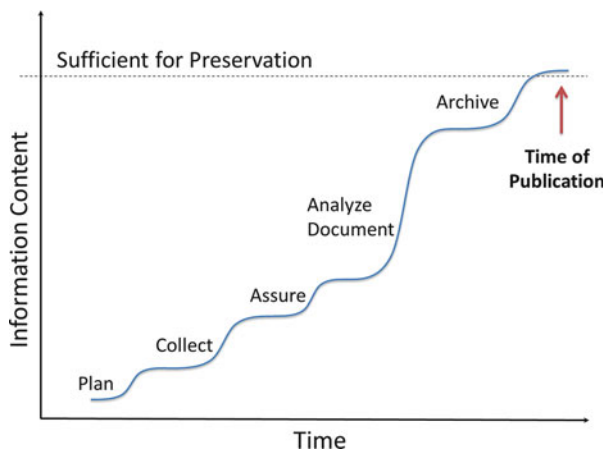
Well-preserved ecological observations will continue to help us understand the functioning of the global ecosystem. More importantly, the data of ecological observations provide the foundation for advancing and sustaining economic, environmental, and social well being (Reid et al. 2010; IGBP 2012; USGEO 2015). Thus, well-preserved ecological data are critically needed to address global sustainability—what could certainly be considered the grand scientific challenge of the twenty-first century (Reid et al. 2010; IGBP 2012).

### 6.1.1 Preservation and Its Benefits

We will define preservation as preparing data packages—data, documentation, and metadata—for a user 20 years into the future (NRC 1991); some advocate even 100 years (Justice et al. 1995). The rationale is that those who generated the data initially or those who worked with the data when the data were first compiled will have forgotten the details of the data within a few years (Michener et al. 1997) (Fig. 5.1). Developing descriptive information for someone 20 or more years out who is unfamiliar with the project, methods, and observations will ensure that the information is preserved and, most importantly, usable (Fig. 6.1) (NRC 1991).

Well-managed and preserved data have many benefits. During the course of a project, investigators who make a habit of preparing organized and well-described data will spend less time doing data management and more time doing research.

**Fig. 6.1** With proper data management and preservation during the course of a project, information about the data is compiled during the data life cycle (plan, collect, analyze, assure, document, and archive; Strasser et al. 2012). Metadata and documentation are recorded so that future users will be able to find and use the data products



Researchers can pick up data files after being away from them for a period and immediately use the data without having to remember what the data means or how filters or analyses were done. Furthermore, researchers can hand off data and documentation to collaborators who can readily understand and use data files, without further explanation.

When the project has been completed and the data are finalized and properly curated, scientists outside your project can find, understand, and use your data to reproduce the findings of your research. Perhaps even more importantly, these data products can be used to address additional broader-scale research questions (Reid et al. 2010; Whitlock 2011; Michener 2017d; Schildhauer 2017). FLUXNET is an example of a project that started out studying the scientific mysteries of individual flux tower sites, but evolved to address larger scale questions across biomes and climate domains. Along with this scientific evolution, FLUXNET has experienced a data evolution in which the community has embraced standard methods for observations and processing, and has come to appreciate the advantages of placing data into common formats, with standard units and parameter names. This standardization facilitates combining data from 10s to 100s of flux towers to address broad questions that cannot be addressed by individual projects (Baldocchi et al. 2012; Papale et al. 2012). A common set of standards ultimately saves time, but requires buy-in, which takes time for investigators to realize the benefits.

Funding agencies protect their investment in Earth science research, through preservation of observations; many funding agencies require that data generated through their grants be shared over the long term (Whitlock 2011). The preserved observations provide the means to understand Earth processes, develop and test models, and provide information for decision makers. Not preserving data products so that they can effectively be used will decrease the return on research investment, and more importantly hinder our ability to advance Earth science.

Some journals (e.g., PNAS, Ecological Monographs), scientific societies (e.g., Ecological Society of America) now require that the data used in a paper be archived before the paper can be published, and others require that the data be shared (PLoS, Nature, Science; Michener 2015). In both cases, data citations with Digital Object Identifier (DOI) locators will allow readers to find the archived data (Cook et al. 2016). Following data management practices for long-term preservation will make it easier for authors to archive their data products associated with a submitted manuscript to meet this requirement.

Another benefit of data preservation is that others will use these well-curated data, resulting in the data producers getting credit. Data repositories have started to provide data product citations, each with a DOI (Parsons et al. 2010; Cook et al. 2016). A benefit of data preservation is that through data product citations (Cook et al. 2009, 2016), data authors get credit for archived data products and their use in other papers, in a manner analogous to article citations. In addition, readers of those articles can obtain the data used in an article (Cook et al. 2016) through the DOI locator.

## 6.2 Practices for Preserving Ecological Data

This chapter is written for a broad audience—for those who are creating data products, for those who may need to prepare the data products for archival, and for those who will access and use the archived data. Accordingly, we will present preservation activities that data contributors, data archives, and data users can perform to preserve data products and make them useful in the future. The focus will be on application of preservation principles, and less so with theoretical/academic aspects of preservation. We are orienting this chapter toward practical aspects, because ecologists may be willing to share their data, but they typically do not have knowledge and training of data management practices that they can use to facilitate sharing (Tenopir et al. 2011; Kervin et al. 2014).

Geospatial, or location, information is a fundamental component of ecological data. The preservation practices described here are primarily for geospatial data products, including tabular data as well as map and image data. Earlier best practices for data sharing focused almost exclusively on tabular data (Olson and McCord 2000; Cook et al. 2001), but the focus has expanded with improvements in sensors, standards, and processing software, and many ecologists are turning to geospatial data.

This chapter builds on the chapter on documentation and metadata (Michener 2017c). Because the metadata descriptors were thoroughly treated there, we will focus on human readable text documents that provide another view into the data. These text documents contain the contextual information about samples—under what conditions was the sample collected, what antecedent conditions influenced the sample, and what do others need to know about the sample context in order to understand the data.

The remainder of Sect 6.2 describes best data management practices that investigators can perform to improve the preservation and usability of their data for themselves and for future users.

### 6.2.1 *Define the Contents of Your Data Files*

The data compiled during a project is derived from the science plan (hypotheses/proposal) for that project. During the proposal writing stage, the investigator should identify the information needed to address the hypotheses and the best way to compile that information. Sometimes that compilation will be to collect samples and make measurements, other times it may be to run models to obtain output, or even fuse data from multiple sources to create a necessary product.

Also during the proposal writing stage, a Data Management Plan (DMP) (Michener 2017a) should be developed that lays out the content and organization of the data based on a comprehensive list of data required for the project. The environmental study will compile a suite of primary measurements along with

contextual and ancillary information that defines the study area (soil, landcover, plant functional types, weather, nutrient status, etc.).

Investigators should keep a set of similar measurements together in one *data file*. The similarity extends to the same investigator, site, methods, instrument, and time basis (all data from a given year, site, and instrument in one file). Data from a continental study of soil respiration at 200 plots could be one data file, but 30-min meteorological data from 30 sites over 5 years could be five data files (one per year) or 30 data files (one per site). We do not have any hard and fast rules about contents of each file, but we suggest that if the documentation/metadata for data are the same, then the data products should all be part of one data set.

## 6.2.2 *Define the Parameters*

Defining the name, units, and format used for each parameter within a project should be done with a clear view to the standards or guidelines of the broader community. Using widely accepted names, units, and formats will enable other researchers to understand and use the data. Ideally, the files, parameter names, and units should be based on standards established with interoperability in mind (Schildhauer 2017).

The SI (International System) should be used for units and ISO be used for formats. The ISO Standard 8601 for dates and time (ISO 2016) recommends the following format for dates:

yyyy-mm-dd or yyyymmdd, e.g., January 2, 2015 is 2015-01-02 or 20150102

which sorts conveniently in chronological order. ISO also recommends that time be reported in 24-h notation (15:30 hours instead of 3:30 p.m. and 04:30 instead of 4:30 a.m.).

In observational records, report in both local time and Coordinated Universal Time (UTC). Avoid the use of daylight savings time because in the spring the instrument record loses 1 h (has a gap of 1 h) and in the autumn, instrument records have a duplicate hour.

The components needed to define temporal information with sufficient accuracy for ecological data include the following: calendar used, overall start and end temporal representation of a data parameter, time point/period that each data value represents, and temporal frequency of a data parameter. As an important example, Daymet (Thornton et al. 2017), a 1-km spatially gridded daily weather data set for North America, uses the standard, or Gregorian, calendar and leap years are considered. But the years within Daymet always contain 365 days; Daymet does this by dropping December 31 from leap years. The documentation for Daymet defines this information (e.g., start and end times of each time step, which days are included and which days are not). Following the Climate and Forecast (CF) Metadata convention (Eaton et al. 2011) and the ISO 8601 Standard (ISO 2016), temporal information of Daymet is accurately defined.

CF Metadata, a convention for netCDF-formatted files, is becoming more common in ecological modeling and in some field studies. These conventions allow combination and ready analysis of data files, and importantly, facilitate the use of field data to parameterize and drive models with a minimum of conversions.

In addition to enabling integration of data files, standard units can be easily converted from one unit to another using a tool such as UDUNITS library (UCAR 2016).

For each data file, investigators should prepare a table that identifies the parameter, provides a detailed description of that parameter, and gives the units and formats (Table 6.1).

**Table 6.1** Portion of a table describing contents and units (dos-Santos and Keller 2016)

Column heading	Units/format	Description
Site		Fazenda Cauaxi or Fazenda Nova Neonita. Both located in the Municipality of Paragominas
Area		Code names given to the site areas. The areas are PAR_A01 for the Fazenda Nova Neonita or CAU_A01 for the Fazenda Cauaxi
Transect		The transect ID number within an area. Transect = plot.
tree_number		Tree number assigned to each tree in each transect
date_measured	yyyy-mm-dd	Date of measurements
UTM_easting	m	X coordinate of tree individual location. Fazenda Cauaxi is in UTM Zone: 22S. Fazenda Nova Neonita is in UTM Zone: 23S
UTM_northing	m	Y coordinate of tree individual location. Fazenda Cauaxi is in UTM Zone: 22S. Fazenda Nova Neonita is in UTM Zone: 23S
common_name		Common name of tree. MORTA = dead tree
scientific_name		Scientific name of tree. NI = not identified. For common_name = MORTA (dead) or LIANA, scientific names are not provided.
DBH	cm	Diameter at breast height (DBH), 1.3 m above the ground. Measured on both live and standing dead trees.
height_total	m	Total Height (m), measured using a clinometer and tape as the height to the highest point of the tree crown. Measured on both alive and standing dead trees. Fazenda Cauaxi site 2012 only—not measured in 2014.

**Table 6.2** Characteristics of sites from the Scholes (2005) study

Site name	Site code	Latitude	Longitude	Elevation	Date
Units		(deg)	(deg)	(m)	
Kataba (Mongu)	K	-15.43892	23.25298	1195	2000-02-21
Pandamatenga	P	-18.65651	25.49955	1138	2000-03-07
Skukuza Flux Tower	skukuza	-31.49688	25.01973	365	2000-06-15

Provide another table that describes each study site or area used in the data product [location, elevation, characteristics (climate or vegetation cover)], along with a formal site name (Table 6.2).

Once the metadata about a record is defined, be sure to use those definitions, abbreviations, units consistently throughout the data set and the project. For air temperature, pick one abbreviation and use it consistently. Do not use T, temp., MAT (mean annual temp), and MDT (mean daily temp) within a data set, if they all mean the same parameter; using one consistently will be much easier for users to understand, particularly as they write code to process the values.

When data values are not present in the data file, investigators should indicate this with a *missing value code*. We suggest that an extreme value never observed (e.g., -9999) be used consistently to indicate that the value is missing.

### 6.2.3 Use Consistent Data Organization

There are several different ways to organize data files. For *tabular data*, one way is similar to a spreadsheet table in which each row in a file represents a complete record, and the columns represent the parameters that make up the record. The table should have a minimum of two header rows, the first of which identifies the parameter names and the second header row identifies the parameter units and format (Table 6.3) (Cook et al. 2001). A suggestion for data files is that a column containing a unique id for each record be included for provenance tracking.

Another perfectly appropriate alternative for *tabular data* is to use the structure found in relational databases. In this arrangement, site, date, parameter name, value, and units are placed in individual rows; unique ids could also be placed in this row. This table is typically skinny (only 5 or 6 columns wide) and long, holding as many records (rows) as needed in the study (Table 6.4). This arrangement allows new parameters to be added to a project in the future without changing the tabular data columns.

For whichever organization chosen, be consistent in file organization and formatting throughout the entire file (Porter 2017). The file should have a separate set of header rows that describes the content of the file. For example, the first row of the file should contain file name, data set title, author, date, and any related companion file names (Table 6.5) (Hook et al. 2010). Within the body of the file,

**Table 6.3** An arrangement of content in which all of the information for a particular site and date (e.g., site, date, parameter name, value and unit) is placed into one row

Station	Date	Temp.	Precip.
Units	YYYYMMDD	C	mm
HOGI	20121001	12	0
HOGI	20121002	14	3
HOGI	20121003	19	-9999

**Table 6.4** An arrangement of information in which each row in a file represents a complete record, and the columns represent the parameters that make up the record

Station	Date	Parameter	Value	Unit
HOGI	20121001	Temp.	12	C
HOGI	20121002	Temp.	14	C
HOGI	20121001	Precip.	0	mm
HOGI	20121002	Precip.	3	mm

do not change or re-arrange the columns or add any notes in marginal cells. Additional features provided by specific software, such as colored highlighting or special fonts (bold, italicized, etc.) that indicate characteristics to humans are not useful for computers, and any information contained in the colors or fonts will not be preserved.

*Spatial data* files containing vector data, such as ESRI’s Shapefile format, treat each point, line, or polygon as a unique record described by a set of common attributes. Records within a shapefile are organized in tabular format where each row corresponds to a feature representing the location or area to which the row’s attributes pertain. Tabular data stored inside ESRI shapefiles are limited by character count and cannot contain special characters so it is good practice to maintain a complementary data dictionary file that defines the parameters, abbreviations, and units.

#### 6.2.4 Use Stable File Formats

A key aspect of preservation is to ensure that computers can read the data file well into the future. Experience has shown that proprietary and non-standard formats often become obsolete and difficult or even impossible to read. Operating systems, the proprietary software, and the file formats will no longer be supported and researchers are left with useless bits.

Over the short term, usually during the course of the research, it is fine to use familiar proprietary data formats. But be sure that those formats can be exported into an appropriate format (without loss of information) suitable for long-term preservation.

Standardized, self-describing, and open data formats are recommended for long-term preservation of ecological data (Table 6.6). Standardized formats increase interoperability of data and lower the barrier of integrating heterogeneous data (Schildhauer 2017). Self-describing formats make data easier to use by a wide range of users. More importantly, open formats ensure consistent support and improvement from user communities and increase longevity of ecological data. Standardized and open formats also serve as a solid basis for developing data access, subsetting, and visualization tools.



**Table 6.5** An example of a well-organized portion of a file with a set of header rows that describe the file (file name, contributor, citation, date, and any relevant notes)

File name	NGEE_Arctic_Barrow_Soil_Incubations_2012										
Date modified:	2015-10-27										
Contact:	Colleen Iversen (iversencm@oml.gov)										
Data set DOI	doi:10.5440/1185213										
Notes	For more information, see data set DOI										
Region	Locale	Latitude	Longitude	Date_sampled	Thaw_Depth	Soil_Horizon	Carbon_concentration_of_soil_layer	Nitrogen_concentration_of_soil_layer			
		Decimal_degrees	Decimal_degrees	yyyy-mm-dd	cm		Percent	Percent			
North Slope	Barrow	71.280867	-156.611606	2012-08-01	36.8	O	42.94	2.66			
North Slope	Barrow	71.280867	-156.611606	2012-08-01	36.8	O	42.94	2.66			
North Slope	Barrow	71.280867	-156.611606	2012-08-01	36.8	O	42.94	2.66			
North Slope	Barrow	71.280867	-156.611606	2012-08-01	36.8	Mi	39.36	2.24			
North Slope	Barrow	71.280867	-156.611606	2012-08-01	36.8	Mi	39.36	2.24			
North Slope	Barrow	71.280867	-156.611606	2012-08-01	36.8	Mi	39.36	2.24			
North Slope	Barrow	71.280867	-156.611606	2012-08-01	36.8	DO	31.6	1.72			
North Slope	Barrow	71.280867	-156.611606	2012-08-01	36.8	DO	31.6	1.72			
North Slope	Barrow	71.280867	-156.611606	2012-08-01	36.8	DO	31.6	1.72			
North Slope	Barrow	71.280891	-156.61164	2012-08-01	39.4	O	34.42	1.76			
North Slope	Barrow	71.280891	-156.61164	2012-08-01	39.4	O	34.42	1.76			
North Slope	Barrow	71.280891	-156.61164	2012-08-01	39.4	O	34.42	1.76			

The body of the file is consistently organized and completed (Iversen et al. 2015)

**Table 6.6** Recommended formats for ecological data preservation (ESO 2016; Edinburgh Data Share 2015)

Format	Description
Text/CSV	Suitable for representing tabular data such as field observations and site characteristics.
Shapefile	Most widely used open format for representing vector data, such as points, lines, and polygons.
GeoTIFF	Open and popular format for storing geospatial raster imageries.
HDF/ HDF-EOS	A feature-rich format suitable for storing complex multi-dimensional and multi-parameter scientific data. The HDF format and its EOS extension (HDF-EOS) have been widely used for NASA earth observation mission data for many years.
netCDF	Similar to HDF but simpler; ideal for storing multi-dimensional and multi-parameter data. Combined with Climate & Forecast (CF) convention, netCDF data files can be standardized and self-describing, which can greatly advance data interoperability. netCDF is gaining popularity in many research communities.

### 6.2.5 Specify Spatial Information

Almost all ecological data are location-relevant and many also have an associated time component. For example, photos taken of field sites should be associated with the accurate location, elevation, direction, and time information; otherwise they will not be suitable for research. There are many other spatial and temporal data types, for example, soil respiration observations across the world, MODIS Leaf Area Index (LAI) maps, and global 0.5-degree monthly Net Ecosystem Exchange (NEE) simulations generated from terrestrial biosphere models. When preparing ecological data for use or long-term preservation, their spatial (“where”) and temporal (“when”) information need to be accurately defined.

Two critical components of spatial information include the Spatial Reference System (SRS) used and the spatial extent, boundary, resolution, and scale under the given SRS. For example, Daymet v3 (Thornton et al. 2017) provides daily weather parameters at 1-km spatial resolution for North America from 1980 to 2016. It uses a special SRS called Lambert Conformal Conic and its definition using the Open Geospatial Consortium (OGC) Well-Known Text (WKT) standard is shown in Table 6.7.

Under this SRS, X and Y coordinates of each of the 1-km grid cells are accurately defined following the CF convention in the netCDF files where Daymet data are stored.

### 6.2.6 Assign Descriptive File Names

Even desktop personal computers can have large hard drives, and it can be very easy to lose files and information on such large drives. To prevent time spent

**Table 6.7** Example Spatial Reference System, showing the projection, spatial extent, boundary, resolution and scale

---

```

PROJCS["North_America_Lambert_Conformal_Conic",
  GEOGCS["GCS_North_American_1983",
    DATUM["North_American_Datum_1983",
      SPHEROID["GRS_1980",6378137,298.257222101]],
    PRIMEM["Greenwich",0],
    UNIT["Degree",0.017453292519943295]],
  PROJECTION["Lambert_Conformal_Conic_2SP"],
  PARAMETER["False_Easting",0],
  PARAMETER["False_Northing",0],
  PARAMETER["Central_Meridian",-96],
  PARAMETER["Standard_Parallel_1",20],
  PARAMETER["Standard_Parallel_2",60],
  PARAMETER["Latitude_Of_Origin",40],
  UNIT["Meter",1],
  AUTHORITY["EPSG","102009"]]

```

---

searching for files, organize the information in a directory or folder structure based on project or activity. The directory structure and file names need to be both human- and machine-readable and so the names should contain text characters only and contain no blank spaces (Cook et al. 2001; Hook et al. 2010). Carefully check for any operating or database system limitations on characters (upper or lowercase, special characters, and file name lengths).

Use descriptive file names that are unique and reflect the contents of the files. In the metadata and documentation define the terms and acronyms in the file names. Examples of good file names include “daymet\_v3\_tmax\_annavg\_1988\_na.nc4”, a Daymet version 3 file containing daily maximum and annual average maximum temperature in 1988 for North America (na) in netCDF-4 format (Thornton et al. 2017).

Names should also be clear both to the user and to those with whom the files will be shared. File names like “Mydata.xls,” “2001\_data.csv,” and “best version.txt” do not adequately describe the file and would not be useful to understand the contents.

While the name should be descriptive and unique, the file name is not the location for all of the metadata associated with a file. A standard metadata record in XML format is a much more useful location for detailed information about a data file, and will be accessible by APIs. See Michener (2017c) on metadata.

### **6.2.7 Document Processing Information**

To preserve your data and its integrity, save your raw data in a “read-only” form (Strasser et al. 2012). By doing so, the raw data will not be affected by any changes, either purposeful or inadvertent. Some spreadsheet type software allows cells to be deleted inadvertently with the slip of a finger on a keyboard. Read only files will prevent those sorts of changes.

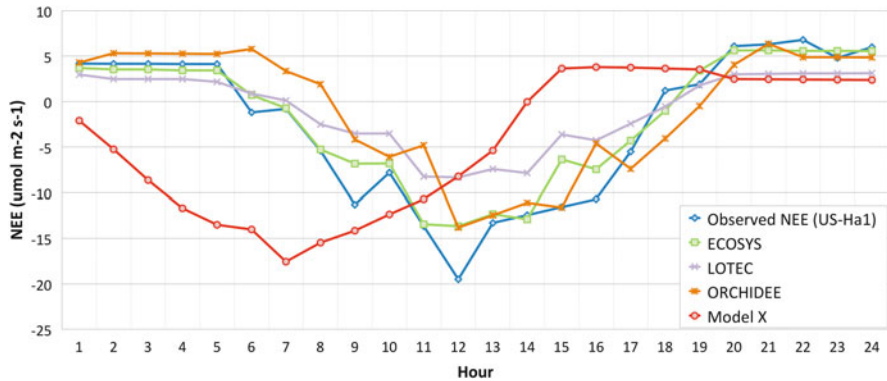
Use a scripted language such as “R”, “SAS” or “MATLAB” to process data in a separate file, located in a separate directory (Hook et al. 2010; Strasser et al. 2012). The scripts you have written are an excellent record of data processing, can also easily and quickly be revised and rerun in the event of data loss or requests for edits, and have the added benefit of allowing a future worker to follow-up or reproduce your processing. The processing scripts serve as the basis for a provenance record. An example R script and some figures generated from the script are captured in Appendix of this chapter.

Scripts can be modified to improve or correct analyses, and then rerun against the raw data file. This approach can be especially beneficial when preparing manuscripts. Two or three months after the analyses have been run and written up, reviewers may want to have changes made (new filtering or statistical analysis, additional data, etc.). Scripts saved along with data files serve as a record of the analysis and can quickly be modified to meet the reviewer’s need. If they were not saved, authors may have difficulty resurrecting the exact formula and perhaps even the data used in the analysis.

### **6.2.8 Perform Quality Assurance**

Quality assurance pertains not only to the data values themselves, but also to the entire data package. All aspects of the data package need to be checked including parameter names, units, documentation, file integrity, and organization, as well as the validity and completeness of data values. One can think of quality assurance of a data set like the careful steps authors go through to finalize an accepted paper for publication.

There are a number of specific checks that researchers can perform to ensure the quality of data products (Cook et al. 2001; Hook et al. 2010; Michener 2017b). The organization within data files has to be consistent. Data should be delimited, lining up in the proper column (Cook et al. 2001). Key descriptors, like sample identifier, station, time, date, and geographic location, should not be missing. Parameter names should follow their definition, and the spelling and punctuation should not vary. Perform an alphabetical sort of the parameter names to identify discrepancies. Check the content of data values through statistical summaries or graphical



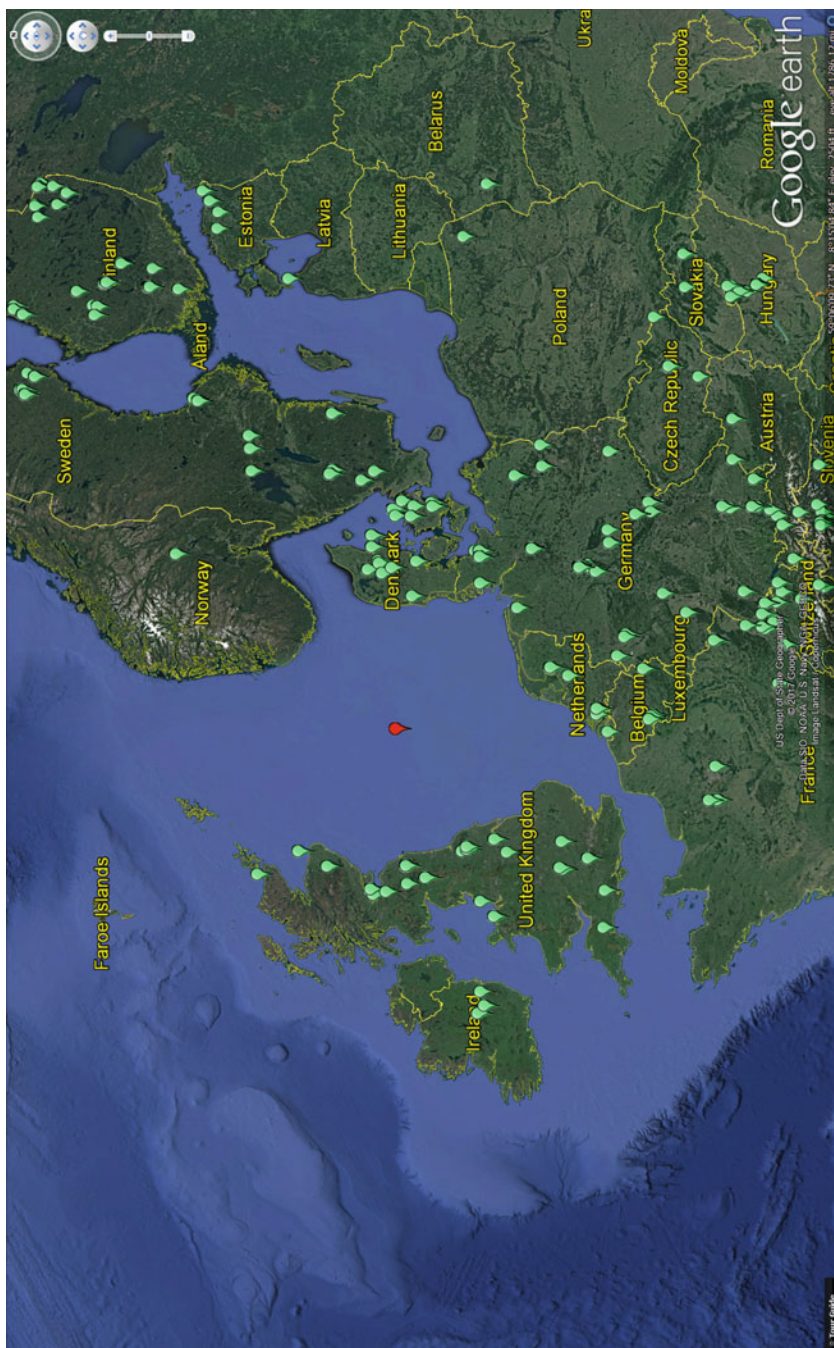
**Fig. 6.2** Comparison of diurnal Net Ecosystem Exchange (NEE) for the Harvard Forest Flux Tower with terrestrial biosphere model output of NEE used to quickly identify quality issues. While most of the models are consistent with the timing and magnitude of noontime NEE, the onset and conclusion of the phytoperiod shows some variation among models, especially Model X, which was an outlier because of improper documentation. It was run with UTC time instead of Eastern US time but was not labeled carefully and was mistakenly plotted with a peak NEE 5 h earlier than the tower or other models (Ricciuto et al. 2013)

approaches to look for anomalous or out of range values. A number of different graphical approaches (leaf diagram, box and whisker diagram, histograms, scatterplots, etc.) are described in Michener (2017b). Another approach is to generate plots of time-series data to check for the physical reasonableness of the values and to ensure that the time zone is correct (Fig. 6.2). Plot the data on a map to make sure that the site locations are as expected (Cook et al. 2001). Common errors in spatial data are placing sites in the wrong hemisphere by not including the correct sign of latitude or longitude or providing the spatial accuracy required to place the site correctly on a shoreline, rather than mistakenly in a lake or coastal ocean (e.g., Fig. 6.3).

There is no better quality assurance than to use the data files in an analysis. Issues with the files, units, parameters, and other aspects of the data products will become evident and draw the attention of the analysts.

### 6.2.9 Provide Documentation

The documentation accompanying a data set should describe the data in sufficient detail to enable users to understand and reuse the data. The documentation should describe the goals of the project, why the data were collected, and the methods used for sample collection and analysis, and data reduction. The description should be detailed enough to allow future researchers to combine that data with other similar data across space, time, and other disciplines (Rüegg et al. 2014).



**Fig. 6.3** Map showing locations of terrestrial sites where soil respiration was measured. The coordinates for the site in the middle of the North Sea need to be checked (Bond-Lamberty and Thomson 2014)

A data set document should contain the following information:

- **What** does the data set describe?
- **Why** was the data set created?
- **Who** produced the data set?
- **When** and how frequently were the data collected?
- **Where** were the data collected and with what spatial resolution?
- **How** was each parameter measured?
- **How** reliable are the data (e.g., what is the uncertainty and measurement precision and accuracy? what problems remain in the data set?)?
- **What** assumptions were used to create the data set (e.g., spatial and temporal representativeness)?
- **What** is the use and distribution policy of the data set?
- **How** can someone get a copy of the data set?
- **Provide** any references to use of data in publication(s)

Often a data set is a collection of multiple files. Each file should be described, including file names, temporal and spatial extent of the data, and parameters and units. If all of the files are the same, this information can be contained in the data set metadata and data set documentation. If each file is unique, in terms of contents, then each should be described separately with file-level metadata record and a file description document. The purpose for such a description is so that an investigator can use an automated method to search for an individual file or even part of the file that is required (e.g., XML metadata record or even self-describing file, like netCDF or HDF; Michener 2017c). If each file is not named in a descriptive manner or described in a document, then a user would have to manually view each file to obtain the required data, something that no one would want to do, especially for big data collections.

### **6.2.10** *Protect Your Data*

Everyone knows the sickening feeling when files are lost, due to hard drive crashes or from other problems. They have either experienced the feeling themselves or know someone who has lost their drives or files. A desktop, laptop, or server is fine one day and the next a problem has come up with the hard drive, and the files have disappeared. Backups are the key to surviving such losses. If you do not have backups, then you cannot retrieve the information and your files are not preserved.

Researchers—really anyone using computers—should create back-up copies often, to ensure that information is not lost. Ideally researchers should create three copies, the original, one on-site, and one off-site (Brunt 2010). The off-site storage prevents against hazards that may affect an institution such as fire, floods, earthquakes, and electrical surges. Data are valuable and need to be treated



accordingly, with appropriate risk mitigation. Cloud-based storage is becoming a valid option for storing and protecting data, especially as an off-site backup solution.

Frequency of the backups is based on need and risk. If you are compiling data from a high frequency sensor, then frequent (e.g., 30-min or hourly) backups are warranted to ensure that the information is not lost due to a disk crash. One can develop a backup strategy that relies on a combination of sub-daily, daily, weekly, and monthly backups, to cut back on the number of individual backups saved but still maintain sufficient backup so that no work is lost.

A critical aspect of any backup is that it be tested periodically, so that you know that you can recover from a data loss. After the initial shock of losing data, there is nothing worse than having the false hope of a backup that is not intact and is corrupted.

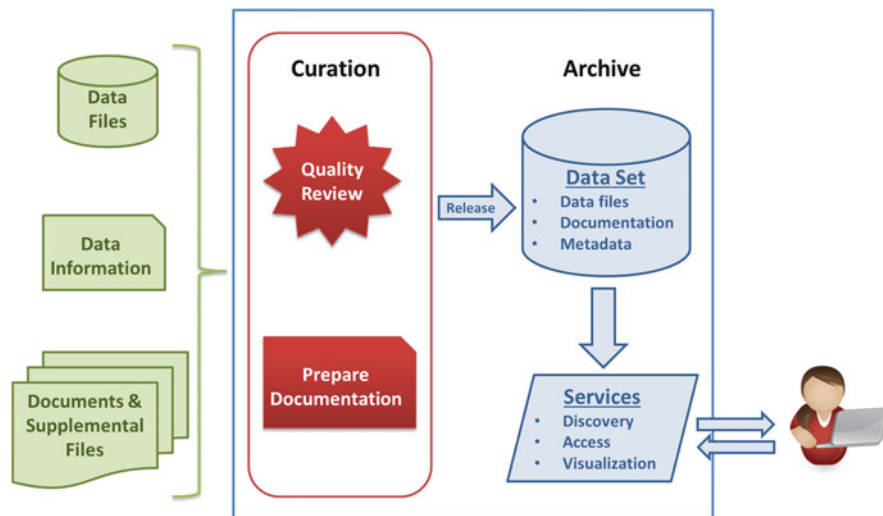
Another aspect of protecting your data deals with data transfers (e.g., over the Internet, such as large files, large numbers of files, or both). Ensure that file transfers are done without error by reconciling what was sent and received, using checksums and lists of files.

### 6.3 Prepare Your Data for Archival

The practices in Sect. 6.2 should have provided the background needed to prepare consistently structured, thoroughly defined, and well-documented data products. During the course of the project, the data should have been easy-to-share with team members, and readily analyzed to address the project's science questions.

At the end of the project, the data products need to be turned over to a data archive for curation and long-term storage (Fig. 6.4). Transitioning the data to an archive should have been part of the initial project planning conducted during the proposal writing stage, when a Data Management Plan (DMP) (Michener 2017a) was developed. The Plan should have identified the data center responsible for curating and archiving the data, and the investigators should have made initial contact with the archive before the proposal was submitted. The DMP should have included some information about the archive, their requirements, and a statement of collaboration by the archive. Because of space restrictions, the two-page DMP would not have included much detailed information. During the research project, the team should have interacted with data center personnel to inform them of the types and formats of data products being produced. Key characteristics that the data center needs to know are the volume and number of files, the delivery dates, and any special needs for the data (viewers or other tools, restricted access, etc.). Suggest a title that is concise in its description of the data set's scientific content and that indicates its spatial and temporal coverage. The data center will have requirements, and the project should identify what those are early in the project to ensure those requirements are incorporated before the data are submitted to the archive.





**Fig. 6.4** Flow of data and documentation from the investigator team to the archive, where quality checks are performed and documentation is compiled. After the data are released, users can search for, download, and use data of interest

Data archives will need data files, documentation that describes the files and the content (Sect. 6.2.9), and, if possible, standardized metadata records (Michener 2017c). As part of the data package, some data centers require supplemental information such as sample design, sample collection and analysis methods, algorithms, code, and data analysis methods, description of field sites, photographs, articles using the data, etc. All of this information will provide context for those who are trying to understand and use the data, especially many years into the future.

## 6.4 What the Archive Does

After the data sets have been finalized, and the project team has transmitted them to the archive, the archive staff begins the process of curation leading to long-term preservation (Lavoie 2000). This section will briefly describe what typically happens to a data set during curation and the services that investigators receive when they archive a data set.

The archive is selected based on a number of factors. The agency that funded the research may have a designated archive, perhaps based on science area. In recent years, a principal investigator's institution, often the library, will provide long-term stewardship. Some journals and scientific societies have preferences for where data associated with published articles should be archived.

### 6.4.1 *Quality Assurance*

A data center goes through the following general steps during curation, summarized in the following list:

1. *Files received as sent*

After the data have been received, the archivist will check the numbers of files and the “checksum” to ensure that the files were received as sent (see Sect. 6.2.10). At this time, staff will also make sure that the file type is appropriate for long-term storage and use (see Sect. 6.2.4).

2. *Documentation describes files*

The archivist will read the documentation and any manuscript associated with the data product to get an understanding of why the data were produced and what the workflow is. If there are a number of unique files, a table will be generated that identifies the contents of each file or group of files. The archivist will check the filenames to ensure they are descriptive and appropriate based on the file content, date, spatial extent, etc. (see Sect. 6.2.6).

3. *Parameters and units defined*

The documentation and the data files should provide the parameter definitions and the units. For tabular data, the data provider should have created a table that defines the column names and units; if not, the archivist could generate this useful table. Often the original investigator will be contacted to identify “mystery” parameters that are not identified, defined, or are unitless (see Sect. 6.2.2).

4. *File content is consistent*

For *spatial data* files, the analyst may view the file or a sample of the files in a GIS tool for consistency. The datum, projection, resolution, and spatial extent will be exported from all files and checked for consistency (see Sect. 6.2.5).

For *tabular data*, the archivist will ensure that the parameter definitions and units are consistent across all files (see Sect. 6.2.2).

5. *Parameter values are physically reasonable*

The maximum and minimum value will be exported and the range checked for reasonableness (see Sect. 6.2.2).

Geospatial tabular data will be loaded onto a basemap for visual inspection of proper overlay (see Sect. 6.2.8).

Staff will check that missing values and other flags are reasonable and consistent (see Sect. 6.2.2). If a scale factor is applied, the archivist will make sure that it is defined.

6. *Reformat and reorganize data files if needed*

The archivist will judge if the formats and organization of the received data files are the most appropriate based on their data stewardship expertise in the relevant research fields and the interactions with data providers. If needed, received data files will be reformatted and reorganized to ease the usage and maximize the future interoperability of data.

## 6.4.2 *Documentation and Metadata*

The archive will often generate two types of documentation. One is a metadata record in standardized format to describe the data and also to find data within a large archive (Michener 2017c). The second is a data set document (a readme type document) that provides a description of the data (what, where, when, why, who) and references to manuscripts using data (see Sect. 6.2.9).

Sometimes the investigator drafts a metadata record (Michener 2017c) using metadata-editing tools, but more often the data center will compile the metadata record.

Often the archivist will generate a data set document that defines all of the parameters and units, based on information or manuscripts provided by the investigators. Each file or type of file in the data set will be described and the spatial and temporal domain and resolution will be provided. The document should also describe the methods and limitations and estimates of quality or uncertainty. The data set document will also include browse images or figures that effectively illustrate the data set contents.

The investigator provides documents with contextual information (see Sect. 6.3) that is archived along with the data files and data set documentation.

A key part of data curation is to generate a data citation that gives credit to the data contributors and the archive, as well as provide a DOI that allows others to find and use the data. Data product citations have structures similar to manuscript citations and include authors, date released, data set title, data center, and DOI (ESIP 2014; Starr et al. 2015; Cook et al. 2016).

## 6.4.3 *Release of a Data Set*

After curating and archiving data, data centers can perform a number of services that benefit both the data users, the data providers, and the funders of the archive as well as funders of the research project. The following list contains a summary of archive activities after the data have been released:

1. Advertise data through email, social media, and website
2. Provide tools to explore, access, visualize, and extract data
3. Provide long-term, secure archiving (back-up and recovery)
4. Address user questions, and serve as a buffer between users and data contributors
5. Provide usage statistics and data citation statistics
6. Notify users when newer versions/updates of data products are available, particularly users who have downloaded the out-of-date data.

Data derived from research is advertised and made available through discovery and access tools. The data can be used to address other hypotheses and when those results are reported in a paper, the original data are cited, which can be used as a measure of the impact of that work and the data center on science.

## 6.5 Data Users

The key responsibility for the users of archived data is to give proper credit to the data contributors. Using other's ideas and research products, including data, requires proper attribution. Data used should be cited in a manner similar to articles, with callouts in the text, tables or figures, and a complete citation with DOI locator in the list of references. Compilation of all of the citations of a data set into a data citation index will ensure that the data authors are given credit for all of the effort associated with making the measurements and compiling a well-preserved data product.

A secondary responsibility of data users is to identify any issues with the data files or documentation or discovery or access tools. Feedback to the data center and to the data contributor on these issues will improve the quality of the data and services at the archive.

## 6.6 Conclusions

Data management is important in today's science, especially with all of the advances in information technology. Sensors and other data sources can generate voluminous data products, storage devices can safely store data files for rapid access, and compute capabilities are sufficient to analyse and mine the big data. Internet transfer speeds are catching up, but in the short-term, cloud computing and storage has enabled access and analysis to occur within the same cluster.

Well-managed and organized data will enable the research team to work more efficiently during the course of the project, including sharing data files with collaborators so that they can pick up the files and begin using them with minimal training. Data that is thoroughly described and documented can potentially be re-used in ways not imagined when originally collected. For example, well-preserved Earth observations are important for understanding the operation of the Earth system and provide a solid foundation for sustaining and advancing economic, environmental, and social well-being.

Because of the importance of data management, it should be included in the research workflow as a habit, and done frequently enough that good data products are generated. The steps outlined in this and related chapters in this book will ensure that the data are preserved for future use.

## Appendix: Example R-Script for Processing Data

This R script (Table 6.8) analyzes a CSV data file of the ORNL DAAC-archived data set: "LBA-ECO CD-02 Forest Canopy Structure, Tapajos National Forest, Brazil: 1999–2003" (Ehleringer et al. 2011).

**Table 6.8** R-script that processes data from a read-only file and generates two figures

```
#####
# Example R script to process data file of an ORNL DAAC-archived data set: "LBA-ECO CD-02" #
# Forest Canopy Structure, Tapajos National Forest, Brazil: 1999-2003" #
# Input data file is stored in directory called "original", which is assigned with #
# read-only permission. #
# All output files, including processed data and plots, will be stored in another #
# directory called "analysis". #
# #
# version: 0.1 #
#####

# Set working directories
# Input CSV data file is in sub-dir "original", on which this script has read permission
# Outputs from this script are stored in sub-dir "analysis", on which this script has both
# read and write permission
setwd('/Users/ywi/Workspace/Temp/R_Scripts/ds1009/')
input_dir <- 'original'
output_dir <- 'analysis'

# Read in original CSV data file and save data into variable "lai"
# Notes: skip first 16 comment lines
#       parse headers on the 17th line
#       column separator is ","
input_file <- 'CD02_LAI_measurements_TNF.csv'
lai <- read.csv(paste(input_dir, '/', input_file, sep=''), header=TRUE, sep=',', quote='\\"',
dec = '.', skip = 16)

# Show summary information about variable lai
summary(lai)

# Select data records associated with site "km 67 - Primary Forest Tower" and save them
# into variable lai_sitel
site_id <- 'km 67 - Primary Forest Tower'
lai_sitel <- subset(lai, lai$Site_ID == site_id)

# Filter out records with non-positive tree height and non-positive LAI value
lai_sitel <- subset(lai_sitel, lai_sitel$Height > 0 & lai_sitel$LAI > 0)

# Save variable lai_sitel into CSV file CD02_LAI_measurements_TNF_Primary_Forest_Tower.csv,
# which is located in the analysis sub-directory.
output_file <- 'CD02_LAI_measurements_TNF_Primary_Forest_Tower.csv'
write.csv(lai_sitel, file=paste(output_dir, '/', output_file, sep=''), quote=TRUE);

# Create a histogram plot of the LAI values of all trees near the Primary Forest Tower
# Also, overlay a density distribution line on top of the histogram plot
# The plot will be saved into file Primary_Forest_Tower_Histogram_LAI.png in the
# "analysis" sub-directory
output_plot1 <- 'Primary_Forest_Tower_Histogram_LAI.png'
par(mar=c(3, 3, 0, 0), mgp=c(2, 1, 0), cex=5)
png(paste(output_dir, '/', output_plot1, sep=''), width=800, height=800, units = 'px')
hist(lai_sitel$LAI, main='', xlab='LAI', col='green', breaks=16, prob=TRUE)
lai_density <- density(lai_sitel$LAI)
lines(lai_density, col='red', lwd=10)
dev.off()

# Create a scatter plot showing the relationship between tree heights and LAI values
# Also, overlay a regression line of tree heights and LAI values on top of the scatter plot
# The plot will be saved into file Primary_Forest_Tower_LAI_Height_Plot.png in the
# "analysis" sub-directory
output_plot2 <- 'Primary_Forest_Tower_LAI_Height_Plot.png'
png(paste(output_dir, '/', output_plot2, sep=''), width=800, height=800, units = 'px')
par(mar=c(3, 3, 1, 1), mgp=c(2, 1, 0), cex=5)
plot(lai_sitel$Height, lai_sitel$LAI, main='', xlab='Tree Height (m)', ylab='LAI',
col='blue', type="p")
lai_height_reg <- lm(lai_sitel$LAI~lai_sitel$Height)
abline(lai_height_reg, col='red', lwd=10)
dev.off()
```

The script retrieves data records with positive height and LAI values for trees near the site designated “Primary Forest Tower.” After determining a frequency histogram (Fig. 6.5), it then analyzes the relationship between tree height and LAI

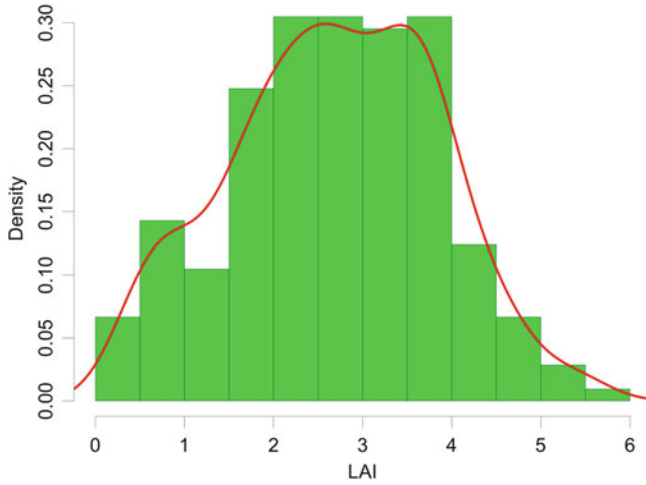


Fig. 6.5 Histogram of LAI for trees in primary forest near flux tower site

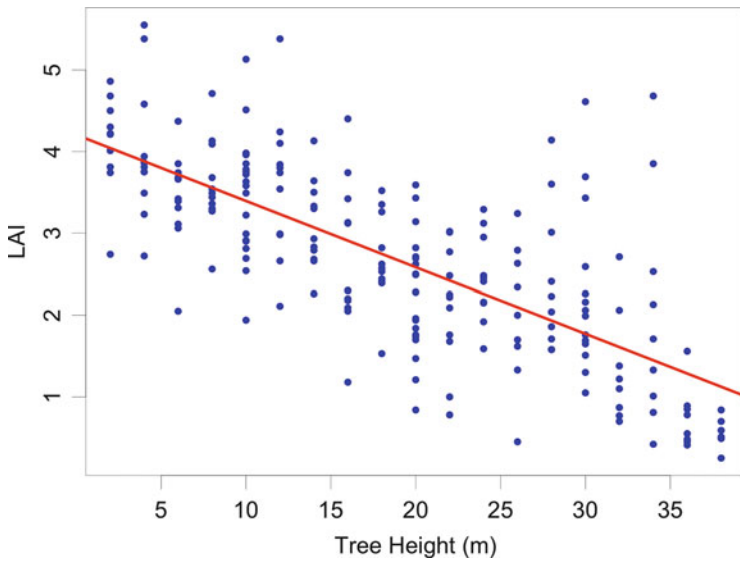


Fig. 6.6 Plot of LAI versus Tree Height for primary forest trees near the flux tower site

values. As revealed by the output plot (Fig. 6.6), height and LAI values have negative correlation for trees near site “Primary Forest Tower”.

Input CSV data file of this R script is stored in directory “original”, on which the script has only read-only permission. All outputs of this script are saved in directory “analysis”, for which the script has both read and write permission.

## References

- Baldocchi D, Reichstein M, Papale D et al (2012) The role of trace gas flux networks in the biogeosciences. *Eos Trans* 93:217–218. doi:[10.1029/2012EO230001](https://doi.org/10.1029/2012EO230001)
- Bond-Lamberty BP, Thomson AM (2014) A global database of soil respiration data, version 3.0. ORNL DAAC, Oak Ridge, TN. doi:[10.3334/ORNLDAAC/1235](https://doi.org/10.3334/ORNLDAAC/1235)
- Brunt JW (2010) Protecting your digital research data and documents: LTER cybersecurity briefing #1. <http://intranet2.lternet.edu/content/protecting-your-digital-research-data-and-documents>. Accessed 25 Jan 2015
- Cook RB, Olson RJ, Kanciruk P et al (2001) Best practices for preparing ecological and ground-based data sets to share and archive. *Bull Ecol Soc Am* 82:138–141. <http://www.jstor.org/stable/20168543>
- Cook RB, Post WM, Hook LA et al (2009) A conceptual framework for management of carbon sequestration data and models. In: McPherson BJ, Sundquist ET (eds) *Carbon sequestration and its role in the global carbon cycle*, AGU Monograph Series 183. American Geophysical Union, Washington, DC, pp 325–334. doi:[10.1029/2008GM000713](https://doi.org/10.1029/2008GM000713)
- Cook RB, Vannan SKS, McMurry BF et al (2016) Implementation of data citations and persistent identifiers at the ORNL DAAC. *Ecol Inf* 33:10–16. doi:[10.1016/j.ecoinf.2016.03.003](https://doi.org/10.1016/j.ecoinf.2016.03.003)
- dos-Santos MN, Keller MM (2016) CMS: forest inventory and biophysical measurements, Para, Brazil, 2012-2014. ORNL DAAC, Oak Ridge, TN. doi:[10.3334/ORNLDAAC/1301](https://doi.org/10.3334/ORNLDAAC/1301)
- Eaton B, Gregory J, Drach R et al (2011) NetCDF climate and forecast (CF) metadata conventions (Vers. 1.6). CF conventions and metadata. <http://cfconventions.org/cf-conventions/v1.6.0/cf-conventions.pdf>. Accessed 10 May 2016
- Edinburgh Data Share (2015) Recommended file formats. [http://www.ed.ac.uk/files/atoms/files/recommended\\_file\\_formats-apr2015.pdf](http://www.ed.ac.uk/files/atoms/files/recommended_file_formats-apr2015.pdf) Accessed 10 May 2016
- Ehleringer J, Martinelli LA, Ometto JP (2011) LBA-ECO CD-02 forest canopy structure, Tapajós National Forest, Brazil: 1999–2003. ORNL DAAC, Oak Ridge, TN. doi:[10.3334/ORNLDAAC/1009](https://doi.org/10.3334/ORNLDAAC/1009)
- ESIP (Earth Science Information Partners) (2014) Data citation guidelines for data providers and archives. doi:[10.7269/P34F1NNJ](https://doi.org/10.7269/P34F1NNJ)
- ESO (ESDIS Directory Standards Office) (2016) Standards, requirements and references. <https://earthdata.nasa.gov/user-resources/standards-and-references>. Accessed 20 Apr 2016
- Hook LA, Vannan SKS, Beaty TW et al (2010) Best practices for preparing environmental data sets to share and archive. Oak Ridge National Laboratory Distributed Active Archive Center, Oak Ridge, TN. doi:[10.3334/ORNLDAAC/BestPractices-2010](https://doi.org/10.3334/ORNLDAAC/BestPractices-2010)
- IGBP (International Geosphere Biosphere Program) (2012) The Merton Initiative: towards a global observing system for the human environment. <http://www.igbp.net/publications/themertoninitiative.4.7815fd3f14373a7f24c256.html>. Accessed 7 Mar 2016
- ISO (2016) Date and time format - ISO 8601. <http://www.iso.org/iso/home/standards/iso8601.htm>. Accessed 18 Apr 2016
- Iversen CM, Vander Stel HM, Norby RJ et al (2015) Active layer soil carbon and nutrient mineralization, Barrow, Alaska, 2012. Next generation ecosystem experiments arctic data collection, Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory, Oak Ridge, TN. doi:[10.5440/1185213](https://doi.org/10.5440/1185213)
- Justice CO, Bailey GB, Maiden ME et al (1995) Recent data and information system initiatives for remotely sensed measurements of the land surface. *Remote Sens Environ* 51:235–244. doi:[10.1016/0034-4257\(94\)00077-Z](https://doi.org/10.1016/0034-4257(94)00077-Z)
- Kervin K, Cook RB, Michener WK (2014) The backstage work of data sharing. In: *Proceedings of the 18th international conference on supporting group work (GROUP)*, Sanibel Island, FL, ACM, New York. doi:[10.1145/2660398.2660406](https://doi.org/10.1145/2660398.2660406)
- Lavoie B (2000) Meeting the challenges of digital preservation: the OAIS reference model. OCLC. <http://www.oclc.org/research/publications/library/2000/lavoie-oais.html>. Accessed 21 Aug 2015

- Michener WK (2015) Ecological data sharing. *Ecol Inform* 29:33–44. doi:[10.1016/j.ecoinf.2015.06.010](https://doi.org/10.1016/j.ecoinf.2015.06.010)
- Michener WK (2017a) Project data management planning, Chapter 2. In: Recknagel F, Michener W (eds) *Ecological informatics. Data management and knowledge discovery*. Springer, Heidelberg
- Michener WK (2017b) Quality assurance and quality control (QA/QC), Chapter 4. In: Recknagel F, Michener W (eds) *Ecological informatics. Data management and knowledge discovery*. Springer, Heidelberg
- Michener WK (2017c) Creating and managing metadata, Chapter 5. In: Recknagel F, Michener W (eds) *Ecological informatics. Data management and knowledge discovery*. Springer, Heidelberg
- Michener WK (2017d) Data discovery, Chapter 7. In: Recknagel F, Michener W (eds) *Ecological informatics. Data management and knowledge discovery*. Springer, Heidelberg
- Michener WK, Brunt JW, Helly J et al (1997) Non-geospatial metadata for ecology. *Ecol Appl* 7:330–342. doi:[10.1890/1051-0761\(1997\)007\[0330:NMFTES\]2.0.CO;2](https://doi.org/10.1890/1051-0761(1997)007[0330:NMFTES]2.0.CO;2)
- NRC (National Research Council) (1991) Solving the global change puzzle: A U.S. strategy for managing data and information, Report by the Committee on Geophysical Data, Geosciences, Environment and Resources, National Research Council. National Academy Press, Washington, DC. <http://dx.doi.org/10.17226/18584>
- Olson RJ, McCord RA (2000) Archiving ecological data and information. In: Michener WK, Brunt JW (eds) *Ecological data: design, management and processing*. Blackwell Science, Oxford, pp 117–130
- Papale D, Agarwal DA, Baldocchi D et al (2012) Database maintenance, data sharing policy, collaboration. In: Aubinet M, Vesala T, Papale D (eds) *Eddy covariance: a practical guide to measurement and data analysis*. Springer, Dordrecht, pp 411–436. doi:[10.1007/978-94-007-2351-1](https://doi.org/10.1007/978-94-007-2351-1)
- Parsons MA, Duerr R, Minster J-B (2010) Data citation and peer-review. *Eos Trans* 91 (34):297–298. doi:[10.1029/2010EO340001](https://doi.org/10.1029/2010EO340001)
- Porter JH (2017) Scientific databases for environmental research, Chapter 3. In: Recknagel F, Michener W (eds) *Ecological informatics. Data management and knowledge discovery*. Springer, Heidelberg
- Reid WV, Chen D, Goldfarb L et al (2010) Earth system science for global sustainability: grand challenges. *Science* 330:916–917. doi:[10.1126/science.1196263](https://doi.org/10.1126/science.1196263)
- Ricciuto DM, Schaefer K, Thornton PE et al (2013) NACP site: terrestrial biosphere model and aggregated flux data in standard format. ORNL DAAC, Oak Ridge, TN. doi:[10.3334/ORNLDAAC/1183](https://doi.org/10.3334/ORNLDAAC/1183)
- Rüegg J, Gries C, Bond-Lamberty B et al (2014) Completing the data life cycle: using information management in macrosystems ecology research. *Front Ecol Environ* 12:24–30. doi:[10.1890/120375](https://doi.org/10.1890/120375)
- Schildhauer M (2017) Data integration: principles and practice, Chapter 8. In: Recknagel F, Michener W (eds) *Ecological informatics. Data management and knowledge discovery*. Springer, Heidelberg
- Scholes RJ (2005) SAFARI 2000 woody vegetation characteristics of Kalahari and Skukuza sites. ORNL DAAC, Oak Ridge, TN. doi:[10.3334/ORNLDAAC/777](https://doi.org/10.3334/ORNLDAAC/777)
- Starr J, Castro E, Crosas M et al (2015) Achieving human and machine accessibility of cited data in scholarly publications. *Peer J Comp Sci* 1:e1. doi:[10.7717/peerj-cs.1](https://doi.org/10.7717/peerj-cs.1)
- Strasser C, Cook RB, Michener WK et al (2012) Primer on data management: what you always wanted to know about data management, but were afraid to ask. California Digital Library. <http://dx.doi.org/doi:10.5060/D2251G48>
- Tenopir C, Allard S, Douglass K et al (2011) Data sharing by scientists: practices and perceptions. *PLoS One* 6:e21101. doi:[10.1371/journal.pone.0021101](https://doi.org/10.1371/journal.pone.0021101)



- Thornton PE, Thornton MM, Mayer BW et al (2017) Daymet: daily surface weather data on a 1-km grid for North America, Version 3. ORNL DAAC, Oak Ridge, TN. doi:[10.3334/ORNLDAAC/1328](https://doi.org/10.3334/ORNLDAAC/1328)
- UCAR (University Corporation for Atmospheric Research) (2016) UDUNITS. <http://www.unidata.ucar.edu/software/udunits/>. Accessed 18 Apr 2016
- USGEO (US Group on Earth Observation) (2015) Common framework for earth – observation data. US Group on Earth Observation, Data Management Working Group, Office of Science and Technology Policy. [https://www.whitehouse.gov/sites/default/files/microsites/ostp/common\\_framework\\_for\\_earth\\_observation\\_data\\_draft\\_120215pdf](https://www.whitehouse.gov/sites/default/files/microsites/ostp/common_framework_for_earth_observation_data_draft_120215pdf). Accessed 25 Jan 2015
- Whitlock MC (2011) Data archiving in ecology and evolution: best practices. *Trends Ecol Evol* 26 (2):61–65. doi:[10.1016/j.tree.2010.11.006](https://doi.org/10.1016/j.tree.2010.11.006)

# Chapter 7

## Data Discovery

William K. Michener

**Abstract** Data may be discovered by searching commercially available internet search engines, institutional and public repositories, online data directories, and the content exposed by data aggregators. Chapter 7 describes these various search approaches and presents seven best practices that can promote data discovery and reuse. It further emphasizes the need for data products to be uniquely identifiable and attributable to the data originators who must also be uniquely identifiable.

### 7.1 Introduction

Data discovery is the act of searching for and finding data that are or may be of particular interest. Prior to the advent of the Internet and World Wide Web, data discovery was often a difficult and laborious process. Researchers discovered that data existed via word-of-mouth and conference presentations as well as through the published literature. Accessing or acquiring such “found” data was often even more difficult as data sharing has only begun to become the norm over the past two to three decades (Michener 2015).

It is presently much easier to discover data. An array of Internet search engines, data repositories, data directories, and data aggregators have been created to facilitate data and information discovery and provide other services. This chapter describes the various tools and approaches that are most commonly used today to discover specific data products (Sect. 7.2) and best practices for promoting data discovery and use (Sect. 7.3).

---

W.K. Michener (✉)  
University of New Mexico, Albuquerque, NM, USA  
e-mail: [william.michener@gmail.com](mailto:william.michener@gmail.com)

## 7.2 Discovering Data Created by Others

The various tools and approaches that are used to discover data differ widely in their efficacy at precisely finding the data that one is hoping to examine or acquire. Commercial Internet search engines are very effective at discovering web sites and web pages that mention research projects and publications resulting from those studies. For instance, Google Scholar is particularly adept at enabling users to discover publications related to a particular topic and that may include descriptions of data collection and analytical methods. Internet search engines are often less useful for precisely discovering data that are held in institutional and public repositories as such data may be insufficiently described, hidden behind institutional firewalls, or disambiguated from their associated metadata.

Searches of institutional and public repositories may quickly lead one to particular data products as such repositories often provide search tools that are tailored to facilitate searches of their data holdings. It may, however, prove challenging to identify the specific repository where one invests the time and effort in conducting individual searches of the repository holdings. Data directories help address this challenge by enabling one to search for particular data by keywords and then be pointed to specific databases or repositories that may be linked to the online data directory; the user may then be directed to a particular repository where the data and metadata can be further examined. Data aggregators are increasingly being developed to provide a mechanism to discover data that originate from many different sources (e.g., individuals, repositories, and institutions such as museums and research networks). Data aggregators often provide additional value-added services and products such as quality assurance, metadata checks, and access to analytical and visualization tools. The different approaches to data discovery and relevant examples are described below.

### 7.2.1 *Internet Search Engines*

Internet search engines are commonly used to search for information, publications, data and other content that is available on web sites that are part of the World Wide Web. Some of the more commonly employed general search engines include Google, Bing, Yahoo! and Baidu. Internet search engines work by: (1) retrieving information about web pages by routinely visiting web sites (i.e., web crawling); (2) indexing the information that is retrieved such as titles and page content based on HTML markup of the content; and (3) allowing users to query the indexed content based on one or more keywords that the user enters. Different search engines use different and, typically, proprietary approaches and algorithms for indexing and caching (i.e., storing content for rapid access and processing) web content. Google's search engine, for example, is based on an algorithm that ranks web pages based on the number and rank of the web sites and pages that link to

them. Commercial internet search engines may derive income by assigning higher rankings to pages from web sites that pay to have their content prioritized in searches, by allowing paid advertisements to appear alongside search results, or by both approaches. Search and indexing algorithms may filter and preferentially rank results based on user characteristics such as location and prior user search history.

The usefulness of a web search engine is related to how relevant results are to the user. For instance, a user will often enter one or more search terms or keywords such as “primary production” and retrieve millions of results. Some internet search engines offer various ways to filter large lists of results and more precisely find desired content. Such approaches include specifying a particular range of dates (e.g., retrieving results that are from the current calendar year) and employing Boolean operators (i.e., AND, OR, and NOT) to refine the query. Using the previous example, after retrieving a daunting list of results after querying “primary production,” a user may refine the search by entering “primary production AND grassland AND data” to more precisely discover content of interest. Nevertheless, such a search may still retrieve millions of results—some that point to specific data products, others that point to publications dealing with topics such as how to calculate net primary productivity from biomass data and a multitude of other related issues.

### **7.2.2 Data Repositories**

Numerous data repositories exist worldwide that hold ecological and environmental data. The Registry of Research Data Repositories (also known as [re3data.org](http://re3data.org); [re3data.org](http://re3data.org) Project Consortium 2016) is a global registry of research data repositories where one can search for and discover relevant research data repositories from various academic disciplines. [re3data.org](http://re3data.org) lists hundreds of repositories and includes many data directories and data aggregators. The listed repositories vary widely in size and scope from archives such as the Macaulay Library<sup>1</sup> which is the largest scientific archive of biodiversity audio and video recordings collected worldwide (Cornell University 2016), to the HJ Andrews Experimental Forest<sup>2</sup> which hosts ecological, environmental and related research data that are primarily associated with a large forest research site in Oregon’s Cascades Mountains in the U.S. Pacific Northwest (Andrews Experimental Forest LTER 2016), to the Dryad

---

<sup>1</sup>[re3data.org](http://re3data.org): Macaulay Library; editing status 2014-06-25; [re3data.org](http://re3data.org)—Registry of Research Data Repositories. <http://doi.org/10.17616/R3CS4N> last accessed: 2016-01-14.

<sup>2</sup>[re3data.org](http://re3data.org): HJ Andrews Experimental Forest; editing status 2015-05-28; [re3data.org](http://re3data.org)—Registry of Research Data Repositories. <http://doi.org/10.17616/R3591T> last accessed: 2016-01-14.

Digital Repository<sup>3</sup> which is a large general purpose, international, curated archive that holds data that underlie scientific and medical publications from hundreds of journals, professional societies and publishers (Dryad 2016).

In addition to variable size and scope, data repositories offer different approaches for discovering and acquiring data. HJ Andrews Experimental Forest data, for example, may be searched by: (1) using a simple “string search” where a word or phrase is entered; (2) “advanced search” where one can specify data associated with a particular researcher, a subset of theme keywords selected from a list and specific study sites that are also selected from a list; or by (3) browsing the list of all data products. Once one has identified data of interest, the metadata and other descriptive information can typically be downloaded immediately, but acquisition of the data requires that one register as a user and state the purpose for which the data are being requested.

Biodiversity audio and video recordings can be easily discovered in the Macaulay Library by searching catalog numbers or common names or species names of the organism(s) of interest. A search for “bluebird” generates a web page that includes a listing of available audio and video recordings and other information about the recordings including links to most of the recordings so they may be listened to or viewed. Acquiring the recordings requires that one license the media and place an order for the recordings that includes catalog number and/or species, a description of the recording, requested data format, and delivery details; use for research and education purposes is free, but commercial and other users may be required to pay a license fee and studio fee for preparing the media.

Data may be discovered in the Dryad repository via several mechanisms including simple text string search or a more advanced search that allows the user to narrow the results set by title, author, subject, publication date and publication name. For example, one may search for “wood density” and then narrow the search further by specifying “Ecology Letters” as the publication name which leads to a seminal paper by Chave and colleagues (2009) and the associated Dryad data package (Zanne et al. 2009); note that the lead author of the journal article and the Dryad data package are different individuals. The Dryad web page describes the contents of the data package (i.e., downloadable file names and file sizes, title, and other details) as well as links to the full metadata, the number of times the data package contents have been downloaded and instructions for citing both the journal article and the data package. The inclusion of a digital object identifier (DOI) in the data package citation makes it possible to easily link to the data from data package citations that are included in the Literature Cited sections of papers by other authors (e.g., Mascaro et al. 2012) that have cited the journal publication that is based on the data (i.e., Chave et al. 2009 in this case) and that have used and cited the data (i.e., Zanne et al. 2009).

---

<sup>3</sup>[re3data.org](http://re3data.org): DRYAD; editing status 2015-11-18; [re3data.org](http://re3data.org)—Registry of Research Data Repositories. <http://doi.org/10.17616/R34S33> last accessed: 2016-01-14.

### 7.2.3 *Data Directories*

The U.S. National Aeronautics and Space Administration's (NASA) Global Change Master Directory (GCMD) makes it easy for scientists and the public to discover and access data relevant to climate change (NASA 2016). The GCMD contains descriptions of tens of thousands of data sets from the Earth and environmental sciences. One can perform searches of science keywords (e.g., atmosphere, biosphere, oceans, paleoclimate), instruments (e.g., Earth remote sensing instruments, in situ/laboratory instruments), platforms (e.g., aircraft, Earth observation satellites, in situ ocean-based platforms), locations (e.g., continent, geographic region, vertical location), providers (e.g., academic, government agencies, non-government organizations), project name or acronym, and free text. Searches lead to records that include project titles and brief abstracts and the records individually link to the more complete metadata file and, frequently, to the data.

The GCMD also provides access to authoring tools and other services that data and service providers can use to describe and facilitate discovery of their data products. Keyword vocabularies are central to the GCMD search capability and provide "controlled" lists of keywords that are accepted by the broader scientific community. The vocabularies enable data providers to describe their data products using standardized terms and are continually being expanded and revised.

The GCMD enables research organizations and other partners to create portals that support discovery of the portion of the GCMD content that is associated with a particular organization or partner (e.g., Antarctic Master Directory, World Water Forum). The GCMD also serves as one of NASA's contributions to the international Committee on Earth Observation Satellites (CEOS), through which it is named the CEOS International Directory Network (IDN) Master Directory (CEOS 2016). The IDN Master Directory provides links to numerous GCMD-associated portals.

DataONE is another related type of service that supports discovery of Earth and environmental science data (DataONE 2016). DataONE harvests and indexes metadata from a large international network of data repositories (Michener et al. 2011, 2012). It provides direct links to 100 s of thousands of data products that are stored in various repositories worldwide. To discover data, a researcher typically enters a keyword or phrase (e.g., "primary productivity") in the DataONE search bar which links to a visual display that enumerates the number of data products that exist in different geographic regions worldwide and includes more advanced search capabilities. The user can then easily narrow down the result set by searching for particular data attributes (e.g., density, length), repositories (e.g., data only from the Dryad Digital Repository), data creators, years, identifiers (e.g., DOIs), taxa (e.g., class, family), and locations (Fig. 7.1).

The screenshot displays the DataONE search interface. At the top left is the DataONE logo. A navigation bar includes links for About, News, Participate, Resources, Education, and Data. Below this, the search bar contains 'DATAONE SEARCH: Search' and a 'Go' button. A 'Clear all filters' button is visible. The search results are filtered by 'primary productivity'. A 'Filter by:' section lists various attributes like Data attribute, Data files, Member Node, Creator, Year, Identifier, Taxon, and Location. The results list three datasets with their titles, authors, and URLs. A map on the right shows the geographic distribution of these datasets across North America, with colored overlays and numerical counts for each state or province.

**Search Criteria:**  
 Search phrase: [ ]  
 Datasets: 1 to 25 of 25,044  
 Sort by: Most recent  
 1 2 3 ... 1,002 Next

**Filter by:**  
 Data attribute  
 Data files  
 Member Node  
 Creator  
 Year  
 Identifier  
 Taxon  
 Location

**Search Results:**

- Dataset 1:** Steven Hamburg, Anne G. Rhoads, and Matt Vadeboncoeur. 2015. Leaf area index following the ice storm of January 1998 at the Hubbard Brook Experimental Forest. U.S. LTER Network. <https://pasta.lternet.edu/package/metadata/ml/knb-lter-hbr/4/6/6>.
- Dataset 2:** Karen Wright. 2015. Core Site Phenology Study from the Chihuahuan Desert Grassland and Shrubland at the Sevilleta National Wildlife Refuge, New Mexico (2000-present ). U.S. LTER Network. <https://pasta.lternet.edu/package/metadata/ml/knb-lter-sev/137/208261>.
- Dataset 3:** Andrew Burton, Jerry Melillo, and Serita Frey. 2010. Root and Mycorrhizal Respiration at Harvard Forest Soil Warming Experiments since 2007. U.S. LTER Network. <https://pasta.lternet.edu/package/metadata/ml/knb-lter-hfr/1/1/7>.

**Map Distribution:**  
 The map shows data points across North America. Key counts include: Canada (8), BC (7), AB (2), SK (15), MB (3), SD (186), ND (3565), IA (708), MO (1083), WI (2428), IL (4), IN (3), OH (1391), WV (486), VA (1), NC (1), TN (106), KY (1707), LA (1632), AL (9), MS (2855), GA (23), SC (14), PA (655), NY (1), CT (9), VT (3), NH (3), ME (3), VT (3), NH (3), ME (3), VT (3), NH (3), ME (3).

**Footer:**  
 DataONE is a collaboration among many partner organizations, and is funded by the US National Science Foundation (NSF) under a Cooperative Agreement. Acknowledgement: This material is based upon work supported by the National Science Foundation under Grant Numbers 0809944 and 1435608. Disclaimer: Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Fig. 7.1 The DataONE search interface showing search criteria (left), datasets matching “primary productivity” (center) and distribution of data sets in portions of North America (right)

### 7.2.4 Data Aggregators

Data aggregation is the process whereby data are gathered from multiple sources and then, typically, presented in a standardized format to users. Some data aggregators perform minimal or no additional processing of the data whereas others provide numerous value-added services to benefit users. Value-added services can include data reformatting, performing quality assurance checks (e.g., duplicate detection, invalid entries), adding taxonomic or geographical location information, and providing statistical and graphical summaries. Those aggregators that provide significant value-added services often work with specific types of data such as meteorological data or data pertaining to particular groups of organisms.

Data aggregators are typically grouped in with data repositories (e.g., listed in the Registry of Research Data Repositories; [re3data.org](https://re3data.org) Project Consortium 2016) although they differ from most repositories with respect to the services provided. For instance, a typical institutional data repository may archive a wide range of data from a large number of contributors; services may be limited to activities such as the provision of a metadata entry tool, addition of a DOI, citation recommendations, and periodic backup. A data aggregator, on the other hand, may accept limited types of data that are in one or a small number of specific formats; the aggregator may then further process the data by summarizing the data, adding additional metadata descriptors, and so on. The examples below highlight a subset of non-commercial data aggregators indicating the types of data they aggregate and some of the services that are provided.

*The Atlas of Living Australia* (ALA 2016) is an online repository that contains data and information about Australia's plants, animals and microbes (e.g., species occurrence records, photos, sound recordings, maps, molecular data and links to pertinent literature). ALA aggregates records and datasets submitted from thousands of sources including citizens, governmental agencies and other groups. It provides access to keys as well as tools that enable data and metadata to be entered in standardized formats. In addition, a variety of value added services and features are provided including: (1) a spatial portal that allows one to view and create maps that show species occurrences relative to climate and numerous other features; (2) "fishmap" which allows one to find Australia's marine fishes; (3) "Explore Your Area" which allows one to see all species within a user-specified radius of your home location (Fig. 7.2); and (4) a "dashboard" that provides updates on numbers of occurrence records and datasets, records submitted by institution/data provider, conservation status, and numbers of records by state and territory, date and taxonomic grouping.

*The Advanced Ecological Knowledge and Observation System (ÆKOS) Data Portal* (ÆKOS 2016) is a data portal that allows one to discover data about Australian plants, animals and their environment (Fig. 7.3). The portal provides detailed information about the research methods employed to facilitate understanding and reuse of the data; such information is associated with various icons that



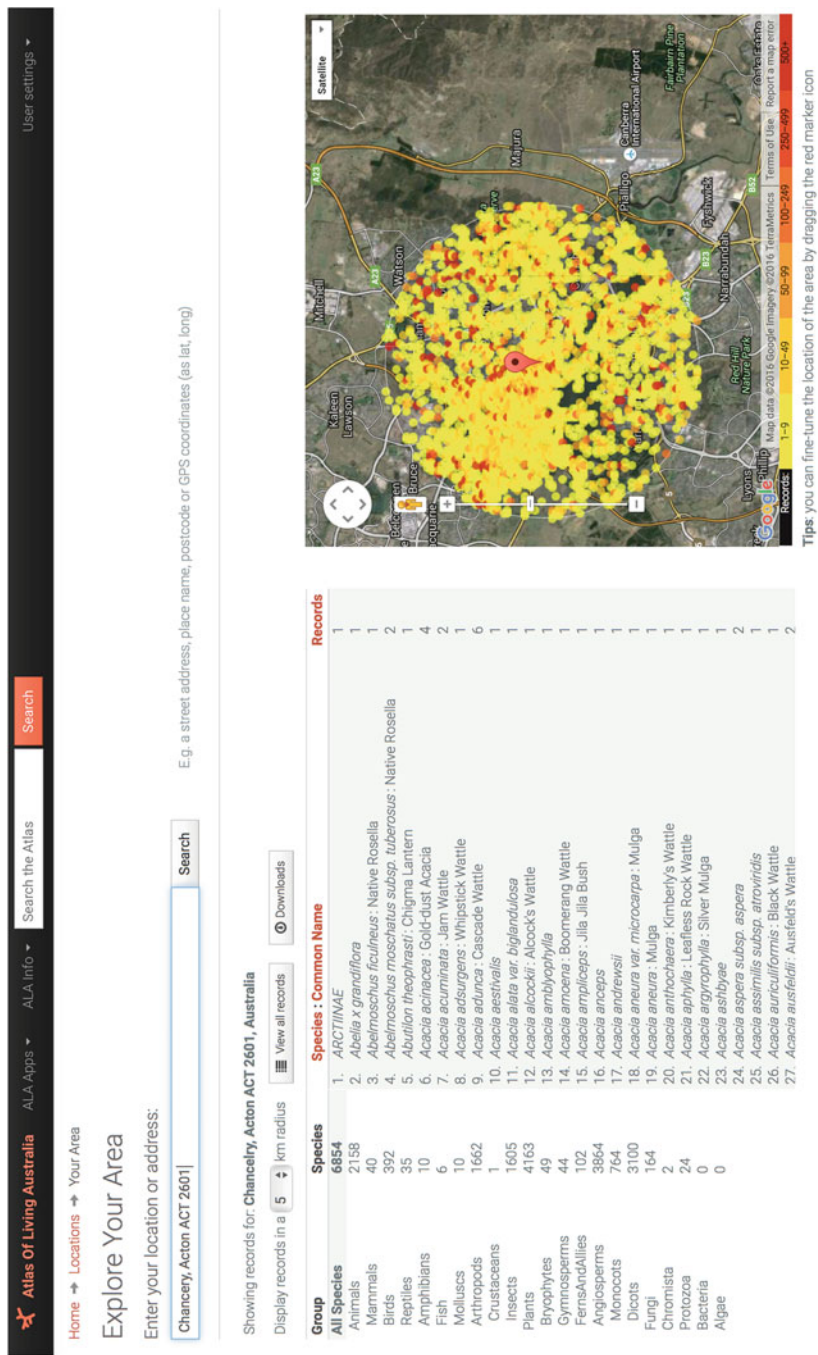


Fig. 7.2 The Atlas of Living Australia repository website illustrating the tool “Explore Your Area” which lists and pinpoints the locations of species within a user-specified radius of a particular location.

The screenshot displays the Aekos Data Portal interface. At the top left is the Aekos logo with the tagline "AUSTRALIAN ECOLOGICAL KNOWLEDGE AND OBSERVATION SYSTEM". The main header "Data Portal" is prominently displayed. A navigation bar includes links for "Help", "What's new", "Feedback", "Legals", "Sponsors", and "v2.4.9", along with a "Data Cart" button. Below the header, a "Search Expression" box contains the text "Free Text free text search of 'Daintree'", with a "Redo Search" button. A map of Queensland, Australia, shows the location of the search results. A legend for the map includes "Base Layer" (Google Physical, Google Satellite, Overlays), "IBRA Regions", "River Basins", and "Drainage". The search results are listed in three panels, each showing a thumbnail map, a Creative Commons license, and a list of icons representing different data layers. The first result is for "Queensland CORVEG Database (DSITI), ID: 26882", with details: Author(s): Queensland Herbarium, Landform: Mountains, Establishment date: 24/01/2003, Number of visits: 1, Location: IBRA Wet Tropics, (Daintree-Bloomfield), Queensland, (-16.51214, 1, Main sampling unit: quadrat. The second result is for "Queensland CORVEG Database (DSITI), ID: 39635", with details: Author(s): Queensland Herbarium, Landform: Alluvial plain, Establishment date: 09/07/2011, Number of visits: 1, Location: IBRA Wet Tropics, (Daintree-Bloomfield), Queensland, (-16.22614, 1, Main sampling unit: quadrat. The third result is for "Queensland CORVEG Database (DSITI), ID: 39637", with details: Author(s): Queensland Herbarium. A "Data Cart" button is visible at the bottom right of the results area, and a page indicator shows "1 - 20 of 45".

Fig. 7.3 The result set obtained by performing a simple search for “Daintree” of the Advanced Ecological Knowledge and Observation System (ÆKOS) Data Portal. Each resulting dataset (*right panel*) is accompanied by icons that represent conditions of use, the types of variables included in the data, duration of the study, and research methods employed

accompany each data set discovered during a user's search. The portal includes numerous features and services that support researchers, educators and resource managers. One can search by location and ecological data themes, and create complex Boolean searches. Figure 7.3 illustrates the result set obtained by performing a simple search for "Daintree." Each resulting dataset is accompanied by icons that represent conditions of use, the types of variables included in the data, duration of the study, and research methods employed. By selecting "More Details", one is taken to a webpage where: (1) an "Observation Diagram" provides a visual representation of the types of observations that are recorded; (2) a "Methods Diagram" that similarly illustrates the sampling methods that are employed with links to methodological details; and (3) "Metadata" where detailed information about all aspects of the data is available. If the data appear suitable, then a user can easily download the data in .csv format.

*VertNet* (2016) aggregates a wide variety of vertebrate biodiversity data from natural history collections worldwide and provides tools that facilitate data discovery, acquisition and publication (Constable et al. 2010; Guralnick and Constable 2010). *VertNet* has integrated, standardized and "cleaned" data derived from previously existing vertebrate data aggregators [i.e., Mammal Networked Information System (MaNIS 2016); Ornithological Information System (ORNIS 2016); HerpNET (2016); and FishNet2 (2016)]. *VertNet* supports publication, indexing, and georeferencing of data and provides training as well as a clear and concise set of norms for data use and publication.

### 7.3 Best Practices for Promoting Data Discovery and Reuse

Data discovery and reuse are most easily accomplished when: (1) data are logically and clearly organized; (2) data quality is assured; (3) data are preserved and discoverable via an open data repository; (4) data are accompanied by comprehensive metadata; (5) algorithms and code used to create data products are readily available; (6) data products can be uniquely identified and associated with specific data originator(s); and (7) the data originator(s) or data repository have provided recommendations for citation of the data product(s). Data organization, data quality, metadata and data preservation were discussed in detail in Porter (2017), Michener (2017a, b) and Cook et al. (2017), respectively.

Good data archiving and sharing policies promote long-term discoverability and accessibility of data and do so in a way that benefits both the data producers and consumers (Duke and Porter 2013; Whitlock et al. 2016). The following discussion focuses on simple steps that can be taken to ensure that data products and scientific code can be easily discovered, reused and cited.

### 7.3.1 Data Products

*Data Products Should Be Uniquely Identifiable and Attributable to Their Originators* “Consumers” or users of data benefit from knowing that a data product exists, that it can be used and cited, that the data originators receive proper attribution, and that others can subsequently discover and use the same data product (e.g., for research transparency and data verification purposes). “Producers” or originators of data benefit from having their data products cited and used by others much like the peer-recognition that is associated with having publications cited by others in the literature.

Persistent Identifiers (PIDs) have emerged as the principal mechanisms to provide a long-lasting reference to datasets and other digital resources. PIDs make it easy to uniquely cite and access research data and other digital resources. Some of the more common PIDs include Archival Resource Keys (ARKs), Digital Object Identifiers (DOIs), Life Science Identifiers (LSIs), and Universal Resource Names (URNs). DOIs are increasingly becoming the norm for citing all types of digital resources and various organizations have emerged to facilitate the creation and management of DOIs. Crossref (2016), for example, commonly provides DOIs for journal articles, books, reports and datasets.

Likewise, DataCite (2016) creates and supports standards for PIDs for data and other digital resources. DataCite member institutions are globally distributed data centers, national libraries, universities and other organizations that serve users by assigning DOIs to data and other objects. DataCite provides specific recommendations for how to cite data (Box 7.1). DataCite also provides various tools for users such as (1) DOI Citation Formatter which creates different citation formats for DataCite and Crossref DOIs; (2) Metadata Search Tool that allows one to search the metadata of datasets registered with DataCite; and (3) Metadata Stats service that provides statistics on datasets that have been uploaded and accessed.

#### **Box 7.1 DataCite Recommendations for Data Citation**

The DataCite “recommended format for data citation is as follows:

Creator (PublicationYear): Title. Publisher. Identifier.

It may also be desirable to include information about two optional properties, Version and ResourceType (as appropriate). If so, the recommended form is as follows:

Creator (PublicationYear): Title. Version. Publisher. ResourceType. Identifier”.

(<http://www.datacite.org/>); accessed 20 Jan 2016).

Many data repositories work with DataCite member institutions to assign DOIs and also provide specific guidelines for citing datasets that are housed in their repository. For example, the dataset citation recommendations for the Dryad Digital Repository are listed in Box 7.2.

### **Box 7.2 Dryad Digital Repository Data Citation Recommendations**

“When referencing data in the text, we recommend the following as a template (substitute your DOI suffix for the xxxxx):

Data available from the Dryad Digital Repository: <http://dx.doi.org/10.5061/dryad.xxxxx>

In the Bibliography, we recommend a citation similar to:

Heneghan C, Thompson M, Billingsley M, Cohen, D (2011) Data from: Medical-device recalls in the UK and the device-regulation process: retrospective review of safety notices and alerts. Dryad Digital Repository. <http://dx.doi.org/10.5061/dryad.585t4>”.

(<https://datadryad.org/pages/faq>; accessed 20 Jan 2016).

*Data Originators and Users Should Be Uniquely Identifiable* It is highly unlikely that any given personal name can be resolved to a single individual; further, some common names like “J. Smith” may be associated with thousands of individuals and many researchers undergo name changes over the course of their careers (e.g., through marriage). This situation presents a real challenge when the goal is to make sure that individuals receive proper attribution for the output of their scholarly and research endeavors. ORCID (Open Research and Contributor ID; ORCID 2016) provides a valuable service that enables researchers to be uniquely identified. ORCID identifiers are unique alphanumeric codes that resolve to a specific individual and can be easily linked to publications, grant proposals, and other outputs and activities. The ORCID organization maintains a registry of unique researcher identifiers and supports mechanisms that enable researchers to link their identifiers to research products. Increasingly, research sponsors and publishers are encouraging or requiring that individuals associate their works with an ORCID identifier.

### **7.3.2 Scientific Code**

Scientific code such as custom software and scripts (e.g., R, Matlab) is used in statistical and graphical analysis, modeling, detecting and correcting errors in data, and creating figures and visualizations. Code precisely records what has been done with the data and the availability of code makes it possible for other scientists to more easily understand and, potentially, reproduce data processing and analytical steps (Maslan et al. 2016; Peng 2011; Barnes 2010; Ince et al. 2012). It is good practice to deposit scientific code in long-term repositories such as Dryad, Figshare, PANGAEA, or Zenodo that provide licenses (e.g. CC0, CC-By) and that assign DOIs so that code is preserved and may be used and properly cited by others (Maslan et al. 2016).

## References

- ÆKOS (2016) ÆKOS: Advanced Ecological Knowledge and Observation System Data Portal. <http://www.aekos.org.au/home>. Accessed 22 Apr 2016
- ALA (2016) Atlas of Living Australia. <http://www.ala.org.au>. Accessed 22 Apr 2016
- Andrews Experimental Forest LTER (2016) HJ Andrews experimental forest long term ecological research. <http://andrewsforest.oregonstate.edu>. Accessed 22 Apr 2016
- Barnes N (2010) Publish your computer code: it is good enough. *Nature* 467:753
- CEOS (2016) CEOS Committee on Earth Observation Satellites. <http://ceos.org/ourwork/workinggroups/wgiss/current-activities/jdn/>. Accessed 22 Apr 2016
- Chave J, Coomes DA, Jansen S et al (2009) Towards a worldwide wood economics spectrum. *Ecol Lett* 12:351–366. doi:10.1111/j.1461-0248.2009.01285.x
- Constable H, Guralnick R, Wieczorek J et al (2010) VertNet: a new model for biodiversity data sharing. *PLoS Biol* 8:e1000309. doi:10.1371/journal.pbio.1000309
- Cook RB, Wie Y, Hook LA et al (2017) Preserve: protecting data for long-term use, Chapter 6. In: Recknagel F, Michener W (eds) *Ecological informatics. Data management and knowledge discovery*. Springer, Heidelberg
- Cornell University (2016) The Cornell Lab of Ornithology Macaulay Library. <http://macaulaylibrary.org/>. Accessed 22 Apr 2016
- Crossref (2016) Crossref.org. <http://www.crossref.org>. Accessed 22 Apr 2016
- DataCite (2016) DataCite. <https://www.datacite.org>. Accessed 22 Apr 2016
- DataONE (2016) DataONE: Data Observation Network for Earth. <http://dataone.org>. Accessed 22 Apr 2016
- Dryad (2016) Dryad. <http://datadryad.org>. Accessed 22 Apr 2016
- Duke CS, Porter JH (2013) The ethics of data sharing and reuse in biology. *BioSci* 63:483–489
- FishNet2 (2016) FishNet2. <http://www.fishnet2.net/aboutFishNet.html>. Accessed 22 Apr 2016
- Guralnick R, Constable H (2010) VertNet: creating a data-sharing community. *BioSci* 60:258–259. doi:10.1525/bio.2010.60.4.2
- HerpNet (2016) HerpNet. <http://herpnet.org>. Accessed 22 Apr 2016
- Ince DC, Hatton L, Graham-Cumming J (2012) The case for open computer programs. *Nature* 482:485–488
- MaNIS (2016) MaNIS: Mammal Networked Information System. <http://manisnet.org>. Accessed 22 Apr 2016
- Mascaro J, Hughes RF, Schnitzer SA (2012) Novel forests maintain ecosystem processes after the decline of native tree species. *Ecol Monogr* 82:221–228
- Maslan KAS, Heer JM, White EP (2016) Elevating the status of code in ecology. *Trends Ecol Evol* 31:4–7. doi:10.1016/j.tree.2015.11.006
- Michener WK (2015) Ecological data sharing. *Ecol Inf* 29:33–44. doi:10.1016/j.ecoinf.2015.06.010
- Michener WK (2017a) Quality assurance and quality control (QA/QC), Chapter 4. In: Recknagel F, Michener W (eds) *Ecological informatics. Data management and knowledge discovery*. Springer, Heidelberg
- Michener WK (2017b) Creating and managing metadata, Chapter 5. In: Recknagel F, Michener W (eds) *Ecological informatics. Data management and knowledge discovery*. Springer, Heidelberg
- Michener W, Vieglais D, Vision T et al (2011) DataONE: Data Observation Network for Earth – preserving data and enabling innovation in the biological and environmental sciences. *D-Lib Mag* 17 (Jan/Feb 2011). doi:10.1045/january2011-michener
- Michener WK, Allard S, Budden A et al (2012) Participatory design of DataONE—enabling cyberinfrastructure for the biological and environmental sciences. *Ecol Inf* 11:5–15
- NASA (2016) NASA Global Change Master Directory. <http://gcmd.nasa.gov/>. Accessed 22 Apr 2016
- ORCID (2016) ORCID. <http://orcid.org>. Accessed 22 Apr 2016

- ORNIS (2016) ORNIS. <http://www.ornisnet.org>. Accessed 22 Apr 2016
- Peng RD (2011) Reproducible research in computational science. *Science* 334:1226–1227
- Porter JH (2017) Scientific databases for environmental research, Chapter 3. In: Recknagel F, Michener W (eds) *Ecological informatics. Data management and knowledge discovery*. Springer, Heidelberg
- re3data.org Project Consortium (2016) re3data.org Registry of Research Data Repositories. <http://www.re3data.org>. Accessed 22 Apr 2016
- VertNet (2016) VertNet. <http://www.vertnet.org/>. Accessed 22 Apr 2016
- Whitlock MC, Bronstein JL, Bruna EM et al (2016) A balanced data archiving policy for long-term studies. *Trends Ecol Evol* 31(2):84–85. doi:[10.1016/j.tree.2015.12.001](https://doi.org/10.1016/j.tree.2015.12.001)
- Zanne AE, Lopez-Gonzalez G, Coomes DA, et al (2009) Data from: towards a worldwide wood economics spectrum. Dryad Digital Repository. doi:[10.5061/dryad.234](https://doi.org/10.5061/dryad.234)



# Chapter 8

## Data Integration: Principles and Practice

Mark Schildhauer

**Abstract** Data integration is the process of combining (also called “merging” or “joining”) data together to create a single unified data object from what were multiple, distinct data objects. The motivation for integrating data is usually to bring together the information needed to jointly analyze or model some phenomena. By producing a single, consistently structured object through data integration, the process of further manipulating those data is vastly simplified, while presumed relationships among the data are clarified.

Data integration is essential for many scientific disciplines, but especially in disciplines such as ecology and the environmental sciences, where processes and patterns of interest often emerge from interactions among numerous complex physical phenomena. Observations of these distinct phenomena are often collected by disparate parties in uncoordinated ways, using different data systems. It is then necessary to gather these data together and appropriately integrate them, to clarify through further modeling and analysis the nature and strength of any relationships among them. Synthesis studies, in particular, often require finding, and then bringing together disparate data in order to integrate them, and reveal new insights.

This chapter describes aspects of data that are critical for determining whether and how data can be integrated, and discusses some of the theoretical considerations and common mechanisms for integrating data.

### 8.1 Introduction

Data integration is the process of combining, merging, or joining data together, in order to make what were distinct, multiple data objects, into a single, unified data object. Data integration is one of the most fundamental operations that researchers and analysts typically must master, as many interesting scientific questions can be investigated only if the data needed to address them are assembled together. A typical motivation for data integration is to bring together data of similar or

---

M. Schildhauer (✉)  
National Center for Ecological Analysis and Synthesis, University of California, Santa Barbara, CA, USA  
e-mail: [schild@nceas.ucsb.edu](mailto:schild@nceas.ucsb.edu)



complementary kinds, in order to better inform models and analyses about phenomena of interest. Data are also integrated in order to discover or test whether potential relationships exist among the pieces, or to expand the spatial, temporal, and thematic ranges over which scientists can explore potential relationships.

There are also many situations in scientific research where experiments are carried out in a controlled environment by manipulating a well-defined set of “variables” so that variability among unmeasured features is considered to be unimportant relative to the outcome of interest. Such experimental studies provide a strong inferential basis for testing and refining hypotheses, particularly when the processes of interest are driven by relatively few factors in strongly determinate ways. Many of the advances in molecular biology over the past decades, for example, have been gained through carefully controlled experiments based on sets of mutually exclusive hypotheses (Platt 1964). The data supporting these types of research is often fully self-contained with minimal need for further integration with other data.

While highly controlled experiments such as those commonly found in laboratory work can advance scientific understanding in powerful ways, studies of the natural environment are often not amenable to such designs as many factors—climate, topography, hydrology, soil composition, or the abundance and identity of surrounding organisms (including humans!)—cannot all be simultaneously manipulated or controlled. Yet these factors may be significant in structuring processes at the ecological population, community, and ecosystems levels. In essence, ecological systems can be highly complex, involving extensive interactions and feedbacks among many potentially critical factors. This leads to a dilemma as to how scientists can collect data on all these potentially critical factors, given the relatively limited resources that individual researchers, and even research teams or projects, typically have available to record observations of *all the potential factors* that could influence the patterns and processes shaping the environment at multiple scales.

Many field sciences (ecology, geology, oceanography, etc.) especially benefit from having additional data available from the particular spatiotemporal region where their observations were acquired: one might combine lists of bird species collected by different individuals from the same area to gain a more comprehensive record of the local avifauna; or one might merge data on tree growth rates with data sets about air temperature, soil type, and precipitation, that were collected from the same geospatial region over the same time period.

The need for data integration is thus pervasive throughout much of the environmental and earth sciences. Indeed, data integration is one of the principal activities involved in doing synthetic analyses. Much of the success of ecological synthesis studies over the past few decades can be attributed to arduous but fruitful efforts that brought together “existing data” to generate novel insights, activities that were often facilitated by synthesis centers that support knowledge and data integration activities (Carpenter et al. 2009). Regardless of the motivation, however, data integration requires that researchers understand structural and semantic aspects of

the data, as well gain some mastery of the mechanics involved in transforming and manipulating data to produce appropriate and re-usable integrated data products.

In this chapter we describe the basic concepts and mechanisms of data integration. As there are many types of specialized data formats and these can require special considerations based on the nature of the phenomena that the data “represent”, we focus here on understanding the structure and mechanics of dealing with *tabular data*. While we will not cover the technical details of dealing with other common data types conventionally stored in other formats—e.g., gene sequence data, or raster or vector formatted geospatial data—many of the conceptual and operational aspects of working with data described here pertain to those formats as well.

## 8.2 Essential Characteristics of All Data

Data integration can take many forms depending on the characteristics of the data to be integrated. There are three essential characteristics of all data, however, that should always be considered by a researcher when planning for data collection, as well as for strategizing how to integrate data. These characteristics are the **semantics** (or *meaning*) of the data, the **structure** (*form* or *format* in which the data are documented or stored) of the data, and finally the choice of **syntax** (*specific programming language* or *application*) used to define, store, query, and retrieve the data.

Many scientists take for granted the meaning and structure of their data because they are intimately acquainted with the theories and methods for investigating their phenomena of interest, the conventional terms used to describe those phenomena, as well as various specific tools for acquiring, storing, and manipulating data about those phenomena. However, this can lead to highly idiosyncratic modeling of data, structured in ways that are only comprehensible to the data creator, and lacking in documentation that would enable others to understand the content or structure of the data. This lack of standardization in terminology and modeling is increasingly problematic in an era where the benefits of preserving and documenting data for interpretation and re-use by others is becoming more and more critical (see Cook et al. 2017). Data can serve several purposes, especially in the environmental and earth sciences: as records of the state of diverse natural phenomena at various times and places; as the empirical basis supporting important research findings and hence “reproducible science”; and as source information having second or third “lives” when re-purposed and integrated into synthesis research activities (Hampton et al. 2015). For these reasons, scientists are increasingly taking greater care in planning the structure of their data, and documenting data contents as well, using available tutorials and published best practices (Michener 2015; Strasser et al. 2014).

The concept of metadata—explored in depth in Chap. 5—was alien and confusing to many researchers as recently as a decade ago. As of 2017, however, most scientists are now aware of the importance of providing critical additional information (*metadata*: “higher order, beyond” data) about their data, if it was not already explicated within their original data structure and documentation. With the continuing growth of Internet technologies and computational power, the potential to integrate and jointly analyze data of extremely high volumes or dimensions has never been greater, and is catalyzing a growing culture of collaboration, synthesis, and data-sharing throughout many of the earth, environmental, and social sciences. But data integration can only proceed with ease and confidence if the data are well described and well structured. This means that adequate documentation or metadata in some form, is critical—describing the *semantic content* of the data (what the data are about); and that the data are organized using standard *structures* (how they are stored), rather than idiosyncratic, custom, or inappropriate ones. Following best practices in defining and structuring one’s data will make them discoverable, accessible and amenable for further integration and analyses *by computers*, rather than arduous, manual methods such as copying and pasting data across worksheets. Finally, the choice of *syntax* by which the data are created and manipulated/integrated, is important. Although there may be many variations of computer language syntax for defining and describing data, there are also many commonalities that we outline in this chapter.

Note that every analytical framework approaches data integration using its own syntax and, to some extent, with specialized conceptual models of the data as well. In the case of tabular data, however, these all conform to a great degree to the *relational model* of data developed by Edgar Codd in 1969, with syntactical representation for creating and modifying relational data through Structured Query Language, or SQL (pronounced as “S-Q-L” or “sequel”) (Codd 2000; also see [https://en.wikipedia.org/wiki/Relational\\_model](https://en.wikipedia.org/wiki/Relational_model); <https://en.wikipedia.org/wiki/SQL>). As SQL is also an ANSI and ISO standard, we primarily frame our data integration examples using SQL syntax. Data analysis software such as “R”, Matlab, and SAS have similar but often more idiosyncratic and less standard terminologies than SQL for describing data integration operations, depending on which specific packages, modules, or libraries of those frameworks one uses. However, there are often close analogues to SQL syntax for manipulating data in each of these frameworks. For these reasons, understanding some basic SQL and the relational data model is very useful to any analyst who will be doing lots of data creation, manipulation, or integration.

### 8.3 Data as Records About Reality

Scientists today use many sensors and instruments to collect their data, but in the not too distant past, individuals mainly used their own keen senses—visual, auditory, tactile—to describe and “measure” aspects of nature, and to record these as

observations. Of course, humans' direct sensory capabilities are highly limited (witness a dog's sense of hearing or smell relative to typical human beings), then filtered through our brains, where further interpretation of what was observed (whether real or imaginary) is (cognitively) conceptualized, and recollected or recorded as "data". Our everyday experience is largely based on such interpretations of sensory-based inputs about physical reality, whether scientifically informed or not. Scientists are specifically trained to be acutely aware or knowledgeable about various aspects of this reality, and typically describe and measure aspects of nature with greater rigor, striving to be more deeply perceptive, objective, and unbiased, compared with a naive observer. While it is debatable whether any observations can be fully objective or unbiased, one can point to science as an empirically-based "way of knowing" that has had unprecedented success in predicting and informing many aspects of the behavior of physical reality, compared with other methods, such as "blind faith" (not the band).

When data are used to support details about some natural event or occurrence, we also call such data "evidence". At this level, the very notion of identifying objects and processes in the natural world, and recording their characteristics and interactions with other objects and processes, touches on the domain of philosophy of science (Quine 1981; Taper and Lele 2004). We routinely draw distinctions among objects in the world, ascribe various characteristics to them, and then group these objects into sets and types (e.g., biomes, chairs, fish, and clouds), based on instances (or individuals) that actually occur and that we have measured or observed. For example, there is a notion of a chair, and there is also the instance of chair that you may be sitting in as you read this. We create a notion of "chair" based on some functional characteristics to which we then attribute membership to individual instances. Is a "table" a "chair" when someone is sitting on it? These issues are brain-teasers, and are discussed largely in the realm of philosophy, specifically, the branch of metaphysics called "ontology"—that inquires as to the nature of reality or being. Nevertheless, such issues have relevance to science, for science and empiricism involve describing, measuring, and classifying ("typing") events, occurrences, or physical objects in the material world and their inter-relationships, so it is important to be thoughtful and critical about the basis for our understanding.

Most scientists probably subscribe to the perspective of "naturalistic realism" or "scientific realism", and do not worry much about the underlying basis in "reality" of the concepts and entities that they measure and describe (Hempel 1970), but there are reasons to be cautious about too much naiveté relative to the ontological status of all scientific phenomena (Kuhn 1996). Some reading about and reflection upon these issues may challenge scientists' confidence about the concreteness of their observations, and provoke greater reflection on the nature of the phenomena they are studying, as well as how they are "documenting" these phenomena as data. Fortunately, scientific conceptualizations of reality are continually challenged and amended by empirical observation and experiment, to the point where today we acknowledge that even our common day-to-day experience of space and time does not necessarily conform to some deeper reality, e.g., quantum entanglement (Greene 2005).

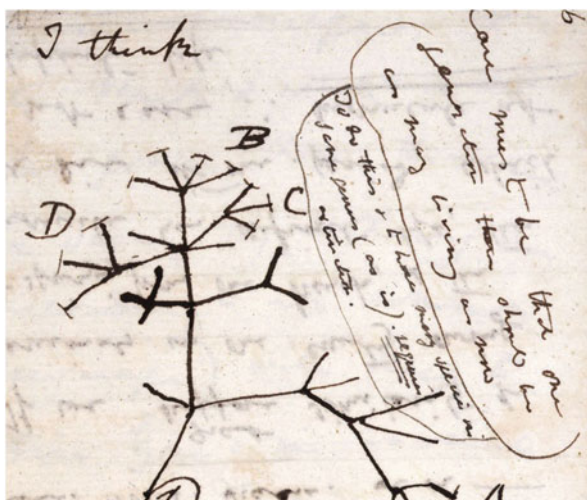
## 8.4 Record-Keeping and Prose Documents as Data Integration Challenges

Until the digital age, much “data” of scientific interest was collected in the form of notes, diaries and sketchbooks. The Codex of Leonardo da Vinci is an outstanding example of this form of scientific recordkeeping. These record-keeping formats typically involve lots of “natural language” descriptions (e.g., in prose English or Chinese)—but also include “richer media” such as illustrations or digital images with captions/descriptions, rough mockups of graphical trends (sometimes looking quite nice by using modern graphics software), and even areas of “well-structured data”—small tables or lists containing observations. These mechanisms for record-keeping are still useful for many field biologists, due to the expressiveness of natural language descriptions. Modern day field notes might also include descriptions of the human or environmental context surrounding the observations (meta-data), along with imagery, additional essential metadata (location, date, observer, etc.) and other pertinent information as “annotations” (Canfield 2011).

Indeed, many great scientific insights from “data integration” in the past did not involve the assistance of computers running analyses on well-structured data, but rather from “data” collected and stored in the form of these lab or field notebooks. The theory of natural selection, for example, emerged from Darwin’s integration of patterns perceived in nature, documented in field notebooks filled with observations of plants and animals, and their interactions and relationships (Fig. 8.1).

Today we can also consider these “natural language” documents as types of data that can be structured and integrated into a “corpus” which can be further mined or analyzed through “Natural Language Processing” (or NLP). The ability to rapidly automate the acquisition of large numbers of relevant text documents by scouring the Internet, and then organizing and “mining” these for patterns and relationships,

**Fig. 8.1** Page from Charles Darwin’s “First Notebook on ‘Transmutation of Species’” (1837)



is a relatively recent capability enabled by modern computational power. NLP is enabled by “Big Data” analysis and involves “integrating” text documents, but will not be touched upon further here (Norvig 2009).

## 8.5 Formal Data Structures Facilitate Integration

Several formalized data structures have existed for millennia. Financial ledgers recording barter/trade transactions were used in ancient Mesopotamia and various highly structured representations of the movements of the sun or moon, viz. calendars, have existed since the Bronze Age. For modern scientific data, however, several common, well-structured data formats exist. The specific implementations and interpretations of these data structures can vary somewhat, depending on what programming languages or analytical tools one is using. Aside from the potentially subtle but significant differences in the semantics of data structures, implementations can also vary relative to the syntaxes (language commands) required for their construction and manipulation. For example, the specific meanings of the following data structures and how they are created or manipulated vary among Python, R, and MATLAB. Nevertheless, these variations are mainly in the fine details and the following common usages of these named structures (Sects. 8.5.1–8.5.4) will generally be accurate.

### 8.5.1 Sets and Sequences

*Sets and sequences, including lists, vectors, and linear arrays*—are (typically) one dimensional data structures consisting of individual *values* (also called *elements*) that can be strictly ordered or not. While any collection of *unique* elements sharing some common property can constitute a *set*, elements in a *sequence* can be *repeated* and referenced by mapping to an *index* position within the set. This is necessary in order to readily refer to elements (or subsets) within the sequence, without having to specify their actual value. By convention, in many programming languages, square brackets [...] or braces {...} or even parentheses (...) are used to delimit vectors, arrays, and lists. Examples of lists are:

- items one must remember to purchase [wine, cheese, crackers, plates, cups]
- types of butterflies seen on a hike [*Danaus plexippus*, *Morpho menelaus*]
- the months in a year
- the height (m) of a plant measured on the first of each month from 2003 to 2015 [0.45, 1.18, 2.46, 5.78, 6.38, 7.72]

If the first list above was called “picnic”, then picnic[2] would equal “cheese” (the second element in the list); although in many computer languages, the origin of indexing starts at “0” not “1”, so the second element would be: picnic[1]. In vectors all elements must be of the same type, while in lists, elements can be of mixed types, e.g., both character and numeric.

## 8.5.2 Matrices

*Matrices* are two dimensional, rectangular arrays of numbers or expressions, where in a formal mathematical sense, index-ordering of the values is fixed. The two-dimensional or rectangular matrix is common, and has a formal structure such that the exact ordering of values is fixed across both the first (often called “row”) and second (often called “column”) dimensions. Individual elements are identified and accessible via an index, e.g.,  $X[i, j]$ , where  $X[i+1, j+2]$  has a specific relationship to all the other matrix elements—referencing the cell one row “below”, and one column to the “right”. Thus, matrices cannot be arbitrarily re-sorted by row or column without changing the nature of the matrix (unlike a table, as described below). Matrices also needn’t be solely numeric, though some language/software tools require that. Finally, matrices are often confused with, or somewhat misleadingly used to describe structures for displaying *cross-classified data*.

## 8.5.3 Cross-classifications

*Cross-classifications* are effective ways of *presenting* data when data elements are described according to two attributes, e.g., *seed shape AND seed color*; or *site AND species*, where the cell values could contain counts or other measures like total weight, etc. (Fig. 8.2). Cross-classifications are not, however, the best way to store and update raw data. One can see from the example in Fig. 8.2, that adding a new category of color or shape, such as “green-yellow”, or “dappled”, might require adding another column or row of data, and possibly re-classifying existing instances into these new categories. The *Table* structure described next provides a much better way for storing these types of data with great flexibility in splitting and lumping categories, and transformation to a Cross-classification format easily achieved a few simple transformation rules.

## 8.5.4 Tables

*Tables* are extremely common data structures used extensively in many natural and social sciences. There is both a common vernacular use of the term to indicate any structured “rectangular” data object, as well as a more formal definition that is grounded in the mathematics of set theory and relational algebra (Codd 1970).

In both cases, a table (equivalently called a “*relation*”), can be envisioned as a set of rows and columns, in which each column contains some particular type of

Fig. 8.2 Cross-classification data structure

SHAPE / COLOR	Green	Yellow
Smooth	323	46
Wrinkled	59	11



measurement or variable (also more formally called “attributes”), while the rows represent individual records or observations consisting of values for each of the column measurements. Note that these values can be “missing” or blank for any number of the columns in a row and any number of rows in the table. It can be a bit confusing to describe the structure of tables, as there is lots of variability in the terminologies that are commonly used (Fig. 8.3).

Tables are typically constructed such that rows collectively represent a set of records about some single “entity type”—where an “entity” represents some thing or process that has features that are measurable or named. Thus, a table might describe entity type=*People*, and measured fields might include a surname and birthdate. Rows would each represent an “instance” of that entity type, or in this case, a *Person*. Tables should be designed such that additions of information are easily made by adding rows, not columns, except for when some new type of attribute or variable about that entity type, needs to be added. In that case, defining a new column may be necessary. When creating a table, one specifies the name of table (usually reflecting the “entity type” it represents), and the names of the types of information contained in the columns. By definition, adding a new column to a table effectively creates a new table, whereas adding a row does not. This is because relational tables are defined by their name and the specific variables they contain.

Most current relational database management systems (RDBMS), including software such as PostgreSQL, MySQL, or Oracle—use SQL, or “Structured Query Language”, to construct and manipulate complex tabular data. SQL is highly standardized such that the *Data Definition Language* (DDL) syntax for creating a Table will look very much like the following in many relational database management systems:

```
CREATE TABLE Seed_Traits
(Shape VARCHAR, Color VARCHAR, Count INTEGER);
```

The Table is given a name (*Seed\_Traits*), and each Column (*Shape*, *Color*, *Count*) is given a name as well as a data type (*VARCHAR*, *VARCHAR*, *INTEGER*). The data type indicates how the computer should treat the cell values in those columns, e.g., as numerical ones amenable to arithmetic operations, or simply as text strings.

In “R” a similar table structure might be defined (as a data frame) as follows:

```
Seed_Traits <- data.frame(Shape=character(), Color=character(), Count=integer
());
```

Database (SQL)	Formal	Analytical term (R, SAS)
Table	Relation	Data frame, Dataset
Column	Attribute, Field	Variable
Row	Tuple	Record, Observation
Cell	Element	Value

**Fig. 8.3** Terms describing table structures; read across rows for “synonyms”



Note the strong similarity in syntaxes between R and SQL—involving naming the table (called a data frame in “R”) “Seed\_Traits”, and naming and typing the variables—Shape, Color, and Count. Despite these syntactical similarities, the underlying model of the data structure is somewhat different: a SQL table is a set of tuples/rows, while an R data frame is a list whose elements are vectors/columns. Technically, this leads to a difference in how data are inserted during the table/data frame creation process: in SQL—INSERT tuple/row; in R—add vector/column).

The DDL statements shown above for defining one’s tables in SQL and R include the name of the table or data frame, names and types of the attributes, and how these, as well as tables, might be related to one another. These statements represent your *data model*, or *schema*, in executable code. It is best to use scripts such as these to define one’s data since scripts can be later reviewed, modified, and re-executed, unlike, for example, simply typing your data into rows and columns in a spreadsheet. If one does not have the DDL statements (SQL, R, or other) for defining the structure and contents of one’s tables, it is nevertheless important to have a good grasp of one’s data schema, even if only represented in text or a diagram.

### 8.5.5 Tables or Spreadsheets?

Not everything that looks like a table *IS* A table. When one constructs a spreadsheet or worksheet in some popular application, such as Excel or Google Sheets, even if these are regularly structured in rows and columns with the first row being a “HEADER” row listing the column variables, one is typically not creating a true “table” by default. This is because there may be no “enforcement” of data typing—it would be permissible if one were to add a character string to a column in which all values should be numeric, or put a numeric value into a column that consists predominantly of text values. It is also common in many naively-constructed spreadsheets (and perniciously so) to find comments and other marginalia outside the “boundaries” of the table—which violate the integrity of the table and make it difficult to differentiate the tabular data from the other presentational, descriptive and summary elements surrounding it. While it is possible to use spreadsheets to create table-like structures, one should be aware that these other features can make later integration of those data highly problematic. One big advantage of using formal database software, like PostgreSQL or MySQL, or analytical software like “R” or SAS, is that when one creates a table or table-like structure (e.g., a data frame in “R”), the software will assist you in making the data conformant to the columns you specified with the data types you defined for them.

Exactly how tables are defined—with rows “collecting” data regarding one single object such as a person, or as often is the case in ecological and environmental studies, a heterogeneous entity, e.g., combining information about a specimen along with context gleaned from other sources, such as spatiotemporal data, information about the collector or methods, etc.—involves the art of *data modeling*. The choice of what variables to include in a table, or separate out into distinct

tables, depends to a great extent on the concerns of the researcher relative to the analyses they are planning for their data, and whether they might be “merging” their data with other data. However, it is always good practice to keep these principles in mind when constructing tables: to create tables that represent some entity type (e.g., a “person” table, or an “institution” table), structured with columns representing variables, and each row representing a set of “linked” or closely related (dependent) measurements. This will lead to data that are well constructed, and readily amenable to ingestion and further manipulation and analysis by most software packages. These are also the characteristics contributing to what is referred to in the “R” statistical world as “Tidy Data” (Wickham 2014).

### 8.5.6 Tables or Cross-classifications?

The cross-classification data format illustrated above in Fig. 8.2 superficially resembles a table. Where it differs is in how a well-modeled table can flexibly accommodate new values, as opposed to the cross-classification. In the cross-classification example above, the naming of “columns” as “Green” or “Yellow” creates a problem if a researcher wants to later track a new color, say “Light Green”, or “Greenish-yellow”. One can see that having a Variable for “Color” is better than using two possible *values* for Color (Green, Yellow) as *names* of Variables. With a well-constructed table, one simply adds Rows (records) to accommodate new values, e.g., for new SHAPES or COLORS (Fig. 8.4). Adding Columns, however, changes the table structure and, in many analytical packages, will require re-defining the table since you are adding a new Variable.

### 8.5.7 Modeling True Tables

Researchers often conceive of their data as residing in tables, but these tables often do not conform well with the formal requirement of representing some single entity type. Rather, researchers’ tables are often inherently heterogeneous objects—coupling together (in a record or tuple; Fig. 8.3) information about several different

SHAPE	COLOR	COUNT
Smooth	Green	323
Wrinkled	Green	59
Smooth	Yellow	46
Wrinkled	Yellow	11
Smooth	<i>Light Green</i>	3
<i>Semi-smooth</i>	Green	1

**Fig. 8.4** First few rows of Table “Seed\_Traits” with more flexible structure than cross-classification depicted in Fig. 8.2

entity types or things—e.g., documenting the taxonomic identities, counts, and heights of **trees**; along with contextual information such as the **location** where those measurements were taken (e.g., place name, geo-coordinates); along with perhaps some **climatological data** (e.g., mean annual precipitation and maximum air temperature); and **additional metadata** (date of collection, data collector's name), etc.

Such *analytically-ready tables* are often products of prior data integration processes even if those integration processes were done manually—merging in ancillary data such as a place name or precipitation data that were in fact derived from other sources. Even these tables, however, should have ALL unique rows—if all the values are identical in two or more rows of a table, this would indicate either a duplication error, or that those seemingly identical records in fact represent measurements that should somehow be further differentiated. For example, if two records from the Table about “Seed\_Traits” have identical values for the measurements of shape, color, and count, there should be some documentation, ideally contained in an additional column, about what “differentiates” those records—different sampling events in time, different specimen IDs, or measurements by different persons, etc.

While most scientists prefer working with *analytically ready tables* as above, *relational databases* take a different approach in modeling data tables, and for different ends. Relational databases are effective at storing complex and voluminous data, enabling a number of different analytically-ready tables to be derived from them. These analytically ready tables often result from the JOINing or integration of relational tables through a *query*. A *query* is a structured statement requesting information from the tables in a database and, as mentioned earlier, is often expressed in SQL. Relational databases are also very efficient at storing information with minimal redundancy. For example, information about taxonomic entities and their placement in a biological classification scheme (Family-Genus-Species) might be stored only once in a Table, but referenced many times by other tables in a relational database through a concise *Key relationship*, described in more detail below. Another advantage of a relational database is that, due to careful modeling and specification of relationships among tables through Keys and other constraints, the integrity of the data can be maintained, allowing multiple users to simultaneously add, delete, and update entries to the database without introducing errors (called “anomalies” in database terminology).

Data (as opposed to “statistical”) normalization is another important concept that is perhaps not familiar to many researchers, unless they have had to model some fairly complex data. Nevertheless, it is a useful concept to know even when creating simple tables. Data normalization involves thinking about the entities that “naturally exist” in your data, and then separating these entities out into separate tables. Information that is closely associated with the same type of entity is kept together as a tuple in its own table—e.g., the information above about *Jane Smith* might be a record in a “People” Table. In database terms, these types of necessary and close associations among attributes are called “dependencies”. When an attribute's value, e.g., an ORCID ID (ORCID, Inc. 2016), uniquely determines the value of some

other attribute (e.g., a birthdate), it is said that the birthdate is *functionally dependent* on the ORCID ID, or that the ORCID ID “functionally determines” the birthdate. Identifying functional dependencies in your data, and grouping these attributes together into the same table, is a big part of normalization. The most basic type of normalization, however, is called 1st normal form, and involves “atomizing” your data. Atomization is when you construct your tables such that the “cell” contents are limited to single values of “one thing” or measurement. Thus, it is not good practice to store values of an entire home address, or a given name and surname, or both latitude and longitude—together in a single cell. These should be separated out into more fundamental components—e.g., one variable (column) to hold values for latitude and one variable (column) to hold values for longitude.

“Keys” are extremely important in integrating data, as they are the set of values in a record that can be used to identify a unique occurrence or row in a table. Keys form the basis for linking one table with another by matching up their values *across* tables. Keys are still typically constructed such that they are only “unique” for rows within a specific database table. Local database keys are often generated as an additional column containing integer numbers, uniquely associated with a table row. A key can, however, also consist of more than one variable, if that is what is needed to uniquely identify a table row (e.g., both *date* and *location* might uniquely identify a record). Locally-scoped keys, however, make it difficult to identify and integrate compatible data *across* databases, which may be distributed around the Internet, or simply appear in multiple spreadsheets on a researcher’s desktop computer.

There are two important roles that Keys play in data integration: as PRIMARY KEY, or FOREIGN KEY. For a given table, one can identify a PRIMARY KEY, which is the attribute or set of attributes that uniquely identify a record (row) *in that table* as distinct from all the others. Each row of a table must have a unique value for its PRIMARY KEY. For example, in a table called “People”, that would contain one record per person, the ORCID Identifier could be specified as the Primary Key *for that table*. Wherever an ORCID Identifier might appear in *some other table*, however, it is called a “Foreign Key” and refers back to the table where that attribute is identified as its Primary Key. Thus, if an ORCID ID is designated as the Primary Key for a table containing a record for “Jane Smith”, wherever that particular ORCID ID appears as a Foreign Key in another table, it will be clear that it refers to the one particular *Jane Smith*, and her associated attributes in the “People” table, and not records of other individuals with that same name of *Jane Smith*. In relational databases Key *relationships* among tables are formally specified by explicitly identifying which columns are Primary Keys or Foreign Keys, and called *constraints*. When using other analytical frameworks, such as “R” or SAS, however, Key constraints are often not explicitly specified, in which case the researcher must be fully aware of which variables can serve as Keys, in order to properly integrate data tables.

There is insufficient room here to describe in detail *relational algebra*, which provides the underlying theory for modeling relational data. There is a strong mathematical logic underlying how tables and their attributes are constructed

using those algebraic principles, how to model and normalize one's data, and how keys should be chosen and used to inter-relate tables for integration. While the syntactical and even structural aspects of creating, manipulating and integrating tables can vary across analytical frameworks, the theoretical bases of information modeling and relational database structures are still very broadly relevant, and can be found in many books that describe the theory and strategies for building databases (e.g., Halpin and Morgan 2008; Connolly and Begg 2014).

### 8.5.8 *Need for Global Keys*

As it becomes easier and easier to integrate data from multiple sources, there is a danger of re-counting certain observations—e.g., in the case where we might have two identical observations for a “Jane Smith”, exhibiting the same height, weight, and birthday. These records might refer to two people with the same name, or be an erroneous repeat or “duplication” of data about the same “Jane Smith” individual, due to some earlier data integration. To differentiate whether these records refer to the same set of measurements or instances, one would need additional information, ideally involving unique identifiers, such as Social Security numbers (in the USA) or ORCID Identifiers.

Although the growth of the Web makes more data readily available to researchers for integration and analysis, it also makes more critical the need for “global keys”—in order to clearly reference the distinct instances of a multitude of objects, including not only obvious entities like people or institutions, but also other entities and phenomena that are singular—such as specimens from natural history museums or ice-core samples, or specific oceanographic cruises or field expeditions. In these cases, we might be able to infer from metadata what specific event or instance of some object took place or was measured—a survey of a vegetation plot, or a count of a herd of elephants—but in each case, we are relying on those metadata to differentiate one event or measurement from another. Critical information that allows such differentiation can often get lost when data are transferred, subsetted, summarized, re-combined, etc.

To afford better possibilities for data integration in general, and particularly over the Internet, Keys need to become *globally unique*. This requires expanding the notion of Keys beyond simply identifying unique rows in a database table, to also enable identifying unique tables, databases, and ideally any *distinct* information resource accessible over the Web. This is not as unfamiliar a notion as it first sounds—ORCID Identifiers essentially represent global Keys, in that they uniquely identify an instance of “something”—in this case individual researcher/scientists. This notion of a *globally unique identifier* may be conceptually simple, but its actual implementation can vary. The specific term “Globally Unique Identifier”, or “GUID”, for example, refers primarily to the assignment of a unique 128-bit digital signature to a digital object, with the capacity to assign over  $10^{38}$  such unique addresses to items (probably enough to last us a while!). On the Web, however, a

*Uniform Resource Identifier*, or *URI*, can also serve as a globally unique identifier (Allemang and Hendler 2011). URIs have a similar format to the more familiar URL (*Uniform Resource Locator*) that we type into the address area of our Web browsers. These two differ in that URLs indicate a *location* of something on the Web, while URIs have a broader meaning indicating a *resource*, which is *any* object referenced and accessible through the Web, including simply “locations” (URLs) for Web sites or pages. Thus, all URLs are URIs, but not necessarily vice-versa. With the growing need to access and integrate distributed data across the Internet, URIs will increasingly serve as global Keys pointing to data resources across the Web.

Journal publishers are already issuing globally unique identifiers for articles, most typically as Digital Object Identifiers, or DOIs (International DOI Foundation 2016), while Social Security numbers and ORCID IDs are other examples of globally unique identifiers for individual persons. As we increasingly use URIs as globally unique identifiers, the URIs will also need to be persistent—in that there is an intention and commitment that these will always “point to” a specific resource that can be accessed over the Web. This process of accessing the value or other representation of the resource pointed to by a URI is called “dereferencing”, and is a critical enabler of the Semantic Web and Linked Data (Berners-Lee et al. 2001; Heath and Bizer 2011). An increasing need for more effective integration of relevant data *distributed across the Internet* makes clear the advantage of using non-local, global Keys such as DOIs or ORCID IDs. Such global identifiers will become more commonplace in other aspects of scientific data and measurements, much more than simply persons or publications. Such new approaches for integrating Web-distributed data are not yet well covered in traditional references about databases, but are developing as blends of those technologies with emerging ones for the Semantic Web and Linked Data.

## 8.6 Merging or JOINing Tables

The most common motivation for data integration arises from researchers’ needs to combine tables to create an enriched set of “coupled” variables that can be further manipulated and analyzed in search of various statistical relationships and other patterns. These techniques often require constructing records in which some of the variables are identified as “predictors”, while other variables are hypothesized “outcomes” conditioned on the values of the predictor variables. Other motivations for integrating data might simply be to test for potential relationships among variables, such as positive or negative correlations with one another. These analyses rely on bringing the data together in ways that clarify which measurements are somehow associated together, which is often indicated by grouping those values together in the same record.

There are a number of ways to merge tables, depending on the needs and interests of the researcher. These various approaches have proper names and, even if a data

analyst often finds it necessary to mix and customize these approaches, it is good to be aware of the main integration processes and their proper names. To keep matters simple, the focus here will be on merging two tables together, but the same principles hold for merging together more than two tables. It is usually clearer, however, to merge tables in an iterative pair-wise fashion, at least initially.

The most basic way of merging or integrating two tables involves a Cartesian product, also known as a *cross join*, such that every row (or tuple) of one table is matched with every row of another table. That is, if Table A has 8 rows and 3 attributes, and Table B has 10 rows and 2 attributes, the resulting Cartesian product, Table C, has 80 ( $8 \times 10$ ) rows with 5 ( $3 + 2$ ) attributes. However, it is rarely the case that researchers will be merging data in this way—usually doing so only if they need to create all possible combinations of the records from one data set with another. More typically, researchers are bringing data together based on some “matching” variable between the tables. These matching variables, which ideally are Keys, must be *domain compatible*—that is, they represent the same “thing” or measurement, and have the same set of allowable values. This generally requires an understanding of the semantic contents of the variables. If two Tables have records with the same value for a Key variable, this would indicate that those tables have some measurements in common.

### 8.6.1 APPENDING or Unioning

A common data integration operation used for combining two or more tables that have the same variables, is “appending”. This involves simply attaching tables to one another, by matching along compatible variables, creating an output table that contains the sum of the rows from the input tables (minus potential duplicates). This is a highly useful operation when one collects numerous data sets of identical structure and column semantics, but that may be housed separately because, for instance, they are collected during different years, or from different places, or by different people. When appending these types of data, it is often necessary then, to afterwards add a column or two of additional information that will provide critical differentiating metadata about time, place, or person.

The two tables A and B in Fig. 8.5 both contain at least two variables of interest, for example, the names and weights of people participating on some project. In

**Fig. 8.5** Tables A and B, to demonstrate behavior of data integration operations

Table “A”	Table “B”																								
<table border="1" style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th style="text-align: left;">ID</th> <th style="text-align: left;">Name</th> <th style="text-align: left;">wt</th> </tr> </thead> <tbody> <tr> <td>1</td> <td>Amber</td> <td>115</td> </tr> <tr> <td>2</td> <td>Bill</td> <td>205</td> </tr> <tr> <td>3</td> <td>Dave</td> <td>175</td> </tr> </tbody> </table>	ID	Name	wt	1	Amber	115	2	Bill	205	3	Dave	175	<table border="1" style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th style="text-align: left;">ID</th> <th style="text-align: left;">name</th> <th style="text-align: left;">wt</th> </tr> </thead> <tbody> <tr> <td>1</td> <td>Mark</td> <td>185</td> </tr> <tr> <td>2</td> <td>Matt</td> <td>205</td> </tr> <tr> <td>3</td> <td>Rebecca</td> <td>115</td> </tr> </tbody> </table>	ID	name	wt	1	Mark	185	2	Matt	205	3	Rebecca	115
ID	Name	wt																							
1	Amber	115																							
2	Bill	205																							
3	Dave	175																							
ID	name	wt																							
1	Mark	185																							
2	Matt	205																							
3	Rebecca	115																							

Table A our variables of interest are name and wt, while in Table B, the variables are also called name and wt. The ID variables in both tables A & B might be simple identifiers, that will not be of interest after the append unless additional information is added, since these will not be unique across tables after the tables are merged—they cannot serve as Keys for joining the two tables. To enable the ID column to retain its potential as a Key value *after* the append operation, we would need to modify the values, e.g., by including the originating table label as an identifier—hence, for Amber the identifier might be “A1”, while for Rebecca it might be “B3”. But, this would not happen automatically with a simple append operation.

In the standardized syntax of SQL, the following statement would be used to append these two tables:

```
CREATE TABLE C (ID, name, wt)
INSERT INTO 'C'
SELECT * FROM (
SELECT * FROM A
UNION ALL
SELECT * FROM B);
```

The “UNION ALL” would not exclude duplicate records from the result of combining the two Tables, whereas use of only a “UNION” statement would eliminate duplicate rows from the resulting table. The “SELECT \*” statement here indicates that all the variables in the “FROM” table will be included in the output. Note that in this example we show the statements needed to create the output “Table C” using the SQL “CREATE TABLE”, and “INSERT INTO” statements. In later examples we leave out this step, which in “R” would be like leaving out the assignment of an output to a new named data frame. The resulting output “Table C” appears in Fig. 8.6. Note that the “matching” structure of our two variables of interest allows us to simply vertically “stack” the two data sets to create one integrated output.

In “R”, the *rbind* command very similarly appends data frames (assuming the Tables A and B above are converted to data frames A and B) if they have matching column names, and would produce the same output as above:

```
base::rbind(A,B) # using base R
dplyr::bind_rows(A,B) # using the dplyr package
```

In SAS, the *SET* command can be used for this and other powerful appending operations.

**Fig. 8.6** Output Table C from appending the Tables A and B shown in Fig. 8.5

ID	name	wt
1	Amber	115
2	Bill	205
3	Dave	175
1	Mark	185
2	Matt	205
3	Rebecca	115



## 8.6.2 JOINS

Perhaps the most common data integration operations involve merging multiple tables based on matching values in specified subsets of variables shared among the tables. These types of operations are collectively called “JOINS” in SQL. Note, however, that other data manipulation frameworks, such as R, Matlab or SAS use different terminologies to refer to some of these data merging processes. For example, R has a number of named functions for doing these types of operations, and these vary from package to package (e.g., in R::base, or the popular *dplyr* and *data.table* packages). The R package *dplyr*, however, does use syntax largely borrowed from SQL for accomplishing many of its data frame “combining” operations. SAS accomplishes JOINS using both the *SET* and *MERGE* commands, and uses the sorted order of variables to determine the exact outcomes. In addition, however, SAS offers a rich set of SQL commands through its PROC SQL procedure.

We will describe four types of JOINS: INNER JOIN; and LEFT, RIGHT, and FULL OUTER JOINS. An understanding of how these joins work in SQL will provide a solid foundation for doing other types of merges, as is possible using powerful data manipulation languages such as R, SAS, or Python’s *pandas* package.

When joining two or more tables, it is typically necessary to identify “matching” variables that represent the relationships or “linkages” among those tables. As discussed earlier, Key variables are often used for this purpose. Recall that modeling your data usually involves normalizing them—reducing data redundancy, clarifying the natural relationships among the data by grouping related attributes together into a Table, and identifying the Key variables. Ideally all functionally dependent information about an entity instance is fully documented in a single Table. For example, for a person named *Jane Smith*, all information tightly associated with that individual would be stored ONCE in a “People” Table—surname, given name, institutional affiliation, birthdate, etc. This greatly reduces the possibilities for errors because, e.g., if the values of any of these attributes need to be changed, the change only requires updating entries within one single record in one table. All other tables that might be linked to Jane Smith (and other peoples’) record through a Key variable would then automatically be updated.

There are basically three ways in which records from separate tables can be related or “matched up”, and these relationships are called *cardinality constraints*. Considering two tables at a time, records from one table to another can be related as: one-to-one (often indicated as “1:1”), one-to-many (“1:M”), or many-to-many (“M:N”). One-to-one relationships usually involve tables that are closely related, and often the attributes in these might appear in the same Table instead of stored separately. But sometimes there are reasons to separate these. For example, we might have a “People Public” Table with an ORCID ID as a *Primary Key*, while also containing other personal details (birthdate, country of birth) about some individual, but a separate “People Private” Table, that might contain a social security number, medical history or other confidential information. This latter table merely needs to contain a person’s ORCID ID to be used as a *Foreign Key* to link back to (“match with”) the appropriate Primary Key ORCID ID in the People

Public Table, or in this case, possibly vice-versa. Since every citizen and permanent resident of the USA should have one and only one Social Security number, and every scientific researcher should have an ORCID ID, there should be a 1:1 relationship among the records from those two tables.

More commonly, data tables are integrated through one-to-many relationships. For example, a relationship linking (“matching”) taxonomic names (stored in a Taxonomy Table) to the taxonomic identity of individual plant stems recorded in a Vegetation Plot Table is 1:M, since a single taxonomic name can be assigned to many stems, but any given stem can only be assigned to a single taxon. One can reverse the ordering of this description—many stems can belong to one taxon, and describe that relationship equivalently as M:1. In contrast, the relationship between a Table listing plant host species and a Table of potential pollinator species would likely be many-to-many (M:N), since a plant might host multiple pollinator species, and a pollinator might visit multiple plant species. Note finally that there may also be no relationship at all between the records in two tables: a table containing measurements of plant heights from sampling plots in Kansas might have no relationship with a table of measurements of barnacle densities around deep-sea hydrothermal vents in the Pacific Ocean! These two tables might have no *Key variables* in common that could provide a basis for matching up and integrating records across them.

Of these potential relationships, integrating tables that have many-to-many relationships with one another are the most complex, and require careful planning in order to create effective and meaningful joins. In relational database systems where the data are highly normalized (i.e., data are atomized, redundancies minimized, and functionally dependent attributes are in the same tables), these situations often require creation of another table, variably called a “linking”, “associative”, “junction”, or “mapping” table, that essentially includes the Key mappings of the matching instances from both of the tables. For example, a familiar situation is the many-to-many relationship between books and their readers (books are read by multiple people; people read multiple books). The *linking table* would contain individuals’ ORCID IDs in a tuple with the ISBNs (a type of global identifier) of the books they’ve read. There would be a record for each “reader/book” combination. Thus the table would contain multiple records per person, as well as multiple records per book. But the person and book tables, if normalized, would each contain only one record per person or book, respectively. The linking or associative table is necessary to document the specific many-to-many relationships among entries from the two normalized tables. Creation and use of such tables in relational databases is a topic that is well covered in books on modeling relational data (Halpin and Morgan 2008; Connolly and Begg 2014).

It is common in many field-collected ecological data sets, however, for the raw data to be entered “as observed”, such that the same taxonomic names might appear in many records, such as in the plant/pollinator case described above. These are essentially already “linking tables” that allow one to bring together, as 1:M JOINS, data from normalized tables that might describe, e.g., the traits of the plant species (Plant Trait Table; 1 row per taxon), and the traits of the pollinator species (Pollinator Trait Table; 1 row per taxon). One should carefully consult the manual for one’s software to be sure of the exact syntax needed for merging tables that have many-to-many matches.

Here we focus on the simpler but very common cases that involve merging tables with 1:1 and 1:M relationships. The latter case is particularly useful when integrating tables based on key relationships that bring together complementary data that enable exploration of new hypotheses and analyses.

The first type of JOIN we will describe is called an INNER JOIN, in that it includes only those rows that are matches of variables from both tables. Here, we do an INNER JOIN on the variable `prodID` (which might represent a purchase code in this example), found in Tables `Customer` and `Item` (Fig. 8.7) using SQL code:

```
SELECT cName, prodID, pName, cost
FROM Customer AS C
INNER JOIN Item AS I
ON C.prodID= I.prodID
```

resulting in the output shown in Fig. 8.8.

- The “**Customer AS C**” statement enables the name of the table to be abbreviated from “`Customer`” to later be referenced as simply “`C`”
- In the statement “**ON C.prodID**”, the `prodID` attribute in table `Customer` is being referenced.
- The “**ON**” statement references the same “linkage” variable, `prodID`, from both the `Customer` and `Item` tables.

Note that the record where `C.prodID=2500`, `ID=3` and `Name=“Dave”` from Table `Customer` is missing from this output; as is the row where `P.prodID=25`, and `Name=“Cheese”` from Table `Item`. That is because there were no matches for the values of the `prodID` variable of those records (`C.prodID=250` or `P.prodID=25`) in both tables. In this case, `Customer Dave` may not have bought `Item Cheese`, but instead purchased something else that was not listed in the `Item` table.

**Fig. 8.7** Sample Tables “`Customer`” and “`Item`” to demonstrate behavior of different data integration operations

Table “Customer”			Table “Item”		
ID	cName	prodID	prodID	pName	cost
1	Amber	15	15	Crackers	24
2	Bill	30	25	Cheese	48
3	Dave	2500	30	Wine	89
1	Amber	201	201	Plates	10
2	Bill	215	215	Cups	12

**Fig. 8.8** Output from INNER JOIN on Tables “`Customer`” and “`Item`” from Fig. 8.7

cName	C.prodID	I.prodID	pName	cost
Amber	15	15	Crackers	24
Bill	30	30	Wine	89
Amber	201	201	Plates	10
Bill	215	215	Cups	12

You could accomplish this type of inner join in “R” using the *merge* command on analogous data frames of Customer and Item:

```
base::merge(Customer, Item, by.x="prodID", by.y="prodID", all=FALSE)
dplyr::inner_join(Customer, Item, by="prodID")
```

For ecologists, one might imagine the above tables linking taxonomic names of Predators with their Prey Items through a matching “food\_for\_ID” (or “eats\_ID”) variable. Or a table consisting of “Stream Chemistry” measurements might be linked through a variable such as “locationID” to a “Site” table that included details about the place that was sampled (stream name, geo-coordinates, mean flow volume, stream depth and width, etc.). In order not to distract researchers about the “correctness” of the measurements and entities JOINed in these examples, we keep them very generic. We challenge the scientist to imagine how these JOIN patterns pertain to their own data integration needs.

Another common case is when someone wants to match a variable or variables in one table with those in another, but doesn’t want to “lose” those records that have no match. Instead, the unmatched values might have missing or NULL values for any additional variables represented in the newly merged output table. These are OUTER JOINS, and there are three types: LEFT, RIGHT, and FULL.

A LEFT (OUTER) JOIN, often called a “LEFT JOIN” follows—

```
SELECT cName, prodID, pName, cost
FROM Customer AS C
LEFT JOIN Item AS P
ON C.prodID= P.prodID
```

resulting in the output shown in Fig. 8.9.

The values of “NULL” for variables pName and cost in the resulting output Table occur where the attributes had no matches for C.prodID=P.prodID for value=2500 in Tables Customer and Item. The “LEFT JOIN”, however, specifies that even unmatched records from the first (LEFT-hand) mentioned table (here Table Customer with cName=“Dave” and C.prodID=2500), are carried into the resulting output table, so every row from Table Customer appears in the output Table.

Similar results using “R” could be generated as:

```
base::merge(Customer, Item, by.x="prodID", by.y="prodID", all.x=TRUE)
```

or

```
dplyr::left_join(Customer, Item, by="prodID")
```

Fig. 8.9 Output from LEFT JOIN on Tables “Customer” and “Item” from Fig. 8.7

cName	C.prodID	P.prodID	pName	cost
Amber	15	15	Crackers	24
Bill	30	30	Wine	89
Dave	2500	NULL	NULL	NULL
Amber	201	201	Plates	10
Bill	215	215	Cups	12

Note that the exact special characters or numbers representing a “NULL” value in a table (or data frame) can vary depending on the database or analytical framework you are using. Also note again that the ordering of rows in a table is arbitrary—a table can be sorted according to the values for any variable of interest, but by definition there is no intrinsic ordering of the rows (or columns for that matter) in a table.

A “RIGHT JOIN” is very similar to a LEFT JOIN, only differing in that ALL the records from the second mentioned (RIGHT-hand) Table are retained, with unmatched variables from the first (LEFT) table filled with NULL values:

```
SELECT cName, prodID, pName, cost
FROM Customer AS C
RIGHT JOIN Item AS P
ON C.prodID= P.prodID
```

resulting in the output shown in Fig. 8.10.

In the case of OUTER JOINS, the exact nature of the output can also vary, depending on what analytical package you are using and the options you specify for the results—such as whether both the “prodID” variables from the “left” and “right” tables are included in the result set or not.

One would produce very similar results in “R” using:

```
base::merge(Customer, Item, by.x="prodID", by.y="prodID", all.y=TRUE) #
BASE
dplyr::right_join(Customer, Item, by="prodID")
```

A “FULL OUTER JOIN” maintains ALL the records from both Tables, and fills in unmatched variables with NULL values as shown in Fig. 8.11.

```
SELECT cName, prodID, pName, cost
```

**Fig. 8.10** Output from RIGHT JOIN on Tables “Customer” and “Item” from Fig. 8.7

cName	C.prodID	P.prodID	pName	cost
Amber	15	15	Crackers	24
Bill	30	30	Wine	89
Bill	215	215	Cups	12
NULL	NULL	25	Cheese	48
Amber	201	201	Plates	10

**Fig. 8.11** Output from FULL JOIN on Tables “Customer” and “Item” from Fig. 8.7

cName	C.prodID	P.prodID	pName	cost
Amber	15	15	Crackers	24
Bill	306	306	Wine	89
NULL	NULL	25	Cheese	48
Amber	201	201	Plates	10
Dave	175	NULL	NULL	NULL
Bill	215	215	Cups	12

```
FROM Customer AS C
FULL JOIN Item AS P
ON C.prodID= P.prodID
```

In “R”, near equivalent result sets can be created by:

```
base::merge(Customer, Item, by.x="prodID", by.y="prodID", all=TRUE)
dplyr::full_join(Customer, Item, by="prodID")
```

The attentive reader will note that the above JOINS also demonstrate one-to-many relationships of the Customer with the Item table—Customers Amber and Bill both are associated with multiple Items. Also, one can usually imagine some implicit “verb” that describes the relationship between tables that are JOINed. For example, in this case, the relationship might be “purchases”. Amber purchases Crackers and Plates, while Bill purchases Wine and Cups.

Be aware that different analytical packages can handle JOINS in very different ways, both in terms of how to specify the desired operation in some computing language (the syntax), as well as variation in how the outputs are presented. For example, the R *base::merge* function operates very differently, both syntactically and semantically, from functions in the *dplyr* or *data.table* packages, for doing join operations. However, as can be seen here, even R’s *base::merge* function is quite flexible and powerful for doing various JOINS. It is necessary to read the manual for your software very carefully when doing JOINS, to be sure that your outputs are what you expect them to be.

## 8.7 The Datum Is the Atom

While tables represent one of the most common and useful structures for storing and integrating data, it is important to understand at the most atomic level: what are the phenomena being named, typed, described, and/or quantified and preserved in digital data? The growing demand for synthesis of data across traditional scientific disciplines further motivates the need to better understand what data and observations are collectively available, so these can be integrated for further analysis. These trends have led many earth and biological science informatics groups to recently converge on a common model for data: as sets or collections of *observations and measurements* (Madin et al. 2008; Cox 2015). These models all seek to enable more efficient data integration by placing the focus on data at its most elemental levels, which are the individual observations and measurements. Scientific observations are decomposed into constituents representing the *entity* (thing or process) that is observed; the *characteristics* of the entity or process that were documented or measured, and assigned values; and the specific scale or *units* associated with those values (Fig. 8.12). In many cases, additional critical metadata

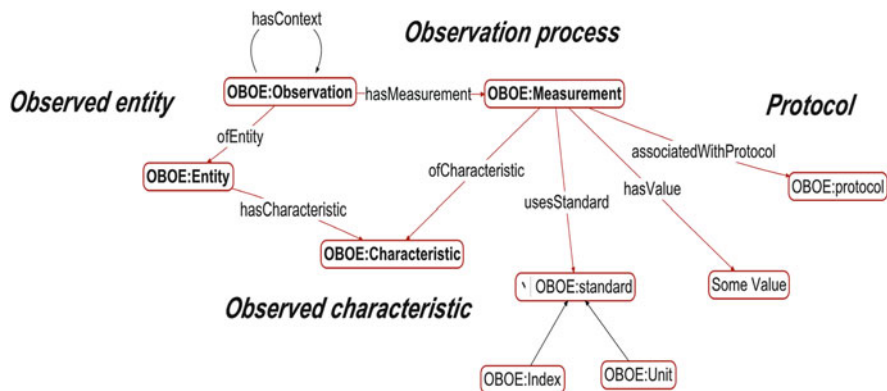
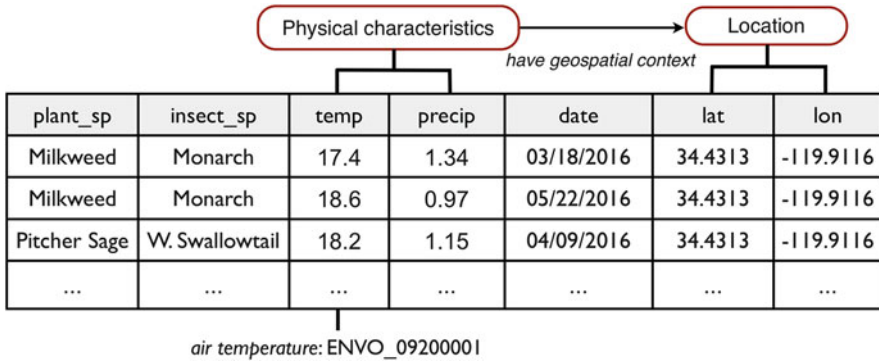


Fig. 8.12 Schema for a typical observational data model

might describe methodological or protocol specifications as well (although these often might not be available, or readily inferred).

While the most common use of the term “observation” still refers to a row or record in a table (e.g., SAS uses this terminology), in many cases calling these row-level accumulations of data “observations” is somewhat misleading. As we have seen, individual values “coupled together” in a record often come from multiple original data sources, and are often associated together by some researcher for some specific purpose, such as for a regression analysis. In most observational data models, the observation is highly atomic—representing an act of measurement of some particular instance, individual, or phenomenon. Then, as depicted in Fig. 8.12, these observations are inter-related through some “has Context” relationship, where “has Context” can be further clarified by using a more specific description, such as “has same geospatial context” or “taken from same specimen”.

For example, a table record might consist of seven observations: (1) mean daily air temperature coupled with (2) monthly precipitation measurements (acquired from a nearby set of weather gages), merged with weekly censuses of the (3) names and (4) abundances of herbivorous insect larvae found on some (5) named plant species, located at (6) some geospatial place name or geo-coordinates, at (7) some point in time. In this case, our table really consists of only five distinct “entities”, some with many potential characteristics that may or may not have been measured or otherwise documented, e.g., (1) characteristics of the local atmospheric conditions, (2) traits of the insect larvae, (3) traits of the plant host species, (4) descriptive aspects of the local habitat and associated terrestrial environment, and (5) the date-time. We could call such a “composite” record a single observation (according to some relational data terminologies), **or** using the terminology of most *observational data models*—it would consist of several linked or coupled observations or measurements—consisting of the *physical characteristics* of some place, that contextualize some measurements on *insect* and *plant entities*, with further information about a place’s *geospatial location* (Fig. 8.13). We have also depicted in Fig. 8.13 how specific measurements might be annotated via a unique identifier, in



**Fig. 8.13** Observational data model showing relationships among table variables, and reference to external identifier

this case clarifying that the column labelled “temp” measures an *air temperature*, as defined by the Environment Ontology’s term ENVO\_09200001, which references a URI (University of Michigan 2016a). The base URI for the “air temperature” measurement in the Environment Ontology would be “[http://purl.obolibrary.org/obo/ENVO\\_09200001](http://purl.obolibrary.org/obo/ENVO_09200001)”, where the “purl” in the address indicates that this is intended to be a *persistent* URL or “PURL”, that is also globally unique.

Careful examination of the “entities” in Fig. 8.13 reveals that for each, only a small set of potential “characteristics” were measured—for weather, only air temperature and precipitation (Fig. 8.13), and not, e.g., relative humidity or air pressure, but these might be linked in from other sources. It further becomes clear that there must be information as to the *units* associated with many of these measurements, e.g., degrees *Celsius* for air temperature, or *centimeters* for precipitation. All recorded values in scientific data should use rigorously defined, highly comparable standards for measurement, as much as possible, to enable consistent, accurate interpretation by researchers regardless of language, location, or time of access.

Regarding the plant/herbivore data in this example table—additional attributes might result from direct measurement such as plant size or age, but others might be inferable simply from knowledge about an entity’s belonging to some class. For example, one of the plant species might be identified as a Milkweed, and through that name linked to an entry in another table providing more complete taxonomic information, such as the scientific name *Asclepias spp.*, as well as plant family, tribe, etc. Milkweed is poisonous to grazing animals, but the toxicity and potential avoidance of Milkweed plants by specific insect species might not be known to the researcher, nor directly assessed by them. Curiosity about how plant chemistry might be structuring plant/insect interactions could motivate further integrating these data with records from another table (possibly prepared by another researcher)—that documents the toxicity of specific chemicals to various insect taxa. From this integrated output table, we might discern that monarch butterflies and several other insect species are immune to Milkweed toxins, but a number of



insect species are not. One can see that many data integration actions are motivated by importing (through a JOIN) a subset of columns from other data sets. The observational data model makes clear that each value in a column represents a distinct observation or measurement, and that the full tuple (row or record) structure in a table might result from integration of multiple heterogeneous data sources, coupled together to help inform some particular analysis of interest.

Observational data models afford maximum flexibility with regards to defining tables by encouraging explicit specification of the semantics necessary to understand the contents of any table “cell” value—in terms of the entity of interest, the characteristics of the entity that are measured, and the units and methods involved in obtaining the measurement(s). This type of information might be contained in a *Data Dictionary*, or more recently, documented using various *metadata standards*, such as the Ecological Metadata Language, EML (Fegeaus et al. 2005), that is used by environmental data repositories such as the U.S. National Science Foundation’s Arctic Data Center (2016), or DataONE (2016). Ironically, however, there are numerous ways in which the “same” entities and their characteristics can be described or defined in data dictionaries or metadata standards, making evaluation of their semantic similarity or equivalence difficult. In the same way that standardization of measurement units, such as agreement on the meaning of a “millimeter” or “kilogram”, has greatly facilitated comparability of data, there is currently a need to increase standardization of the names of entities and measurements, so that scientific data are more effectively discoverable, interpretable, and amenable to potential integration with other data.

One potential solution to reducing the current “babel” of scientific terminologies involves developing and agreeing upon community vocabularies. Entire data objects, as well as individual measurements, are increasingly semantically defined in this way, by referencing terms in web-accessible vocabularies (known as “ontologies”) through dereferenceable globally unique URIs. For example, if one wants to indicate that the measurements in the table depicted in Fig. 8.13 were taken from a “Mediterranean grassland biome”, one can indicate that by referencing a persistent URI, such as one found in the Environment Ontology, ENVO (Buttigieg et al. 2016; University of Michigan 2016b). Syntaxes used to define data in this way are typically expressed in Resource Description Framework (RDF) and Web Ontology Language (OWL)—which are the formal languages recommended for the Semantic Web (Berners-Lee et al. 2001; W3C OWL Working Group 2016). The exact mechanisms for linking ontology terms to digital object structures, e.g., tables, table columns, or even cell values in databases, are currently under development, and involve a process known as “semantic annotation” (Madin et al. 2008). In any case, when it becomes easier and more common for researchers to reference terms from shared vocabularies to unambiguously describe their data via Web-based globally unique identifiers such as URIs, data integration will become much easier. In addition, this practice will enable more precise semantic searches, such that measurements from multiple tables of data distributed around the Web can be more effectively discovered, and “matched” with other measurements.

## 8.8 Conclusion

Data integration involves bringing together distinct, often heterogeneous data, in order to enable more powerful and comprehensive modeling and analysis of phenomena of interest. Data integration is a key data manipulation process that is driving much scientific synthesis by enabling the coupling together of additional and complementary information to derive more holistic or robust understanding.

Our focus here has been on tabular data—one of the most common data structures that field scientists in particular must contend with. We have tried to convey that there is a strong theoretical basis for constructing and interpreting data (for tables in particular). We have emphasized here some of the basic principles, and encourage the reader to investigate the subject further, as this will not only enhance understanding of tabular data structures, but should lead to more robust and accurate coding of diverse data integration processes, that can often get quite complex.

The specific syntactical mechanisms for integrating data can vary considerably from one analytical package to another, but the relational data model and the standardized SQL syntax provide a strong foundational perspective for understanding how to flexibly create, transform and merge tabular data. This understanding can transfer into better practice as well when using free-form tools such as spreadsheets to capture and manipulate scientific data. We hope, however, the reader is convinced that more specialized data creation and manipulation tools that require explicit specification (e.g., in coded statements) of table structures, their key attributes, and how tables are related to one another provide huge advantages with regards to the flexibility and transparency of creating and operating on one's own, as well as others' data, in both the short and long term.

Finally, it is important to recognize that scientific data are collections of “assertions” about the state of physical reality at some time and place, taken by observers who are trained in appropriate, objective methods, often using sophisticated, specialized instruments. As such, scientific data should be regarded with special respect, and in many cases, carefully preserved in ways that can be accessed and interpreted in the future. This is particularly true for data about the state of the environment, which we now know to be rapidly changing. Well-conceived, well-structured records about the state of our environment in the past, present, and into the future—of biological, physical, chemical and other measurements—may prove invaluable for understanding both the long-term and short-term processes that are governing the state of our biosphere. Facilitating the discovery and integration of these records through sound data management practices will provide great benefits to the scientific research enterprise when trying to uncover the complex relationships that govern ecosystems and their components in this age of increasing global human impacts.

**Acknowledgements** I would like to thank Julien Brun for suggesting several useful changes and corrections to the text. Shawn Bowers was a stalwart companion while dissecting the structure of numerous scientific datasets. But I want to acknowledge especially many years of fruitful and stimulating discussions with Matthew B. Jones on matters regarding the nature of ecological data,

and the need for better software tools and cyberinfrastructure to support synthesis and collaboration in the environmental sciences. The National Center for Ecological Analysis and Synthesis, NCEAS, has provided a strongly supportive environment for advancing ecoinformatics practice, and still represents, to my mind, a beacon for promoting and facilitating synthesis in the ecological and conservation sciences. Finally, I want to thank colleagues from several past and ongoing NSF-sponsored Cyberinfrastructure projects, including DataONE (NSF #1430508), SEEK (NSF #0225676), SONet (NSF #0753144), and the KNB (NSF #9980154). It has been a continual and pleasurable collaborative learning process with many bright and selfless colleagues.

## References

- Allemang D, Hendler J (2011) Semantic web for the working ontologist, Effective modeling in RDFS and OWL, 2nd edn. Morgan Kaufmann, Waltham, MA
- Arctic Data Center (2016) NSF Arctic Data Center. <https://arcticdata.io>. Accessed 5 Dec 2016
- Berners-Lee T, Hendler J, Lassila O (2001) The semantic web: a new form of web content that is meaningful to computers will unleash a revolution of new possibilities. *Sci Am* 284:34–43
- Buttigieg PL, Pafilis E, Lewis S et al (2016) The environment ontology in 2016: bridging domains with increased scope, semantic density, and interoperability. *J Biomed Semantics* 7:57. doi:10.1186/s13326-016-0097-6
- Canfield MR (ed) (2011) Field notes on science and nature. Harvard University Press, Cambridge, MA
- Carpenter SR, Armbrust EV, Arzberger PW et al (2009) Accelerate synthesis in ecology and environmental sciences. *BioSci* 59:699–701. doi:10.1525/bio.2009.59.8.11
- Codd EF (1970) A relational model of data for large shared data banks. *Commun ACM* 13:377–387. doi:10.1145/362384.362685
- Codd EF (2000) The relational model for database management: version 2. Addison Wesley, Reading, MA
- Connolly T, Begg C (2014) Database systems: a practical approach to design, implementation, and management, 6th edn. Pearson, Upper Saddle River, NJ
- Cook RB, Wei Y, Hook LA et al (2017) Preserve: protecting data for long-term use, Chapter 6. In: Recknagel F, Michener W (eds) Ecological informatics. Data management and knowledge discovery. Springer, Heidelberg
- Cox S (2015) Ontology for observations and sampling features, with alignments to existing models. <http://www.semantic-web-journal.net/content/ontology-observations-and-sampling-features-alignments-existing-models-0>. Accessed 5 Dec 2016
- Darwin C (1837) Darwin's first diagram of an evolutionary tree *from* First Notebook on “Transmutation of Species”. [https://commons.wikimedia.org/wiki/File:Darwins\\_first\\_tree.jpg](https://commons.wikimedia.org/wiki/File:Darwins_first_tree.jpg). In the public domain {PD-US}. Accessed 24 Jan 2017
- DataONE (2016) DataONE. <https://www.dataone.org>. Accessed 5 Dec 2016
- Fegraus E, Andelman S, Jones MB et al (2005) Maximizing the value of ecological data with structured metadata: an introduction to Ecological Metadata Language (EML) and principles for metadata creation. *Bull Ecol Soc Am* 86:158–168. doi:10.1890/0012-9623(2005)86[158:MTVOED]2.0.CO;2
- Greene B (2005) The fabric of the cosmos. Vintage, New York
- Halpin T, Morgan T (2008) Information modeling and relational databases: from conceptual analysis to logical design, 2nd edn. Morgan Kaufman, Burlington, MA
- Hampton SE, Anderson SS, Bagby SC et al (2015) The Tao of open science for ecology. *Ecosphere* 6:1–13

- Heath T, Bizer C (2011) *Linked data: evolving the web into a global data space* (1st edn). Synthesis lectures on the semantic web: theory and technology 1:1, 1–136. Morgan & Claypool, London
- Hempel C (1970) *Aspects of scientific explanation, and other essays in the philosophy of science*. The Free Press, New York
- International DOI Foundation (2016) <https://www.doi.org>. Accessed 5 Dec 2016
- Kuhn T (1996) *Structure of scientific revolutions*. University of Chicago Press, Chicago
- Madin J, Bowers S, Schildhauer M et al (2008) Advancing ecological research with ontologies. *TREE* 23(3):159–168. doi:[10.1016/j.tree.2007.11.007](https://doi.org/10.1016/j.tree.2007.11.007)
- Michener WK (2015) Ten simple rules for creating a good data management plan. *PLoS Comput Biol* 11(10):e1004525. doi:[10.1371/journal.pcbi.1004525](https://doi.org/10.1371/journal.pcbi.1004525)
- Norvig P (2009) Natural language corpus data, Chapter 14. In: Segaran T, Hammerbacher J (eds) *Beautiful data: the stories behind elegant data solutions*. O'Reilly Media, Sebastopol, CA, pp 219–242
- ORCID, Inc. (2016) ORCID. <http://orcid.org>. Accessed 5 Dec 2016
- Platt JR (1964) Strong inference. *Science* 146(3642):347–353. doi:[10.1126/science.146.3642.347](https://doi.org/10.1126/science.146.3642.347)
- Quine WV (1981) *Theories and things*. Belknap Press of Harvard University Press, Cambridge, MA
- Strasser C, Abrams S, Cruse P (2014) DMPTool2: expanding functionality for better data management planning. *Int J Digital Curation* 9(1):324–330. doi:[10.2218/ijdc.v9i1.319](https://doi.org/10.2218/ijdc.v9i1.319)
- Taper ML, Lele SR (eds) (2004) *The nature of scientific evidence: statistical, philosophical, and empirical considerations*. The University of Chicago Press, Chicago
- University of Michigan (2016a) Ontobee: environment ontology. [http://www.ontobee.org/ontology/ENVO?iri=http://purl.obolibrary.org/obo/ENVO\\_09200001](http://www.ontobee.org/ontology/ENVO?iri=http://purl.obolibrary.org/obo/ENVO_09200001). Accessed 5 Dec 2016
- University of Michigan (2016b) Ontobee: environment ontology. [http://www.ontobee.org/ontology/ENVO?iri=http://purl.obolibrary.org/obo/ENVO\\_01000224](http://www.ontobee.org/ontology/ENVO?iri=http://purl.obolibrary.org/obo/ENVO_01000224). Accessed 5 Dec 2016
- W3C OWL Working Group (2016) Web Ontology Language (OWL). <https://www.w3.org/2001/sw/wiki/OWL>. Accessed 5 Dec 2016
- Wickham H (2014) Tidy data. *J Stat Softw* 59(10):1–23. doi:[10.18637/jss.v059.i10](https://doi.org/10.18637/jss.v059.i10)

**Part III**  
**Analysis, Synthesis and Forecasting of**  
**Ecological Data**

# Chapter 9

## Inferential Modelling of Population Dynamics

Friedrich Recknagel, Dragi Kocev, Hongqing Cao, Christina Castelo Branco, Ricardo Minoti, and Saso Dzeroski

**Abstract** This chapter introduces the design and applications of evolutionary algorithms and regression trees for inferential modelling of complex ecological data. Evolutionary algorithms prove to be superior tools for developing short-term forecasting models, revealing ecological thresholds and supporting quantitative meta-analyses as demonstrated exemplarily by means of the hybrid evolutionary algorithm (HEA). A case study of Lake Müggelsee (Germany) illustrates that models developed by HEA enable one to identify ecological thresholds and driving forces that perform short-term forecasting of population growth. The meta-analysis of Lakes Wivenhoe (Australia) and Lake Paranoa (Brazil) exemplifies the capability of models developed by HEA to test hypotheses on forcing functions of population growth across different environmental and climate conditions. Regression trees display fully transparent correlations between habitat properties and ecological entities. The tree induction process does not require prior assumptions, is fast and is not influenced by redundant variables and noise. Case studies for Lake Prespa (Macedonia) and land areas in Victoria (Australia) illustrate the capacity of regression trees to unravel complex ecological relationships.

---

F. Recknagel (✉) • H. Cao  
University of Adelaide, Adelaide, SA, Australia  
e-mail: [friedrich.recknagel@adelaide.edu.au](mailto:friedrich.recknagel@adelaide.edu.au); [hongqing.cao@adelaide.edu.au](mailto:hongqing.cao@adelaide.edu.au)

D. Kocev • S. Dzeroski  
Jozef Stefan Institute, Ljubljana, Slovenia  
e-mail: [dragi.kocev@ijs.si](mailto:dragi.kocev@ijs.si); [Saso.Dzeroski@ijs.si](mailto:Saso.Dzeroski@ijs.si)

C.C. Branco  
Federal University of the State of Rio de Janeiro, Rio de Janeiro, Brazil  
e-mail: [cbranco.unirio@gmail.com](mailto:cbranco.unirio@gmail.com)

R. Minoti  
University of Brasilia, Brasilia, Brazil  
e-mail: [rminoti@gmail.com](mailto:rminoti@gmail.com)

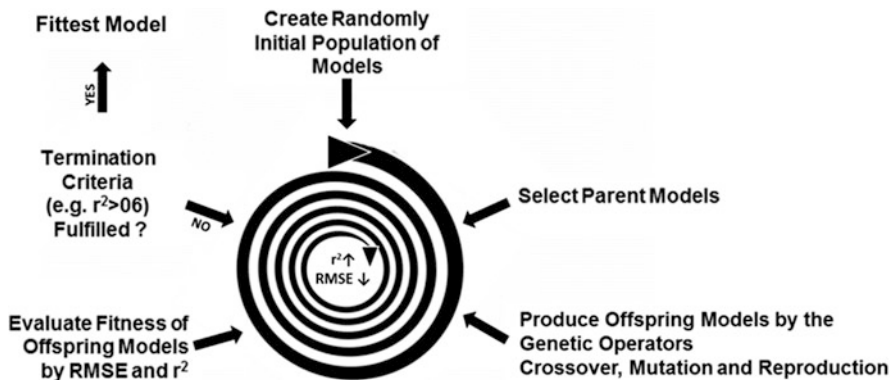
## 9.1 Introduction

Two novel computational techniques for inferential modelling will be presented that are suitable for unravelling and forecasting the complexity and evolving nature of ecosystems: evolutionary computation and regression trees.

Evolutionary computation (Holland 1975) is based on the cognitive principles of ‘generative creation’ and ‘choices over open-ended possibilities’ (Holland et al. 1986). It applies concepts of natural selection and evolution to induce multivariate models represented as IF-THEN-ELSE rules from complex data patterns where the IF part represents the condition and the THEN part the action. These IF-THEN-ELSE models are cyclically redesigned by genetic operations such as ‘cross-over’, ‘mutation’ and ‘reproduction’ until the fittest (best matching) model has been discovered. The fittest model represents relationships between targeted predictor and output variables that suit both elucidation and prediction. Figure 9.1 illustrates the concept of inferential modelling by evolutionary computation.

Holland’s two cognitive principles of evolutionary computation (Holland et al. 1986) appear highly relevant for studying ecological systems where ‘generative creation’ suits the evolving nature of ecosystems, and ‘choices over open-ended possibilities’ meets requirements to comprehend the complex stochastic nature of ecosystems. Resulting models are considered to be inferential (or empirical) and lack detailed process descriptions. However, the IF conditions of these models reveal explicit thresholds, and sensitivity functions between predictor and output variables quantify ecological relationships that represent events subject to modelling such as outbreaks of population density (e.g. Recknagel et al. 2014).

Regression trees are machine-learning methods for constructing predictive models by recursive partitioning of the data space and fitting a simple predictive model within each partition (Breiman et al. 1984; Loh 2011). As a result, the partitioning can be represented graphically as a decision tree. Regression trees



**Fig. 9.1** Inferential modelling by evolutionary computation (RMSE: root mean squared error;  $r^2$ : coefficient of determination)

suit output variables with either continuous or ordered discrete values, whereby its validity is typically measured by the root mean squared error (RMSE).

## 9.2 Inferential Modelling of Ecological Data by the Hybrid Evolutionary Algorithm

The hybrid evolutionary algorithm (HEA) (Cao et al. 2006, 2014, 2016; Recknagel and Ostrovsky 2016) has been designed to evolve fittest IF-THEN-ELSE models from ecological data by integrating genetic programming (GP) and differential evolution (DE) (Fig. 9.2). The fittest model is determined by the lowest root mean squared error (RMSE) and highest coefficient of determination ( $r^2$ ).

HEA applies GP according to Koza (1992) to evolve the optimum structure of the rule model, and DE according to Storn and Price (1997) to optimise the parameters of the rule model. Since GP typically operates on parse trees rather than on bit strings, it is well suited to evolve IF-THEN-ELSE rules for multivariate relationships. GP uses the logic functions  $FL = \{AND, OR\}$ , comparison functions  $FC = \{>, <, \geq, \leq\}$ , and arithmetic functions  $FA = \{+, -, *, /, \exp, \ln\}$  to represent IF-THEN-ELSE rules as a vector of multiple trees. Tree1 denotes the IF condition with the function set  $tree1 = FL \cup FC \cup FA$ , tree2 and tree3 respectively denote the THEN and ELSE branches with the function set  $F_{tree2/tree3} = FA$ .

Figure 9.3 illustrates one crossover step by GP for the optimisation of the IF trees of two parent models for 10-day-ahead forecasts of the abundance of the cyanobacterium *Anabaena* (cells  $mL^{-1}$ ) in Lake Wivenhoe, Australia (see also Chap. 15). Figure 9.3a, e represent two parent models. Figure 9.3b, c, f, g illustrate the selection of crossover points and the crossover between the IF trees of the parent models. Figure 9.3d, h represent the offspring models after the crossover.

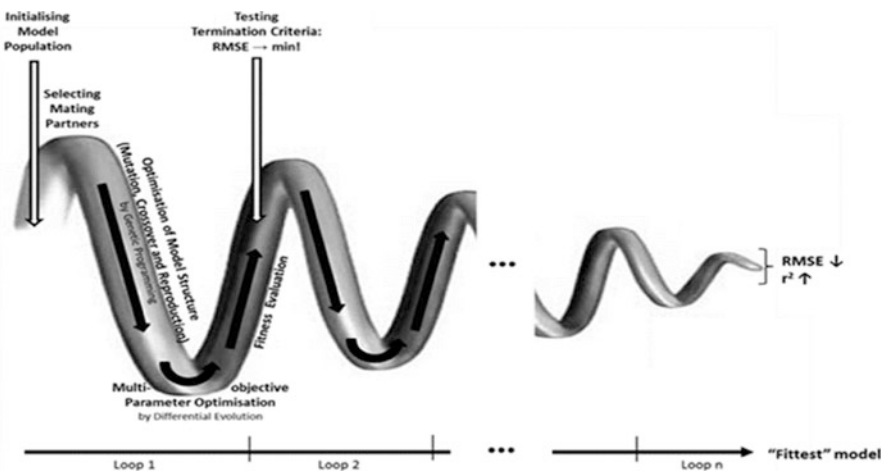
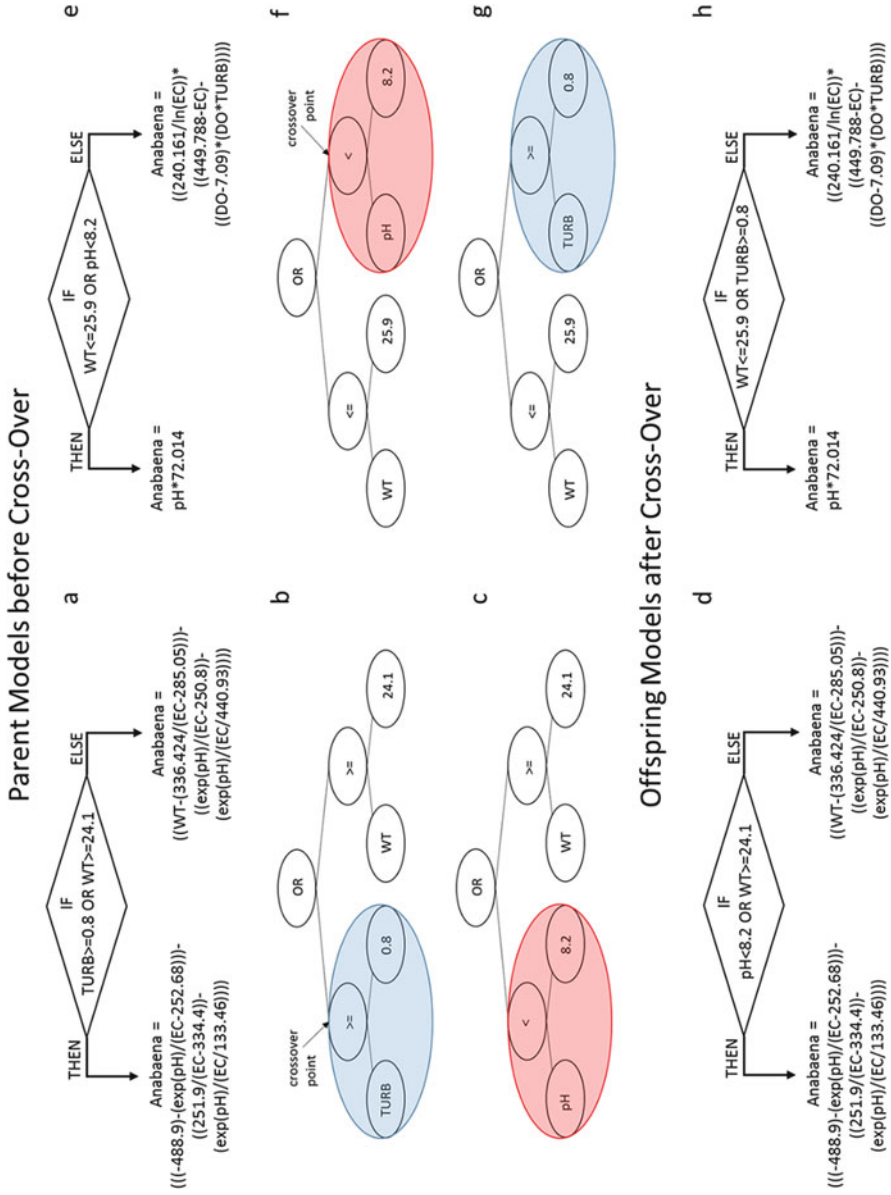


Fig. 9.2 Conceptual diagram of the functioning of the hybrid evolutionary algorithm (HEA) (Recknagel et al. 2013)





**Fig. 9.3** Genetic programming produces two offspring models by cross-over between the IF-trees of two parent models. **(a)** and **(e)**: parent models; **(b)** and **(f)**: selection of crossover points of IF-trees; **(c)** and **(g)**: IF-trees after crossover; **(d)** and **(h)**: offspring models (WT = water temperature °C, TURB = turbidity NTU, EC = electrical conductivity  $\mu\text{S cm}^{-1}$ , DO = dissolved oxygen  $\text{mg L}^{-1}$ )

Differential evolution (DE) extracts information on distance and direction of the current population of solutions towards a global optimum to guide the search for optimal parameters in the IF-THEN-ELSE rules. Since DE does not require separate probability distributions, the scheme becomes completely self-organizing. DE has been implemented in HEA for multi-objective parameter optimization as described by Cao et al. (2014).

A boot-strap training scheme is applied in HEA that selects randomly  $r_{max}$  data-subsets for training (75%) and testing (25%) for each of which  $t_{max}$  generations of models are evolved (Fig. 9.4). After  $r_{max}$  boot-strap runs are completed, it determines the overall ‘fittest model’ of all generations evolved by genetic programming and differential evolution. As a rule, 100 generations ( $r_{max} = 100$ ) prove to be sufficient for minimising the RMSE and approximating global optima of modelling experiments conducted by a Phoenix HPC Supercomputer.

The root mean squared error (RMSE) between the measured training data  $\hat{y}$  and predicted data  $y$  assesses the model fitness as follows:

$$\text{fitness} = \sqrt{\frac{1}{k} \sum_{i=1}^k (\hat{y}_i - y_i)^2}$$

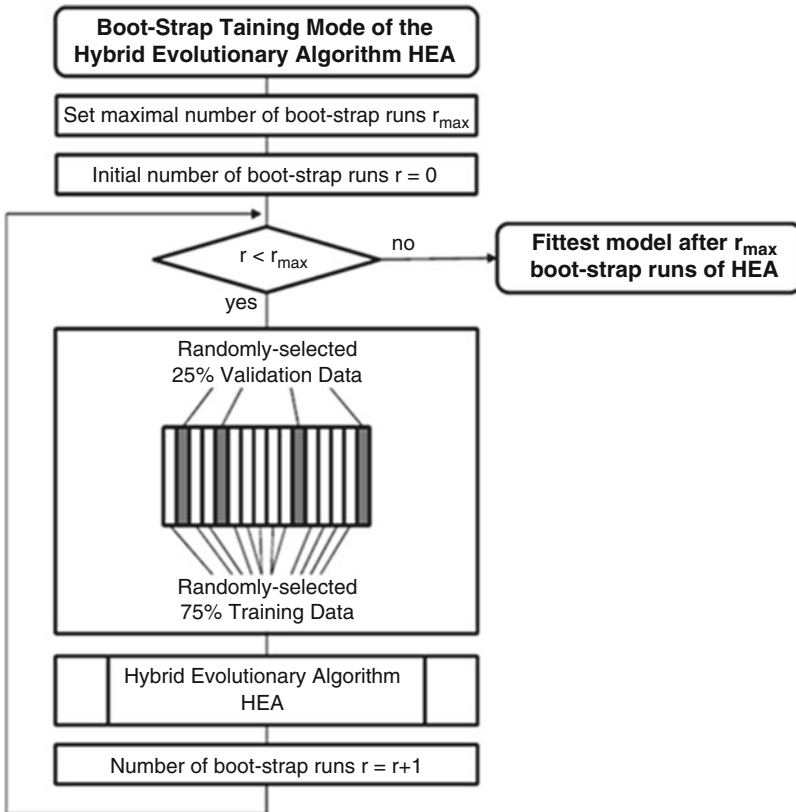


Fig. 9.4 Boot-strap training scheme of the hybrid evolutionary algorithm HEA

The models' performances are measured by means of coefficients of determination ( $r^2$ ). However, in accordance with Bennett et al. (2013), the visual comparison between measured and calculated data proved to be the most relevant approach for the validation of models related to this highly complex data.

HEA automatically carries out sensitivity analyses for the input variables of each discovered model. It calculates output trajectories separately for each input range (mean  $\pm$  SD) by keeping remaining input variables constant at mean values. Resulting sensitivity curves allow one to visualise the output trajectories in percentage terms (0–100%) within their range of each input.

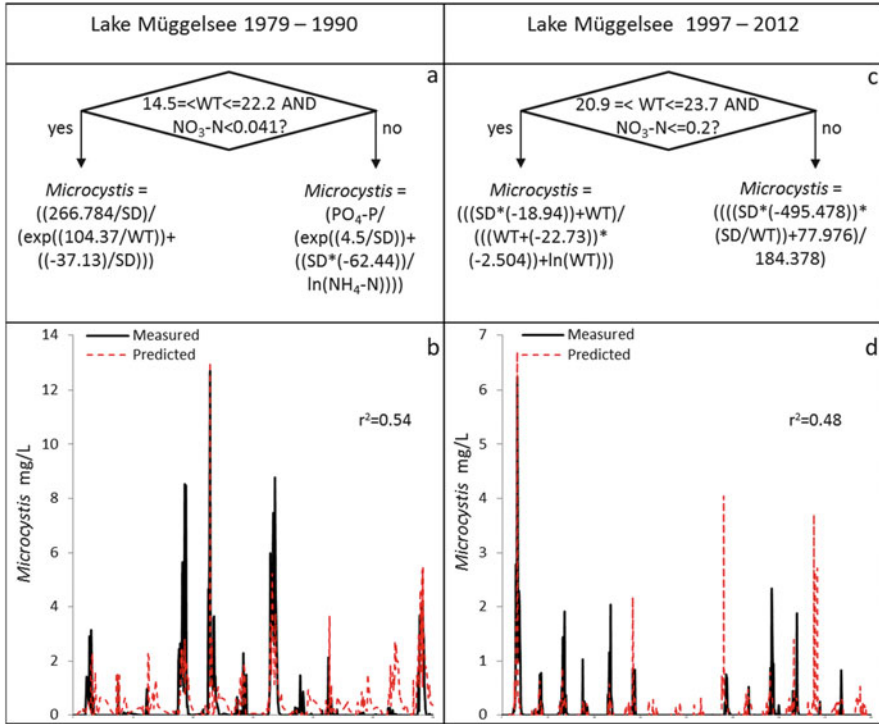
Since HEA infers models from data patterns, it requires cleansed and formatted cross-sectional or time-series data that are representative for the system to be modelled both in terms of number of observations and of relevance for the modelling purpose. Daily data interpolation is required to match dissimilar monitoring frequencies between physical, chemical and biological data, and to allow short-term forecasting for days ahead. Ecosystem evolution requires that models become regularly updated with the most recent data.

### 9.2.1 Population Dynamics of the Cyanobacterium *Microcystis* in Lake Müggelsee (Germany)

The hybrid evolutionary algorithm HEA has been applied to model population dynamics of the cyanobacteria *Anabaena*, *Aphanizomenon* and *Microcystis* in Lake Müggelsee (Germany) for two time periods that differed in the lake's trophic states (Recknagel et al. 2016). In Phase I from 1979 to 1990, the lake appeared to be hypertrophic; in Phase II from 1997 to 2012, decreasing external nutrient loads transformed the lake into a eutrophic state. The aim of this study was to model population dynamics of the three cyanobacteria for the two phases, and identify shifts in the models' IF conditions (thresholds) and input sensitivities in response to differences between the two phases. Here we present the results for *Microcystis* of this study. Table 9.1 summarizes weekly and biweekly measured limnological data of Lake Müggelsee that were used for this study.

**Table 9.1** Limnological data of Lake Müggelsee measured from 1979 to 1990 and 1997 to 2012

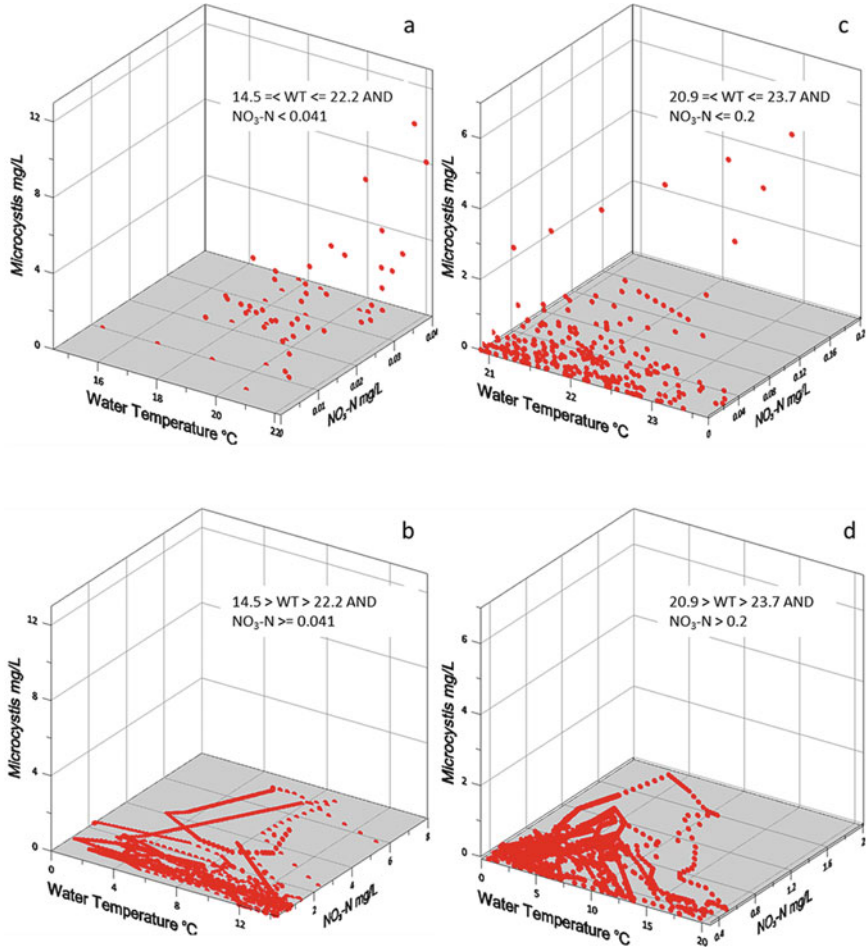
Limnological Variable	1979–90 Hypertrophic	1997–2012 Eutrophic
	Avg/Min/Max	
Water temperature (°C)	11.3/0.1/24.6	11.6/0.2/26.6
Secchi depth (m)	1.5/0.4/5.1	2/0.5/6.3
PH	8.2/7/9.4	8.3/7.1/9.5
PO <sub>4</sub> -P (µg L <sup>-1</sup> )	58.5/1/474	70.6/2/521
NO <sub>3</sub> -N (mg L <sup>-1</sup> )	1.2/0.004/7.4	0.33/0.01/1.91
NH <sub>4</sub> -N (mg L <sup>-1</sup> )	0.2/0.005/4.2	0.096/0.01/0.81
SiO <sub>2</sub> (mg L <sup>-1</sup> )	3.58/0.01/8.1	4.41/0.05/10.5
<i>Microcystis</i> (mg L <sup>-1</sup> )	0.36/0.001/12.67	0.082/0.0001/6.23



**Fig. 9.5** 7-day-ahead forecasting of *Microcystis* in Lake Müggelsee. Phase I: (a) IF-THEN-ELSE model, (b) validation of the model; Phase II: (c) IF-THEN-ELSE model, (d) validation of the model

One hundred models of *Microcystis* have been evolved by HEA for each phase based on repeated bootstrap runs. Figure 9.5 documents the *Microcystis* models with highest coefficients of determination ( $r^2$ ) and p-values less than 0.05 of the two phases. The IF conditions of the model for Phase I identified water temperatures between 14.5 and 22.2 °C and NO<sub>3</sub>-N concentrations less than 0.041 mg L<sup>-1</sup> as indicative for biomass greater than 1 mg/L. The model underestimated the biomass measured for *Microcystis* in 1982 but corresponded with the timing and magnitudes of peak biomass for the remaining years with an  $r^2 = 0.54$  (Fig. 9.5b). Even though the *Microcystis* model for Phase II selected the same threshold criteria their ranges were quite different compared to the model for Phase I with water temperatures between 20.9 and 23.7 °C and NO<sub>3</sub>-N concentrations less than 0.2 mg L<sup>-1</sup> (Fig. 9.5c) that distinguished *Microcystis* values below and above 0.5 mg L<sup>-1</sup> (Fig. 9.6c, d). It achieved an  $r^2 = 0.48$  but failed to forecast the high *Microcystis* biomass observed in 2002 and 2006 (Fig. 9.5d). When the IF conditions of the two models were tested with measured *Microcystis* data of the two phases as illustrated in Fig. 9.6, higher and lower biomass were clearly selected by these conditions.

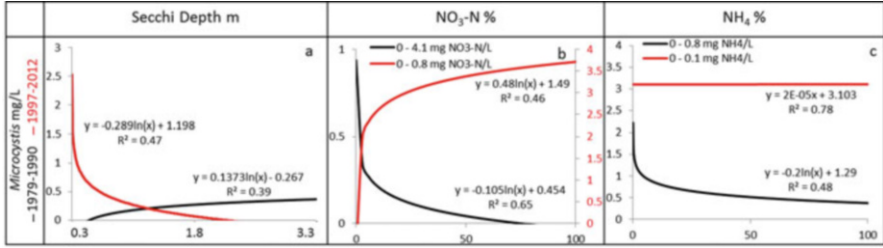
Results illustrated in Fig. 9.7 suggest that *Microcystis* becomes extinct at Secchi depths greater than 2 m in Phase II, but responds slightly positively to increasing



**Fig. 9.6** Functioning of the IF condition as threshold for forecasting of: (a) high population densities by the THEN equation and (b) low population densities by the ELSE equation of the model for *Microcystis* for Phase I; (c) high population densities by the THEN equation and (d) low population densities by the ELSE equation of the model for *Microcystis* for Phase II (Recknagel et al. 2016)

Secchi depths in Phase I (Fig. 9.7a). The finding for Phase I may indicate that *Microcystis* withstands underwater light limitation at lower nutrient limiting conditions by buoyancy enabled by its internal gas vesicles also reflected by almost neutral sensitivity to Secchi depth.

Even though NO<sub>3</sub>-N concentrations were up to 4 times higher in Phase I as compared to Phase II, *Microcystis* displayed highest biomass at lowest NO<sub>3</sub>-N concentrations and *vice versa* in Phase I (Fig. 9.7b) indicating seasonally high N consumption by all phytoplankton phyla. In Phase II *Microcystis* biomass grew



**Fig. 9.7** Average sensitivity functions of 20 cyanobacterium-specific models for phase I and phase II revealing relationships between: (a) Secchi depth and *Microcystis*; (b) NO<sub>3</sub>-N and *Microcystis*; and (c) NH<sub>4</sub> and *Microcystis*

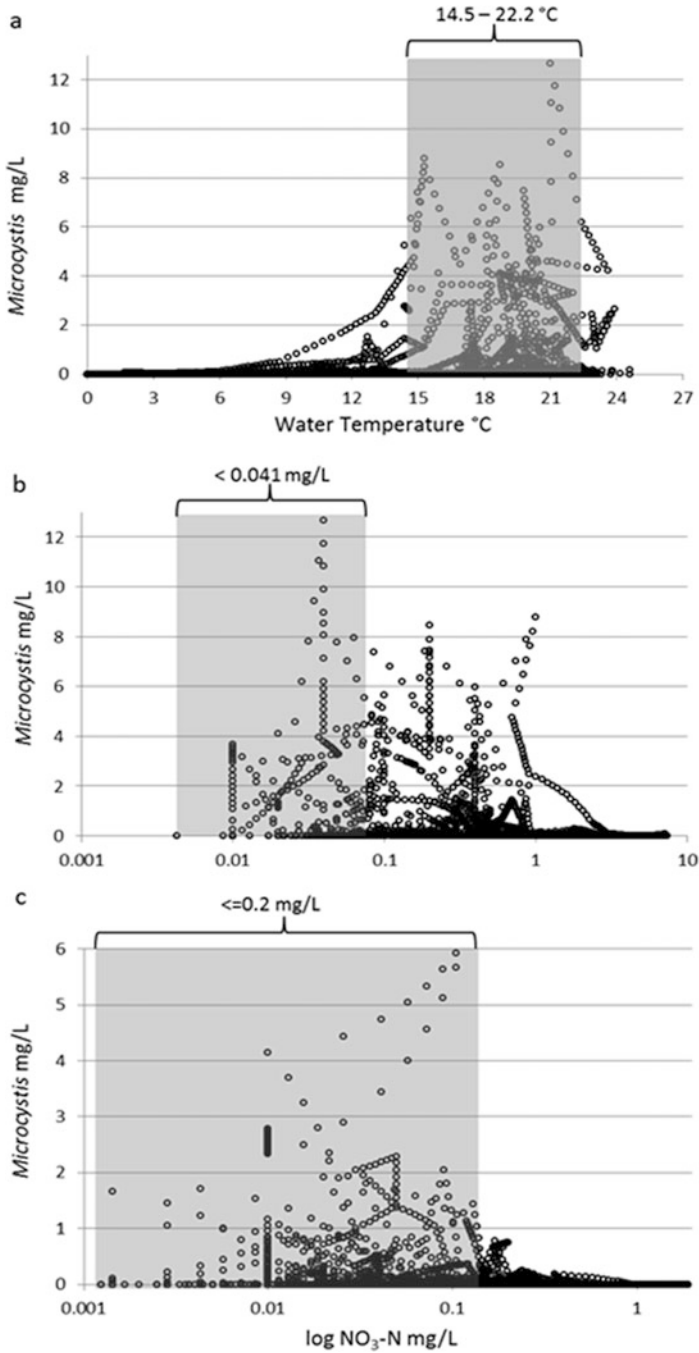
hyperbolically with increasing NO<sub>3</sub>-N concentration towards a plateau near 3.5 mg L<sup>-1</sup>. Figure 9.7b demonstrates that the growth of *Microcystis* is inhibited by the extended nitrogen limitation in Phase II. *Microcystis* appeared to respond neutrally to changes in ammonium concentrations in Phase II.

In summary, the models for *Microcystis* forecast the timing of peaking biomass in both phases. However magnitudes of high peak events were sometimes underestimated and magnitudes of low peak events overestimated.

Threshold conditions of *Microcystis* models indicating biomass greater than 4 mg L<sup>-1</sup> in Phase I and greater than 0.5 mg L<sup>-1</sup> in Phase II included water temperature and concentrations of NO<sub>3</sub>-N. The temperature range between 14.5 and 22.2 °C in Phase I was somewhat surprising since optimum temperatures for *Microcystis* growth are known to be near 23 °C (Reynolds 1984) which is well matched by the temperature range of 20.9 to 23.7 °C in Phase II. However Fig. 9.8a confirms that biomass of 6 to 8 mg L<sup>-1</sup> has been observed at temperatures near 15 °C in Phase I and is most likely related to the transition from late summer to autumn, and thus been recognised as ‘high biomass’ by the model. The NO<sub>3</sub>-N concentrations that were identified by the models to coincide with high biomass of *Microcystis* in summer were less than 0.041 mg L<sup>-1</sup> for Phase I and smaller than 0.2 mg L<sup>-1</sup> for Phase II. These findings by the models correspond with the observed relationships between *Microcystis* and NO<sub>3</sub>-N as shown in Fig. 9.8b, c. In Phase I there is a pattern of biomass greater than 8 mg L<sup>-1</sup> at NO<sub>3</sub>-N concentrations below 0.05 mg L<sup>-1</sup>, reflecting the greatest N-consumption at highest biomass of *Microcystis* whilst highest biomass of greater than 1 mg L<sup>-1</sup> occurs at NO<sub>3</sub>-N concentrations below 0.2 mg L<sup>-1</sup> in Phase II.

The case study of Lake Müggelsee allows the following conclusions to be drawn:

1. inferential models developed by evolutionary computation via HEA can achieve good forecasting accuracy for fast growing harmful populations such as the cyanobacterium *Microcystis*;
2. IF conditions of such inferential models provide information on thresholds that indicate rapid population growth; and



**Fig. 9.8** Relationship between: (a) observed water temperatures and biovolumes of *Microcystis* in Lake Müggelsee in Phase I, (b) observed NO<sub>3</sub>-N concentrations and biovolumes of *Microcystis* in Lake Müggelsee in Phase I, and (c) observed NO<sub>3</sub>-N concentrations and biovolumes of *Microcystis* in Lake Müggelsee in Phase II (Recknagel et al. 2016)





3. sensitivity functions of such models provide information on key driving variables of population growth.

### 9.2.2 Meta-Analysis of Population Dynamics of the Cyanobacterium

**Cylindrospermopsis in Lake Wivenhoe (Australia) and Lake Paranoa (Brazil)**  
 Meta-analysis becomes more conclusive for ecological applications if time-series data and quantitative models can be included (Osenberg et al. 1999). Here we carry out a meta-analysis of population dynamics of the tropical cyanobacterium *Cylindrospermopsis* with models developed by HEA from 6 years of time-series data of the subtropical Lake Wivenhoe and the tropical Lake Paranoa. Basic properties of the two lakes are summarised in Table 9.2. Both lakes experience annual recurring blooms of *Cylindrospermopsis* that severely disrupt water supply and cause higher water treatment costs.

Three years with low and 3 years with high abundances of *Cylindrospermopsis* were selected for modelling from 17 years of time-series data at each lake. Figure 9.9 indicates the 6 years of data from each lake, and displays the mean annual trajectory as well as the min-max envelope of *Cylindrospermopsis*. It

**Table 9.2** Basic characteristics of the Lakes Wivenhoe and Paranoa

	Lake Wivenhoe 1997–2015 Australia	Lake Paranoa 1980–1997 Brazil
		
Morphometry	Max depth = 28m Surface area = 107.5 km <sup>2</sup>	Mean/max depth = 12/38m Surface area = 38 km <sup>2</sup>
Climate	Subtropical	Tropical
Circulation type	Warm-monomictic	Warm-monomictic
Trophic state	Mesotrophic	Eutrophic
Water temperature °C Min/Max/Avg	14.8/30.2/22.7	19.4/29/23.8
TN/TP Min/Max/Avg	6.2/71/28.2	
DIN/DIP Min/Max/Avg		3.1/1120.5/186.35
EC μS cm <sup>-1</sup> Min/Max/Avg	191.7/575.25/350.2	27/103.6/59.9
<i>Cylindrospermopsis</i> Min/Max/Avg	1/173166/9557.4 cells mL <sup>-1</sup>	1 / 5919965 / 416832.7 cells mL <sup>-1</sup>



indicates that the eutrophic Lake Paranoa encounters much higher abundances of *Cylindrospermopsis* dominated by the extreme bloom event in summer 1990/91 with 6 million cells mL<sup>-1</sup> than the mesotrophic Lake Wivenhoe. While the typical growing season in Lake Wivenhoe lasts from November through February, Lake Paranoa experiences high abundances from September through April.

The meta-analysis focused on the question of whether growth of *Cylindrospermopsis* appeared to be favoured by similar water temperatures and nutrient thresholds in spite of the fact that the two lakes differ in climate and trophic state. Models have therefore been developed by HEA from the 6 years of each lake, solely driven by water temperature and by the N/P ratio. As the results in Fig. 9.10 show, coefficients of determination of  $r^2 = 0.47$  and  $r^2 = 0.36$  have been achieved by the water temperature-driven models for Wivenhoe and Paranoa, respectively (see Fig. 9.10b, e). In both cases the models corresponded in terms of average timing and magnitudes of population growth, and revealed that Paranoa seems to have a lower temperature threshold of 25 °C for fast growth of *Cylindrospermopsis* compared to 27.7 °C in Wivenhoe (Fig. 9.10c, f). This might be due to the fact that its annual average temperature gradient of 9.6 °C is noticeably smaller than that of Wivenhoe with 15.4 °C. Figure 9.12a, c display relationships between water temperature and cell division rates of *Cylindrospermopsis* simulated by the two models. In the case of Wivenhoe, it shows that fast rates correspond with rising temperatures, and cell division ceases when temperatures drop below 20 °C. The same trend can be observed for Paranoa.

Figure 9.11 compares the N/P-driven models developed by HEA whereby total nitrogen (TN) and total phosphorus (TP) data were available for Wivenhoe, and dissolved inorganic nitrogen (DIN) and dissolved inorganic phosphorus (DIP) were available for Paranoa. Simulation results for both lakes corresponded well with observed population dynamics of *Cylindrospermopsis* even though lower coefficients of determination of  $r^2 = 0.34$  and  $r^2 = 0.15$  were achieved. Whereas no distinct TN/TP threshold for fast population growth has been discovered for Wivenhoe (Fig. 9.11c), the surpassing of the DIN/DIP threshold of 332.2 seemed to be defining for the distinct bloom event in Paranoa during the summer 1990/91 (Fig. 9.11f).

Figure 9.12b, d display relationships between N/P ratios and cell division rates of *Cylindrospermopsis* simulated by the two models where criteria for P- or N-deficient growth are denoted based on TN/TP-values defined by Sterner (2008) and for DIN/DIP-values defined by Redfield (1958). Most years at both lakes show P-deficient conditions with no obvious limiting effects on cell division rates. However, periods of no cell division seem to correspond with declining N/P ratios.

Based on the outcomes of the meta-analysis, several conclusions can be drawn:

1. Inferential models enhance quantitative meta-analysis.
2. Models solely driven by water temperature simulated major growth events of *Cylindrospermopsis* in both lakes. Results indicate that the mesotrophic Lake Wivenhoe has a higher temperature threshold of 27.7 °C for driving fast growth than Lake Paranoa at 25 °C. Higher water temperatures might be required to

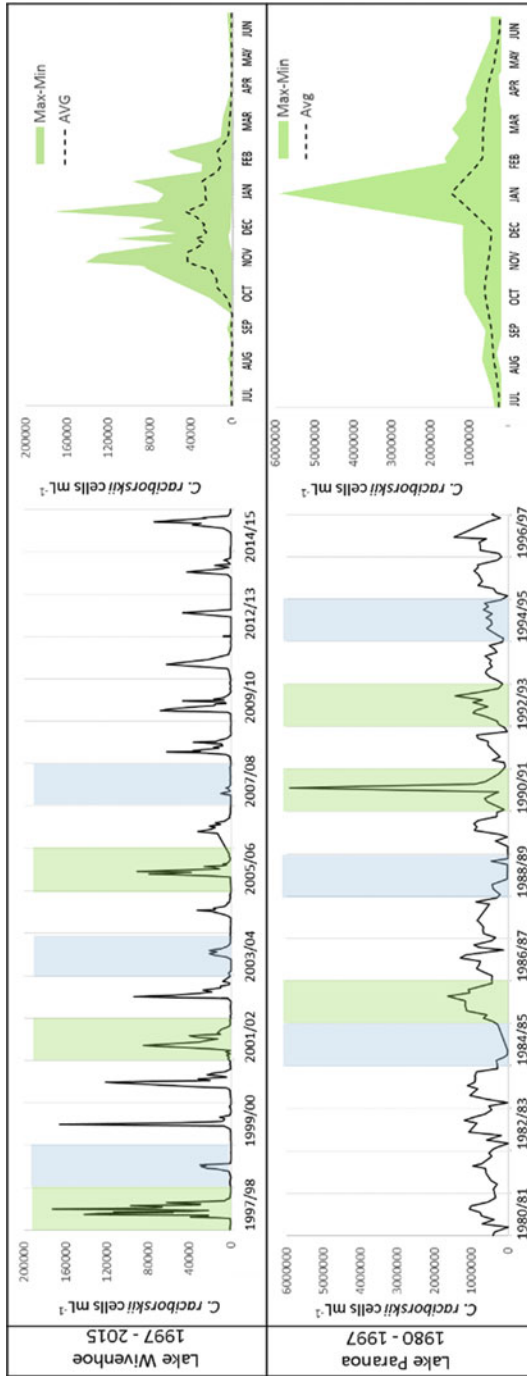
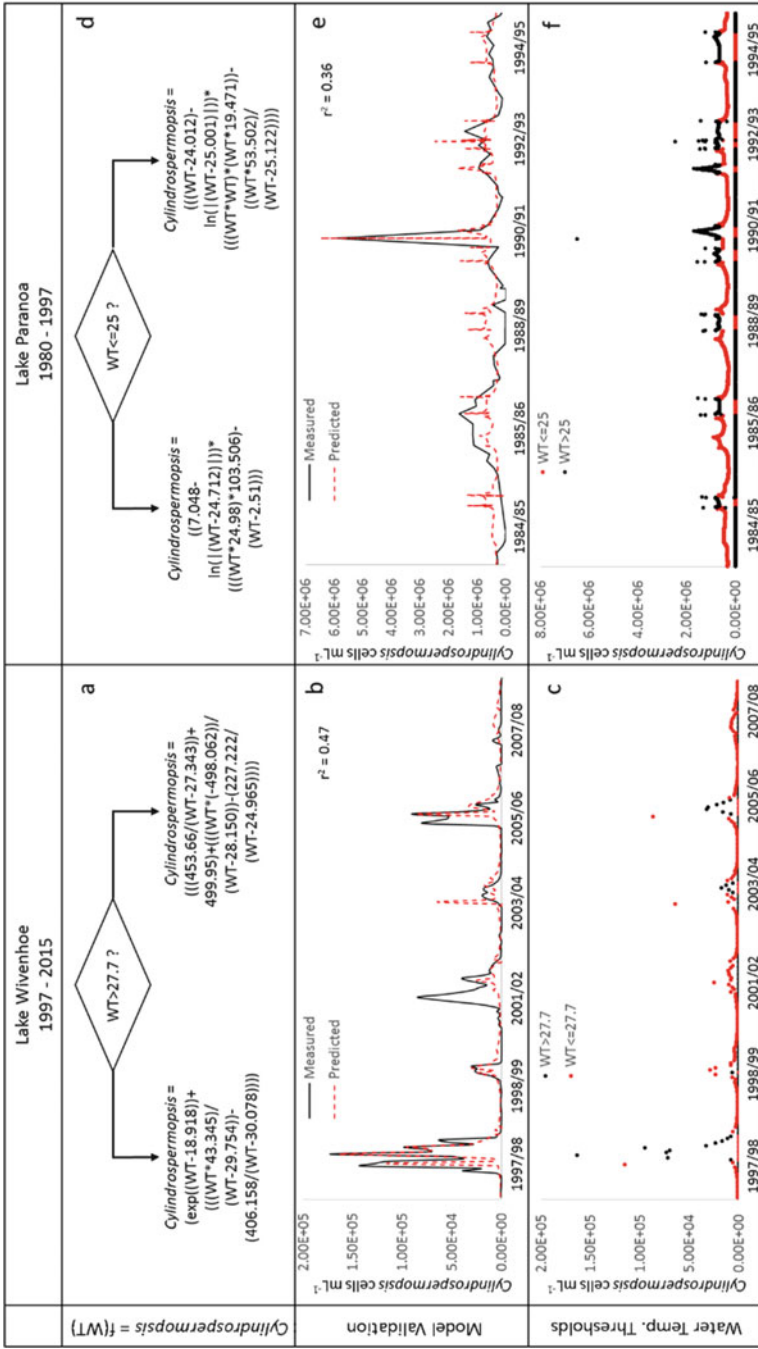


Fig. 9.9 Measured trajectories of 17 years and averaged annual trajectories of 3 selected 'high abundance' years (light green) and 3 selected 'low abundance' years (light blue) of Lake Wivenhoe (top row) and Lake Paranao (bottom row)



**Fig. 9.10** Water temperature dependent *Cylindrospermopsis* models for Lakes Wivenhoe and Paranao: (a) and (d) IF-THEN-ELSE model; (b) and (e) model validation; and (c) and (f) water temperature thresholds

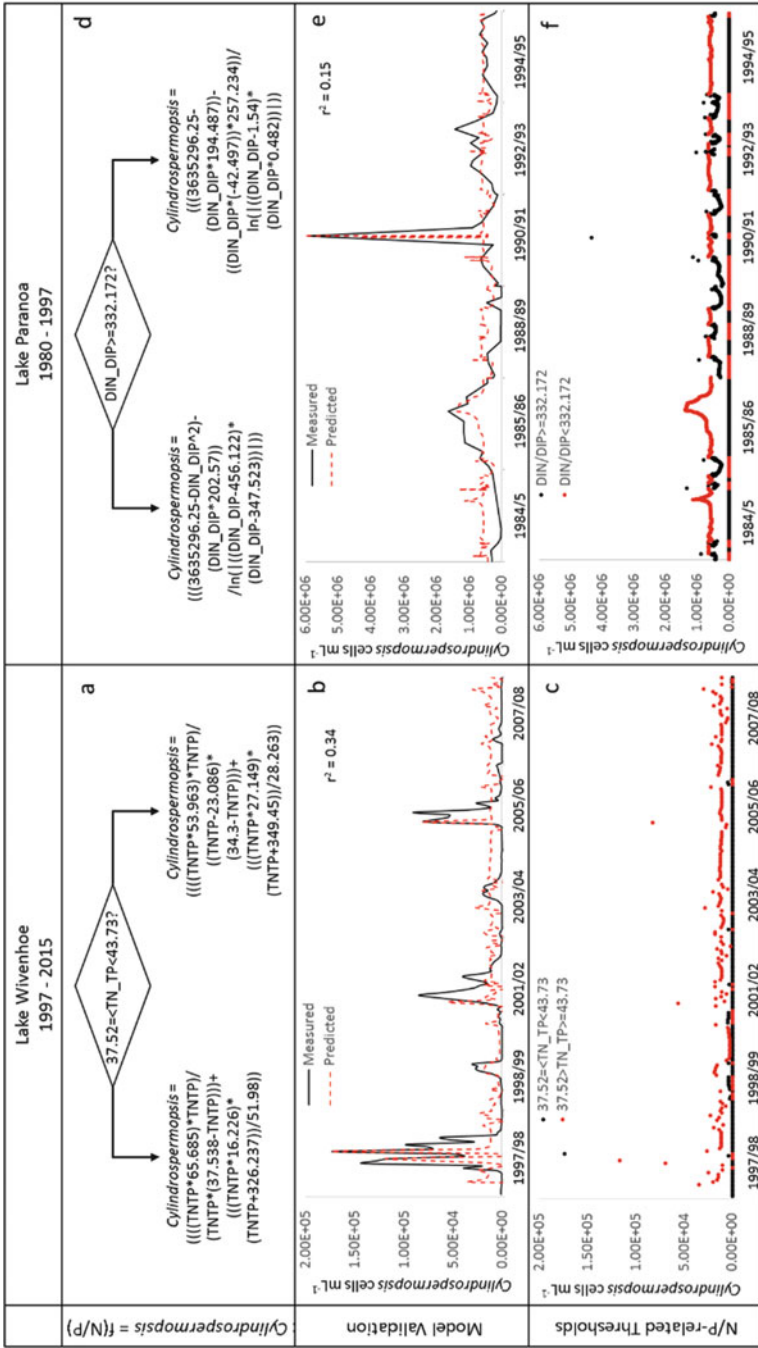
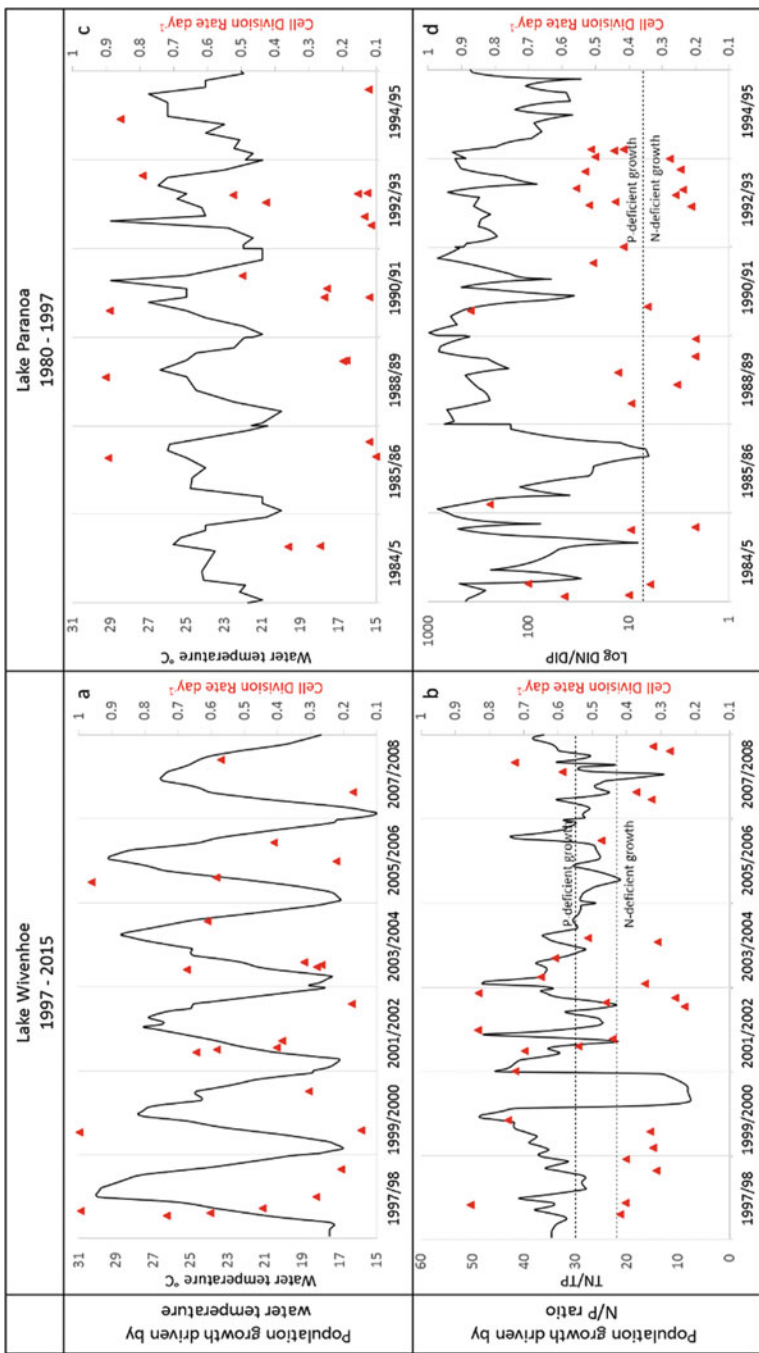


Fig. 9.11 N/P-dependent *Cylindrospermopsis* models for Lakes Wivenhoe and Paranao: (a) and (d) IF-THEN-ELSE model; (b) and (e) model validation; and (c) and (f) N/P thresholds



**Fig. 9.12** Forecasted *Cylindrospermopsis* cell division rates for Lakes Wivenhoe and Paranoa: (a) and (c) relationships between water temperature and cell division rates; (b) and (d) relationships between N/P and cell division rates

stimulate growth at **comparatively** low nutrient concentrations, whilst the eutrophic Lake Paranoa provides *a priori* favourable nutrient conditions for growth.

3. Models solely driven by N/P ratios simulated major growth events of *Cylindrospermopsis* in both lakes reasonably well, but failed to simulate some of the seasonal events. Irrespective of the fact that both lakes experienced phosphorus limitation during most of the 6 years, growth of *Cylindrospermopsis* was seemingly not affected. This suggests that *Cylindrospermopsis* may utilize alternative nutrient sources at low phosphorus concentrations such as sulfolipids—i.e. known as sulphur-for-phosphorus strategy (Van Mooy et al. 2006), and equally likely, tend to have large internal P-stores (Mayberly, pers. comm.).

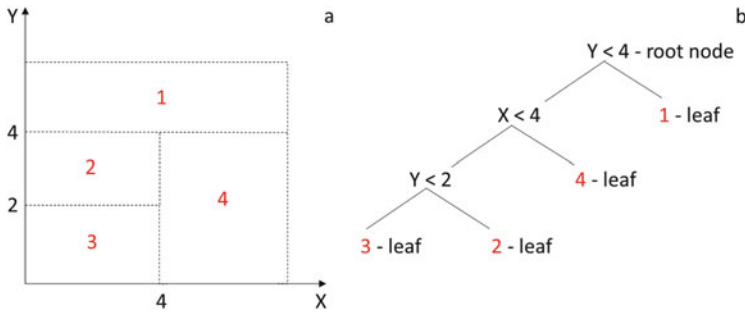
### 9.3 Inferential Modelling of Ecological Data by Regression Trees

Regression trees are hierarchical structures, where the internal nodes contain tests on the independent variables. Each branch of an internal test corresponds to an outcome of the test, and the prediction for the value of the dependent variable is stored in a leaf. Each leaf of a regression tree contains a constant value as a prediction for the dependent variable (regression trees represent piece-wise constant functions).

To obtain the prediction for a new set of data, the data is sorted down the tree, starting from the root node. For each internal node that is encountered on the path, the test that is stored in the node is applied. Depending on the outcome of the test, the path continues along the corresponding branch (to the corresponding subtree). The resulting prediction of the tree is taken from the leaf at the end of the path. The tests in the internal nodes can have more than two outcomes (this is usually the case when the test is on discrete-valued attributes where a separate branch/subtree is created for each value). Typically each test has two outcomes: the test has succeeded or the test has failed. The trees in this case are also called binary trees. Figure 9.13 displays schematically how a regression tree is constructed for data of the function  $Y = f(X)$ .

#### 9.3.1 Induction Algorithm of Regression Trees

A regression tree is usually constructed by a recursive partitioning algorithm from a training set of data (Breiman et al. 1984). The algorithm is known as top-down induction of decision trees (TDIDT). The data include measured values of the independent and dependent variables. The tests in the internal nodes of the tree



**Fig. 9.13** Representing the function  $Y = f(X)$  (a) as regression tree (b)

refer to the independent variables, while the predicted values in the leaves refer to the dependent variables.

The TDIDT algorithm starts by selecting a test for the root node. Based on this test, the training set is partitioned into subsets according to the test outcome. In the case of binary trees, the training set is split into two subsets: one containing data for which the test succeeds (typically the left subtree) and the other containing data for which the test fails (typically the right subtree). This procedure is recursively repeated to construct the subtrees.

The partitioning process stops if a stopping criterion is satisfied (e.g., the number of data in the induced subsets is smaller than the predefined depth/size of the tree exceeds some predefined value, etc.). In that case, the prediction vector is calculated and stored in a leaf. The components of the prediction vector are the mean values of the dependent variables calculated over the records that are sorted into leaves.

One of the most important steps in the tree induction algorithm is the test selection procedure. For each node, a test is selected by using a heuristic function computed on the training data. The goal of the heuristic is to guide the algorithm towards small trees with good predictive performance. The heuristic used typically for selecting the attribute tests in the internal nodes is intra-cluster variation summed over the subsets induced by the test. Intra-cluster variation of a group/cluster of samples  $C$  is defined as

$$\text{Var}(C) = \frac{1}{2|C|^2} \sum_{X \in C} \sum_{Y \in C} d^2(X, Y)$$

with  $N$  the number of samples in the cluster and  $d$  is the distance function between the samples (Euclidean distance is typically used). Lower intra-subset variance results in predictions that are more accurate.

The tree induction algorithm is illustrated in Fig. 9.14. It starts by finding the best split on the complete space (top-right corner) and then recursively partitions the

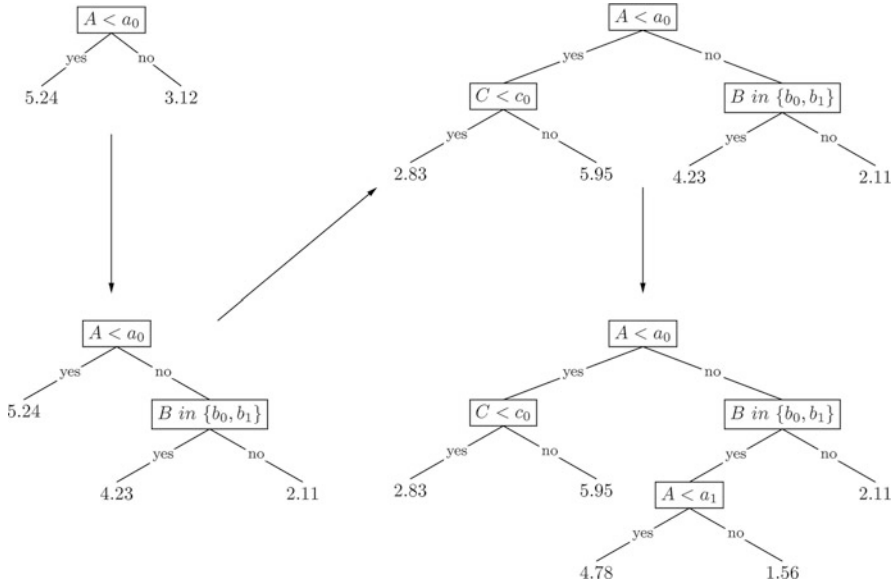


Fig. 9.14 The top-down induction of regression trees algorithm

examples into smaller groups: first adds an additional split instead on the right split, then on the left split. The tree learning stops when some stopping criterion is met.

### 9.3.2 Pruning of Regression Trees

The different stopping criteria can be used to deal with noise and other types of imperfections in the data. This is also known as tree pruning. The most widely used tree pruning approaches include: (1) reduced error pruning; (2) maximal depth; (3) minimal instances in a leaf; and (4) F-test pruning. First, reduced error pruning is one of the most straightforward forms of pruning. It uses an additional validation set of examples to remove the subtrees that are not improving the performance. Namely, it starts at the leaves, and replaces each node with a leaf if the predictive performance (as estimated with the validation set) is not affected.

Second, the maximal depth algorithm takes as input a user defined integer value that specifies the maximally allowed depth for the leaves in the tree. Third, the minimal instances in a leaf algorithm also takes as input a user defined integer value that specifies the minimal number of instances that each leaf of the tree must contain. Fourth, the F-test pruning algorithm checks whether the addition of a split at a given leaf of the tree significantly reduces the intra-cluster variance for the examples in that leaf: The significance level is specified by the user. These pruning algorithms increase the interpretability of trees, while maintaining (or even increasing) their predictive performance.



### 9.3.3 Diatom Populations in Lake Prespa (Mazedonia)

Lake Prespa is located at the border intersection of Macedonia, Albania and Greece. Water quality and diatom populations have been monitored biweekly at 14 sites of Lake Prespa from March 2005 to September 2006. The sampling sites were located in Macedonia (8), Albania (3) and Greece (3), and were considered to be representative for assessing eutrophication effects (Krstić 2005).

The following water quality parameters were measured: temperature, dissolved oxygen, Secchi depth, conductivity, alkalinity ( $pH$ ), nitrogen compounds ( $NO_2$ ,  $NO_3$ ,  $NH_4$ , organic and total nitrogen), sulphur oxide ions ( $SO_4$ ), phosphorus ( $P$ ), sodium ( $Na$ ), potassium ( $K$ ), magnesium ( $Mg$ ), copper ( $Cu$ ), manganese ( $Mn$ ) and zinc ( $Zn$ ). A summary of the water quality data is included in Table 9.3.

Pelagic diatoms were sampled by plankton nets and benthic diatoms collected from submerged plants, rocks and sediments. The relative abundance of identified species was related to the total cell count per sampling site (see also Kocev et al. 2010), and the 10 most abundant diatom species are listed in Table 9.4.

The dataset was used to learn regression trees whereby water quality parameters were considered as independent variables and the 10 most abundant diatom species were considered as dependent variables. The modelling aimed to identify habitat conditions that best suit the 10 diatom species. The learned regression trees are illustrated in Fig. 9.15. We used three pruning algorithms to obtain these trees as follows. First, we set the maximal depth parameter to 3. Second, we set the minimum number of examples in a leaf to 16. Third, we set the significance level for the F-test pruning at 0.05.

Here, we discuss the models for *Cyclotella ocellata* (COCE), *Cavinula scutelloides* (CSCU) and *Navicula prespanense* (NPRE). The most abundant diatom according to the measured data, *Cyclotella ocellata* (COCE), is mostly influenced by the nitrogen compounds, the temperature and the conductivity of the water and the potassium ( $K$ ) concentration. The concentration of nitrates (i.e., nutrients) positively influences the abundance: the leaves of the tree with higher nitrate concentration (the first three leaves) contain samples with higher abundances of *Cyclotella ocellata* than the leaves on the right-hand side of the tree.

The habitat model for *Cavinula scutelloides* shows that the temperature, nitrates and metal concentrations are most influential for this diatom species. Higher temperatures favor this specific diatom species, while optimal concentrations of magnesium (between 6.13 and 9.44  $\mu\text{g}/\text{dm}^3$ ) are needed for the highest abundance. This habitat model also reveals the limiting role of manganese and copper for this species: the higher values of these metals at lower temperatures reduce the abundance of the species.

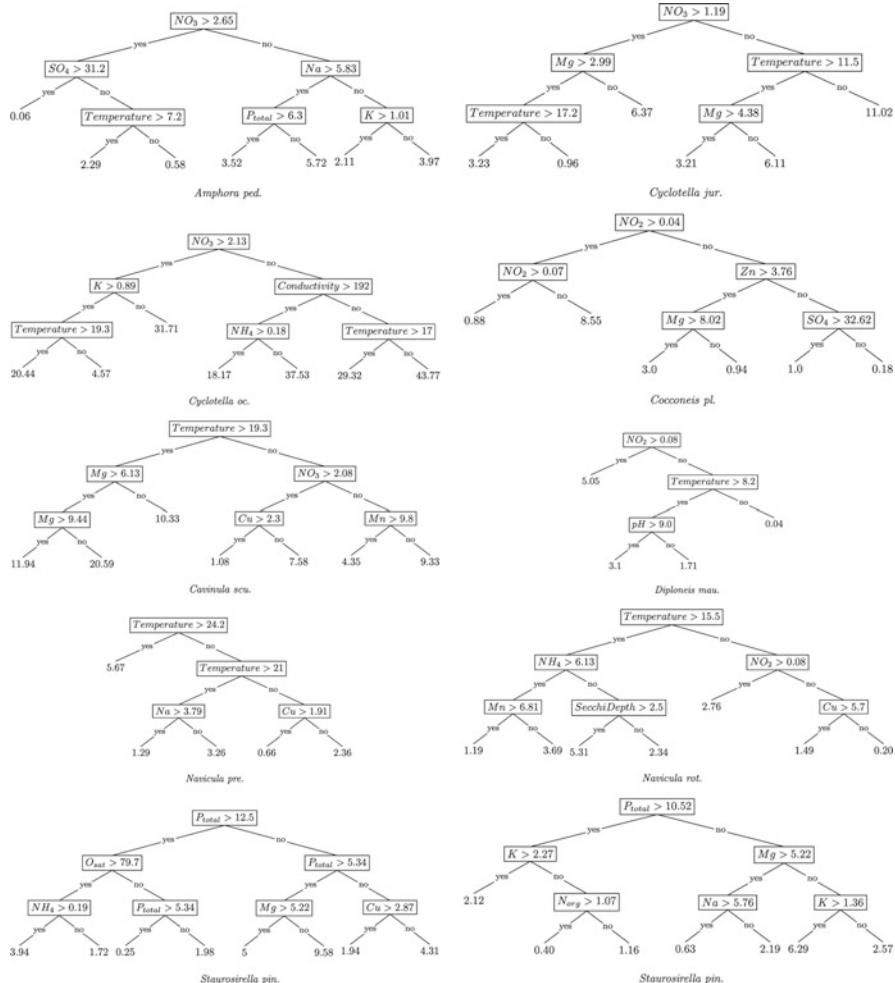
The temperature and concentrations of sodium ( $Na$ ) and copper ( $Cu$ ) are most important for the abundance of the *Navicula prespanense* (NPRE) diatoms. These diatoms are most abundant at water temperatures higher than 24.2 °C—temperatures characterize summer periods. These diatoms are absent (or present in very small numbers) at temperatures lower than 21 °C and when the concentration of copper exceeds 1.91  $\mu\text{g}/\text{dm}^3$ .

**Table 9.3** Statistical summary of water quality data of 218 samples collected at 14 locations across Lake Prespa, Macedonia

	Minimum	Maximum	Mean value	Standard deviation		Minimum	Maximum	Mean value	Standard deviation
Temperature (°C)	2.9	26.8	15.56	6.61		0.01	0.83	0.22	0.14
Dissolved oxygen (mg/dm <sup>3</sup> )	0.7	12.6	8.04	1.99	Inorganic N (mg/dm <sup>3</sup> )	0.02	8.41	1.83	1.10
Saturated oxygen (%)	6.6	114.19	83.07	19.54	Organic N (mg/dm <sup>3</sup> )	2.68	266.1	29.47	22.98
Deficit oxygen (mg/dm <sup>3</sup> )	-9.32	1.33	-1.73	2.02	SO <sub>4</sub> (mg/dm <sup>3</sup> )	1.15	83.13	18.63	15.31
Secchi depth (m)	1.8	5.4	3.09	0.76	Total P (µg/dm <sup>3</sup> )	0.75	13.15	4.36	2.10
Conductivity (µS/cm)	142.5	318	196.23	27.84	Na (mg/dm <sup>3</sup> )	0.23	4.8	1.50	0.66
pH factor	5.5	24.8	8.68	2.86	K (mg/dm <sup>3</sup> )	1.11	19.45	5.70	2.84
NO <sub>2</sub> (mg/dm <sup>3</sup> )	0	0.44	0.03	0.05	Mg (µg/dm <sup>3</sup> )	1.04	23.3	3.97	2.79
NO <sub>3</sub> (mg/dm <sup>3</sup> )	0	13.4	2.07	2.13	Cu (µg/dm <sup>3</sup> )	0.88	230	7.88	16.79
NH <sub>4</sub> (mg/dm <sup>3</sup> )	0.01	1.07	0.29	0.18	Mn (µg/dm <sup>3</sup> )	0.27	22.7	5.23	4.42
Total N (mg/dm <sup>3</sup> )	0.09	9.07	2.07	1.12	Zn (µ/dm <sup>3</sup> )				

**Table 9.4** Acronyms of the 10 most abundant diatoms species

Acronym	Diatom species	Acronym	Diatom species
APED	Amphora pediculus	DMAU	Diploneis mauleri
CJUR	Cyclotella juriljii	NPRE	Navicula prespanense
COCE	Cyclotella ocellata	NROT	Navicula rotunda
CPLA	Cocconeis placentula	NSROT	Navicula subrotundata
CSCU	Cavinula scutelloides	STPNN	Stausosirella pinnata



**Fig. 9.15** Individual habitat models for the 10 most abundant diatom species in Lake Prespa

### 9.3.4 *Vegetation Status of Selected Land Sites in Victoria (Australia)*

In this study, we used vegetation data from 16,967 terrestrial sites in Victoria (Australia) acquired using the habitat hectares approach (Parkes et al. 2003)—a rapid assessment technique of vegetation condition developed primarily for biodiversity conservation planning. ‘Vegetation quality’ in the habitat hectares approach is defined as the degree to which the current vegetation differs from a ‘benchmark’ that represents the average characteristics of a mature and long-undisturbed stand of the same plant community. Against the benchmark, the decline in quality can be estimated for each vegetation type and dissimilar community assemblages; e.g. rainforests and savannahs can be compared by employing the same general index. This general approach has become a standard method used to quantify the condition of habitat within the state of Victoria and has been emulated to some degree by other jurisdictions within Australia (Gibbons et al. 2009).

The habitat hectares score is the weighted sum of 7 site and 3 landscape scale metrics. The landscape components of the ‘habitat hectares’ score can be readily rendered spatially within a GIS using tools such as FRAGSTATS (McGarigal et al. 2002) and have not been further considered in this study. The objective is to make spatially explicit predictions of the 7 site scale components of the habitat hectares score (hereafter referred to as the habitat hectares site score or HHSS).

Each of the 16,967 sampling point based on the ‘habitat hectares’ approach is described by 40 independent (or feature) variables (GIS and remote-sensed data with a pixel resolution of  $30 \times 30$  m) and 7 dependent (or target) variables (the HHSS). The HHSS is a numeric variable composed as a weighted average of the following components: Large Trees; Tree (canopy) Cover; Understorey (non-tree) Strata; Lack of Weeds; Recruitment; Organic Litter; and, Logs. Apart from Lack of Weeds, each component score was calculated comparing the current status of the vegetation with a benchmark.

The *Large trees score* represents the number of large trees (both living and dead) that are present at the measuring site (compared to the ‘benchmark’ archetype). The *Tree Canopy score* assesses the projective foliage cover of canopy trees in the stand, while the *Understorey score* assesses the abundance and diversity of various shrubs and forb/herb strata of a community. The understorey assessment includes only indigenous plant species. The *Lack of weeds score* is calculated from the cover of non-indigenous weed species. The *Recruitment score* provides an indication of the level of regeneration of woody plant species and could be seen as a surrogate measure of the long-term viability of the site’s structural characteristics. *Litter* represents both fine and coarse plant debris less than 10 cm diameter, while *Logs* represent the fallen timber or branches of trees that are substantially detached from the parent tree. An unabridged description of the habitat hectares scores and methods can be found in Parkes et al. (2003). The 40 independent variables include 39 continuous variables and one categorical variable (Table 9.5). The categorical variable *LandCover* surface was derived from Landsat 7 TM spectral data. Classes

Table 9.5 Basic statistics for the remotely sensed variables

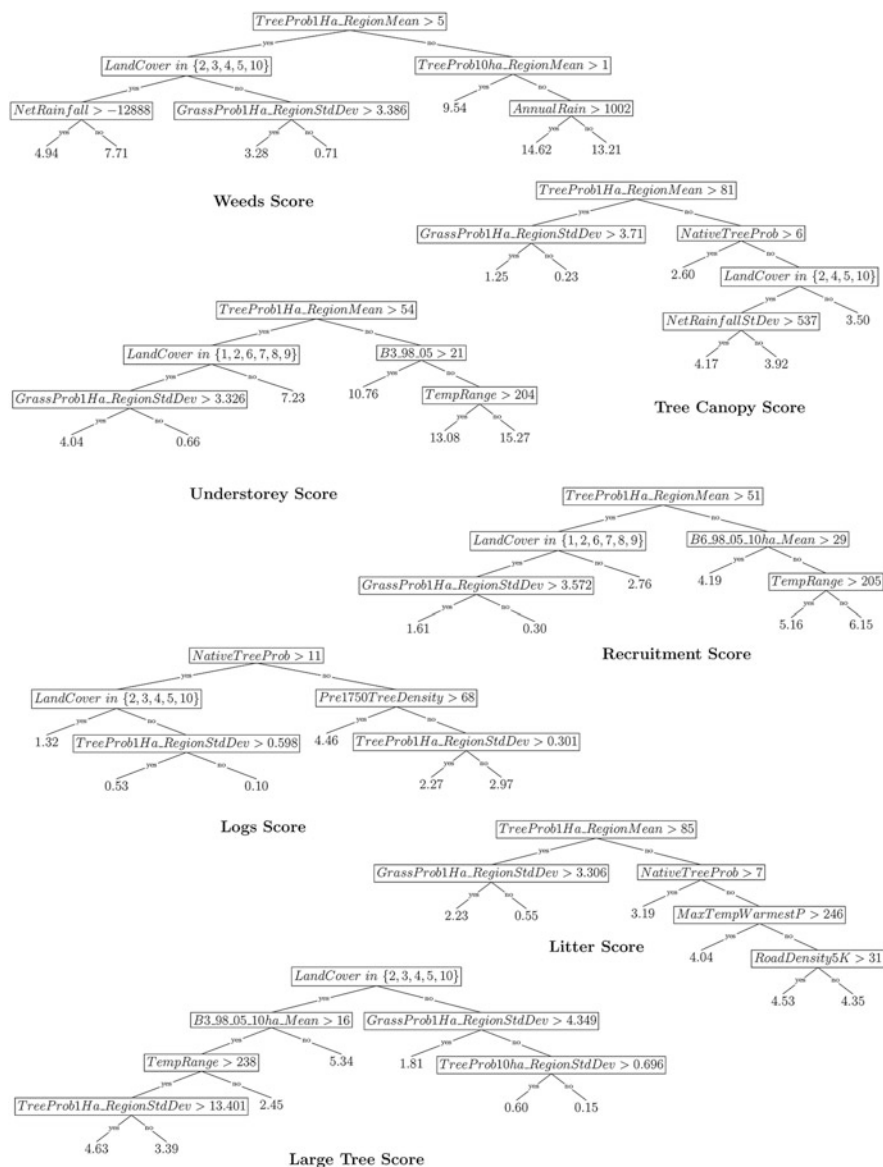
	Minimum	Maximum	Mean value	Standard deviation		Minimum	Maximum	Mean value	Standard deviation
Pre750TreeDensity	0	88	51.12	21.26	B7_98_05	5	148	37.5	20.71
TW1x1000	5335	12000	8126.69	1123.63	Ndvi_89_05	-0.191	0.771	0.4	0.18
ThorPot07	122	972	391.77	75.98	Ndwi_89_15_Mean	-0.625	0.449	-0.08	0.21
ThorInvPot07	169	956	432.09	68.51	Ndwi_89_05_StdError	0.006	0.411	0.09	0.056
RoadDensity5K	0	1380	82.21	150.56	B3_98_05_10ha_Mean	9	131	31.35	14.7
AnnualRain	254	1862	728.81	310.26	B4_98_05_10ha_Mean	13	140	66.15	11.93
NetRainfall	-19,863	-6149	-12,208.23	2321.09	B5_98_05_10ha_Mean	9	202	87.36	35.68
NetRainfallStdev	448	928	633.72	112.89	B6_98_05_10ha_Mean	6	118	39	19.79
MaxTempWarmestP	182	323	265.16	31.29	Nvdi_98_05_10haMean	-0.206	0.764	0.38	0.18
TempRange	149	295	234.49	35.01	B3_98_05_10ha_StdDev	0	34	3.85	3.15
Rad_Direct	578,125	1,296,306	1,017,713.28	51,135.83	B4_98_05_10ha_StdDev	1	33	5.44	3.55
NativeTreeProb	0	100	45.77	47.07	B5_98_05_10ha_StdDev	1	61	10.64	7.21
TreeProb1Ha_RegionMean	0	100	47.54	44.18	B6_98_05_10ha_StdDev	0	40	5.39	3.93
TreeProb10ha_RegionMean	0	100	51.99	41.36	LandCover	(1) Dryland agriculture (including cereal cropping and pasture); (2) Dense forest cover; (3) Woodlands and open forests; (4) Open woodlands and mallee shrublands; (5) Temperate grasslands and chenopod shrublands; (6) Urban/suburban; (7) Urban/industrial and commercial; (8) Irrigated crops, pasture and horticulture; (9) Plantation forestry; (10) Waterbodies and wetlands			
TreeProb1Ha_RegionStdDev	0	49.953	9.32	14.2					
TreeProb10ha_RegionStdDev	0	49.476	15.89	16.78					
NativeGrassProb	1	94	18.59	21.16					
GrassProb1Ha_RegionMean	1	93	18.63	19.66					
GrassProb1Ha_RegionStdDev	0	37.295	5.11	5.7					
GrassProb10ha_RegionMean	1	93	18.56	17.89					
GrassProb10ha_RegionStdDev	0	36.257	7.72	7.08					
BL_98_05	15	198	39.27	15.18					
B2_98_05	12	133	28.54	10.19					
B3_98_05	8	152	30.33	15.6					
B4_98_05	11	161	65.45	13.39					
B5_98_05	8	233	84.30	37.21					

were obtained by applying a  $k$ -means clustering procedure to a stack of median values for all Landsat 7 TM spectral bands and the Normalised Difference Vegetation Index across the years spanning 1989–2005. The 50 classes that emerged from the unsupervised classification were ‘lumped’ into 10 bins that were partially informed by a landuse model similarly derived using an ANN process. This procedure allowed for temporal states consequent of clearing, wildfire and forest harvesting to remain evident within broad landuse classes.

Using this dataset, we learned regression trees that predict the 7 HHSS vegetation condition scores separately. The learned regression trees are illustrated in Fig. 9.16. We used one pruning algorithms to obtain these trees, setting the minimum number of examples in a leaf to 2048.

Here, we discuss the major variables influencing the regression trees for the 7 HHSS scores. To begin with, we follow the positive or far left-hand side of the tree predicting *Weed Score*. It initially partitions the data on the basis of *TreeProb1HaRegionMean*: mean probability of detecting no tree cover within a 1 ha area around the subject pixel. This variable effectively divides the landscape into forests and treeless areas or areas with only scattered trees. Following the positive or left-hand side of the tree the data is further partitioned by the land cover classes. Classes 2, 3, 4, 5, and 10 represent natural or semi-natural areas and we should expect these areas to have a higher weed score (a high positive score reflects the absence of weeds rather than weed infestation) relative to other thinly treed areas. This is borne out by the regression tree. The final node is controlled by *NetRainfall*. *NetRainfall* is a variable that is derived from both mean monthly rainfall and mean evaporation rates. In essence it reflects the amount of effective rainfall (rainfall less evaporation) over an entire year. Once we have reached this node the model predicts that the drier and hotter a place is, the higher the weed score (provided we have satisfied earlier criteria). This reflects the current on-ground ecological reality in south-eastern Australia where there have been few deliberate introductions of exotic plant species into specialist habitat types, such as semi-arid regions, in comparison with temperate and sub-humid climatic regions that have been favoured by human settlement and intensive agriculture. Furthermore, it is apparent that *Recruitment score* and *Understorey score* are positively related. The regression trees of these scores are structurally identical and both employ very similar explanatory variables. Again, this is consistent with both field observation and ecological theory: a diverse and structurally intact understorey implies an adequate level of shrub and tree regeneration. The reverse is also likely. Within defined ecosystem types and states, a positive relationship between ecosystem function and structure is generally accepted by ecologists (Cortina et al. 2006).

Overall, the most important variables influencing all components of the HHSS are those immediately related to (the probability) of (indigenous and non-native) tree cover (such as *NativeTreeProb* that appears in the root of the multi-target tree, and *TreeProb1HaRegionMean*, which appears in the roots of 5/7 single-target trees). It is interesting to note that this is also the case for the sub-components that do not depend directly on the presence of tree cover, e.g. *Weeds Score*.



**Fig. 9.16** Regression trees for each Habitat Hectares site score; the sum of these attributes comprises the overall Habitat Hectares site score

Following closely is *LandCover* (as modelled from satellite images), with dense forest cover (category 2) yielding high HHSS scores. Finally, climate plays an important role, with variables describing temperatures, rainfall and their variability appearing in most of the models.

## 9.4 Conclusions

The two inferential modelling techniques—evolutionary algorithms and regression trees—are designed to extract and synthesise information from complex data patterns of real-world ecological data that improves our understanding of retro- and prospective ecosystem dynamics. Inferential models are typically represented as IF-THEN-ELSE rules that are fully transparent and can easily be updated with newly emerging data. They are suitable for short-term forecasting applications (see Chap. 15) and the identification of ecological thresholds. By contrast, mechanistic or process-based models are represented by rigid algebraic equations based on Michaelis-Menten-type kinetics and empirical relations, and are suitable for long-term forecasting applications (see also Chaps. 10 and 16).

Inferential models developed by evolutionary algorithms such as HEA suit as tools for determining thresholds and key driving variables of fast population growth and up to 30 days forecasting of population dynamics as demonstrated by the case study of Lake Müggelsee and discussed in Chap. 15. They also enhance meta-analysis of time-series data by quantifying phenological indicators as shown by the case study of lakes Wivenhoe and Paranoa. HEA is part of the NETLAKE toolbox (Recknagel and Ostrovsky 2016) recommended for the analysis of high-frequency data from lakes.

Inferential models developed by regression trees are flexible and fully transparent tools for revealing correlations between habitat properties and ecological entities. The tree induction process does not require prior assumptions, is fast and is not influenced by redundant variables and noise.

**Acknowledgements** We thank SeqWater, Brisbane (Australia) for making data available from Lake Wivenhoe, CAESB Brasilia (Brazil) for data from Lake Paranoa, and the Leibniz-Institute of Freshwater Ecology and Inland Fisheries, Berlin (Germany) for data from Lake Müggelsee.

## References

- Bennett ND, Croke BFW et al (2013) Characterising performance of environmental models. *Environ Model Softw* 40:1–20
- Breiman L, Friedman J, Stone CJ et al (1984) Classification and regression trees. Chapman & Hall, London
- Cao H, Recknagel F, Welk A et al (2006) Hybrid evolutionary algorithm for rule set discovery in time-series data to forecast and explain algal population dynamics in two lakes different in morphometry and eutrophication. In: Recknagel F (ed) *Ecological informatics*, 2nd edn. Springer, Berlin, pp 347–368
- Cao H, Recknagel F, Orr PT (2014) Parameter optimisation algorithms for evolving rule models applied to freshwater ecosystem. *IEEE Trans Evol Comput* 18:793–806
- Cao H, Recknagel F, Bartkow M (2016) Spatially-explicit forecasting of cyanobacteria assemblages in freshwater lakes by multi-objective hybrid evolutionary algorithms. *Ecol Model* 342:97–112



- Cortina J, Maestre FT, Vallejo R et al (2006) Ecosystem structure, function, and restoration success: are they related? *J Nat Conserv* 14:152–160
- Gibbons P, Briggs SV, Ayers D et al (2009) An operational method to assess impacts of land clearing on terrestrial biodiversity. *Ecol Indic* 9:26–40
- Holland JH (1975) *Adaptation in natural and artificial systems*. University of Michigan Press, Ann Arbor, MI
- Holland JH, Holyoak KJ, Nisbett RE et al (1986) *Induction. Process of inference, learning, and discovery*. MIT Press, Cambridge
- Kocev D, Naumoski A, Mitreski K et al (2010) Learning habitat models for the diatom community in Lake Prespa. *Ecol Model* 221:330–337
- Koza JR (1992) *Genetic programming: on the programming of computers by means of natural selection*. MIT Press, Cambridge, MA
- Krstić S (2005) Description of sampling sites. Report on baseline data for water (surface and groundwater) including waste related data for the target region – EC-FP6 project TRABOREMA, EC-Project Contract No. INCO-CT-2004-509177, Deliverable 2.2
- Loh W-Y (2011) Classification and regression trees. *WIREs Data Mining Knowl Disc* 16:14–23
- McGarigal K, Cushman SA, Neel MC (2002) FRAGSTATS: spatial pattern analysis program for categorical maps. University of Massachusetts, Amherst (Computer program produced by the authors at the University of Massachusetts, Amherst)
- Osenberg CW, Sarnelle O, Cooper SD et al (1999) Resolving ecological questions through meta-analysis: goals, metrics, and models. *Ecology* 80:1105–1117
- Parkes D, Newell G, Cheal D (2003) Assessing the quality of native vegetation: the habitat hectares approach. *Ecol Manag Restor* 4:29–38
- Recknagel F, Ostrovsky I (2016) Inferential modelling of time series by evolutionary computation. In: Obrador B, Jones ID, Jennings E (eds) NETLAKE toolbox for the analysis of high-frequency data from lakes (Factsheet 11). Technical report. NETLAKE COST Action ES1201. pp 57–60. <http://eprints.dkit.ie/id/eprint/542>
- Recknagel F, Ostrovsky I, Cao H et al (2013) Ecological relationships, thresholds and time-lags determining phytoplankton community dynamics of Lake Kinneret, Israel elucidated by evolutionary computation and wavelets. *Ecol Model* 255:70–86
- Recknagel F, Ostrovsky I, Cao H et al (2014) Hybrid evolutionary computation quantifies environmental thresholds for recurrent outbreaks of population density. *Ecol Inform* 24:85–89
- Recknagel F, Adrian R, Köhler J et al (2016) Threshold quantification and short-term forecasting of *Anabaena*, *Aphanizomenon* and *Microcystis* in the polymictic eutrophic Lake Müggelsee (Germany) by inferential modelling using the hybrid evolutionary algorithm HEA. *Hydrobiologia* 778:61–74
- Redfield AC (1958) Biological control of chemical factors in the environment. *Am Sci* 46: 205–221
- Reynolds CS (1984) *The ecology of freshwater phytoplankton*. Cambridge University Press, Cambridge
- Serner RW (2008) On the phosphorus limitation paradigm for lakes. *Int Rev Hydrobiol* 93: 433–445
- Storn R, Price K (1997) Differential evolution – a simple and efficient heuristic for global optimization over continuous spaces. *J Glob Optim* 11:341–359
- Van Mooy BAS, Rocap G, Fredericks HF, Evans CT, Devol AH (2006) Sulfolipids dramatically decrease phosphorus demand by picocyanobacteria in oligotrophic marine environments. *Proc Natl Acad Sci USA* 103:8607–8612

# Chapter 10

## Process-Based Modeling of Nutrient Cycles and Food-Web Dynamics

George Arhonditsis, Friedrich Recknagel, and Klaus Joehnk

**Abstract** Mathematical models are indispensable for addressing pressing aquatic ecosystem management issues, such as understanding the oceanic response to climate change, the interplay between plankton dynamics and atmospheric CO<sub>2</sub> levels, and alternative management plans for eutrophication control. The appeal of process-based (mechanistic) models mainly stems from their ability to synthesize among different types of information reflecting our best understanding of the ecosystem functioning, to identify the key individual relationships and feedback loops from a complex array of intertwined ecological processes, and to probe ecosystem behavior using a range of model application domains. Significant progress in developing and applying mechanistic aquatic biogeochemical models has been made during the last three decades. Many of these ecological models have been coupled with hydrodynamic models and include detailed biogeochemical/biological processes that enable comprehensive assessment of system behavior under various conditions. In this chapter, case studies illustrate ecological models with different spatial configurations. Given that each segmentation depicts different trade-offs among model complexity, information gained, and predictive uncertainty, our objective is to draw parallels and ultimately identify the strengths and weaknesses of each strategy.

---

G. Arhonditsis (✉)  
University of Toronto Scarborough, Scarborough, ON, Canada  
e-mail: [georgea@utsc.utoronto.ca](mailto:georgea@utsc.utoronto.ca)

F. Recknagel  
University of Adelaide, Adelaide, SA, Australia  
e-mail: [friedrich.recknagel@adelaide.edu.au](mailto:friedrich.recknagel@adelaide.edu.au)

K. Joehnk  
CSIRO Land and Water, Canberra, ACT, Australia  
e-mail: [klaus.joehnk@csiro.au](mailto:klaus.joehnk@csiro.au)

## 10.1 Introduction

Mechanistic aquatic biogeochemical models have formed the scientific basis for environmental management decisions by providing a predictive link between management actions and ecosystem responses. An appealing feature for their extensive use is their role as “information integrators” in that they can be used to synthesize across different types of information that reflect our best understanding of ecosystem functioning (Spear 1997). Their main foundation consists of causal mechanisms, complex interrelationships, and direct and indirect ecological paths that are mathematically depicted in the form of nonlinear differential equations. Model endpoints (*state variables*) usually coincide with routinely monitored environmental variables that, in turn, are considered reliable surrogates of the physics, chemistry and biology of the aquatic ecosystem under investigation. Scientific knowledge, expert judgment, and experimental/field data are used to assign realistic values to model inputs. Such inputs can either be ecologically meaningful *parameters*, representing physical or chemical processes, physiological rates, and partition coefficients, or factors that externally influence the biotic and abiotic components of the system, also known as *forcing functions*. The latter model input could be essential in linking an externally-introduced pollutant (e.g., herbicide application or nutrient loading rates) with a key ecosystem attribute (e.g., biodiversity, total phytoplankton or cyanobacteria levels) or may not be directly subject to anthropogenic control, e.g., temperature and solar radiation.

In the context of environmental decision-making and management, the state variables of a model typically represent components or attributes of the system that we consider to be relevant to the research question being examined. For example, in lake eutrophication problems, the state variables can be the various forms of phosphorus (phosphate, dissolved and particulate organic phosphorus) and the different phytoplankton (diatoms, green algae, cyanobacteria) or zooplankton (copepods, cladocerans) groups. If we are interested in predicting the success of a strategy to reduce fish contamination levels in the Great Lakes, then logical state variables of the model will be the contaminant concentrations in the tissues of several fish species along with their corresponding biomass levels. The state variables can be expressed in units of mass or concentration (*biogeochemical models*), energy (*bioenergetic models*), and number of species or individuals per unit of volume or area (*biodemographic models*). Consequently, the physical, chemical, and biological processes considered by the model account for the transfer of mass, energy and/or individuals and drive the variability of the state variables. Examples of physical processes are diffusive and advective transport of a fluid such as air or water; chemical reactions typically modeled are hydrolysis, photolysis, oxidation, and reduction; biological processes that are essential in modulating the dynamics of biotic components are growth, metabolism, mortality, excretion, predation, and emigration or immigration.

An interesting feature of mathematical models is the existence of *feedback loops* (defined as closed-loop circles of cause and effect in ecological conditions in one

part of the system that shape processes elsewhere in the system) that amplify (*positive feedbacks*) or counteract (*negative feedbacks*) the original change. A characteristic example is the positive feedback of bacteria-mediated mineralization of the excreta of zooplankton basal metabolism that replenishes the summer epilimnetic phosphate pool, which stimulates phytoplankton growth and offsets the herbivorous control of autotroph biomass. Then, this increase of the phytoplankton biomass reinforces the zooplankton growth, thereby preventing an undesirable collapse at the second trophic level. An example of a negative feedback in the system is when excessive phosphorus is added causing excessive phytoplankton growth which, in turn, causes shading that reduces sunlight penetration to lower water depths and therefore the reduced primary production along with the decomposition of the sinking phytoplankton cells result in gradual oxygen depletion and possibly hypolimnetic anoxia which could kill off benthic organisms. Thus, the ability of mathematical models to consider a series of intermingled ecological mechanisms allows reproducing non-linear response patterns induced by distant (and presumably unrelated) causal factors.

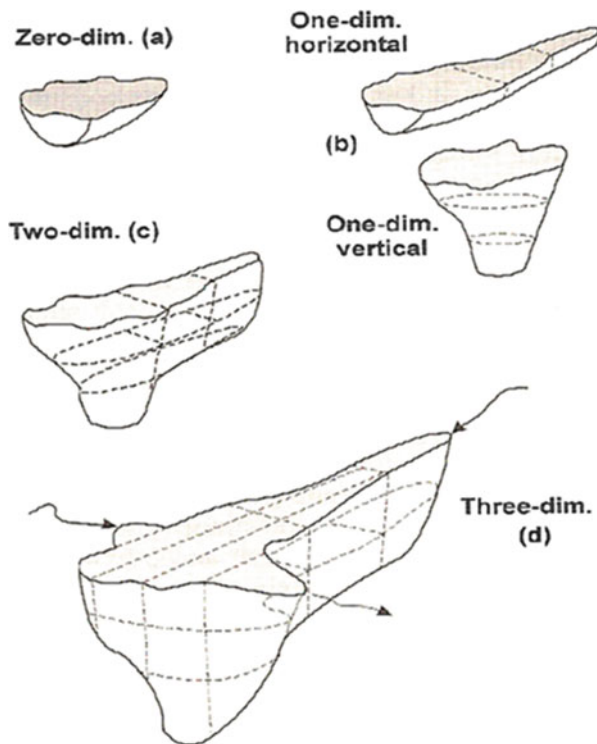
Fundamental to mathematical models is the *Lomonosov-Lavoisier Law of Conservation of Mass*. In quantitative terms, the principle is translated into a *mass balance model* that is built from *mass balance equations* that account for all the inputs and outputs of mass across the system's boundaries and all the transport and transformation processes occurring within the system. For a finite period of time, this concept can be mathematically expressed as:

$$\begin{aligned} \text{Accumulation} &= \Delta \text{Mass} / \Delta t = \Delta(\text{Volume} \cdot \text{Concentration}) / \Delta t \\ &= \text{Input} \pm \text{Reactions} - \text{Output} \end{aligned}$$

This differential equation is solved to get model output (concentration) as a function of time. The solution provides us with a *time dependent, dynamic or unsteady-state model* (these terms are used interchangeably). If the equation is relatively simple, then we can mathematically solve the equation to get an *explicit analytical solution* where infinitesimal time steps are implied. However, often the equation(s) are difficult to solve, in which case we use *numerical approximations*. The simplest of these numerical methods is a *finite difference* approach whereby the computer algorithm steps through time according to defined and discrete time steps. An alternative solution to explicitly solving a differential equation is to assume no change in the state variable value with time or that the system is at a *steady state*, i.e.,  $\Delta \text{Mass} / \Delta t = 0$ .

Mathematical models must also consider variations in space (e.g., geographic variation). If we do not consider spatial variation, then we have a *lumped* model, e.g.,  $\Delta \text{Mass} / \Delta x = 0$  where  $x$  is distance. An example of a lumped model would be treating a lake's water column as a single well-mixed compartment, i.e., the contents are sufficiently well mixed as to be uniformly distributed (Fig. 10.1a). Such characterization is often used to model shallow and small lakes, where stratification does not occur and spatial homogeneity can be assumed. A common example of this type of models is the *continuously stirred tank reactor* (CSTR) that

**Fig. 10.1** Zero-, one-, two-, and three-dimensional strategies for accommodating the spatial variability in lake systems



simulates the system as a single, well-mixed or homogeneous compartment, where its properties can only vary in time according to the following equation:

$$\frac{dC}{dt} = f(C, \theta, t)$$

where the quantity  $C$  (e.g. chemical concentration) being differentiated is called the *dependent variable*; the quantity  $t$  (time in zero-dimensional systems) with respect to which  $C$  is differentiated is called the *independent variable*, and  $\theta$  corresponds to the various inputs of the equation (e.g. external forcing, parameters). When the function involves one independent variable the equation is called *ordinary differential equation* (or ODE). Alternatively, we may consider spatial variations, with the simplest formulation of translating geographic differences into discrete, well-mixed (homogeneous) boxes or compartments, typically defined according to physical properties of the studied system. This type of model is often called a *box model*. Using a lake as an example, a box model may have a warmer upper water layer or epilimnion and a cooler lower water layer or hypolimnion to treat thermal stratification (Fig. 10.1c). In this example, the *discretisation* is defined according to the temperature vertical profiles. The model

then includes heat and/or chemical transfer between the two compartments according to heat and mass transfer coefficients.

The most sophisticated treatment of spatial variation is to have an analytical solution to the differential  $dMass/dx$ . This would quantify the continuous variation in the output as a function of space or location and hence would offer a *continuous model*. Similarly to solutions of the time-varying differential equation, we can solve the equations using an explicit analytical solution or a numerical approximation. In this case, we deal with *partial differential equations* (or PDE) that involve two or more independent variables. For example, such equations can be useful for systems with a prevailing one-directional flow, e.g., rivers where the physical, chemical, and biological properties are determined by this flow, and thus we may opt for a one-dimensional representation that accommodates variability in the  $x$  axis (Fig. 10.1b). Namely, the advection-diffusion equation that combines the two main processes of mass transport, advection and diffusion, along with a first-order reaction, will be suitable to describe the spatiotemporal distribution of a substance in a river:

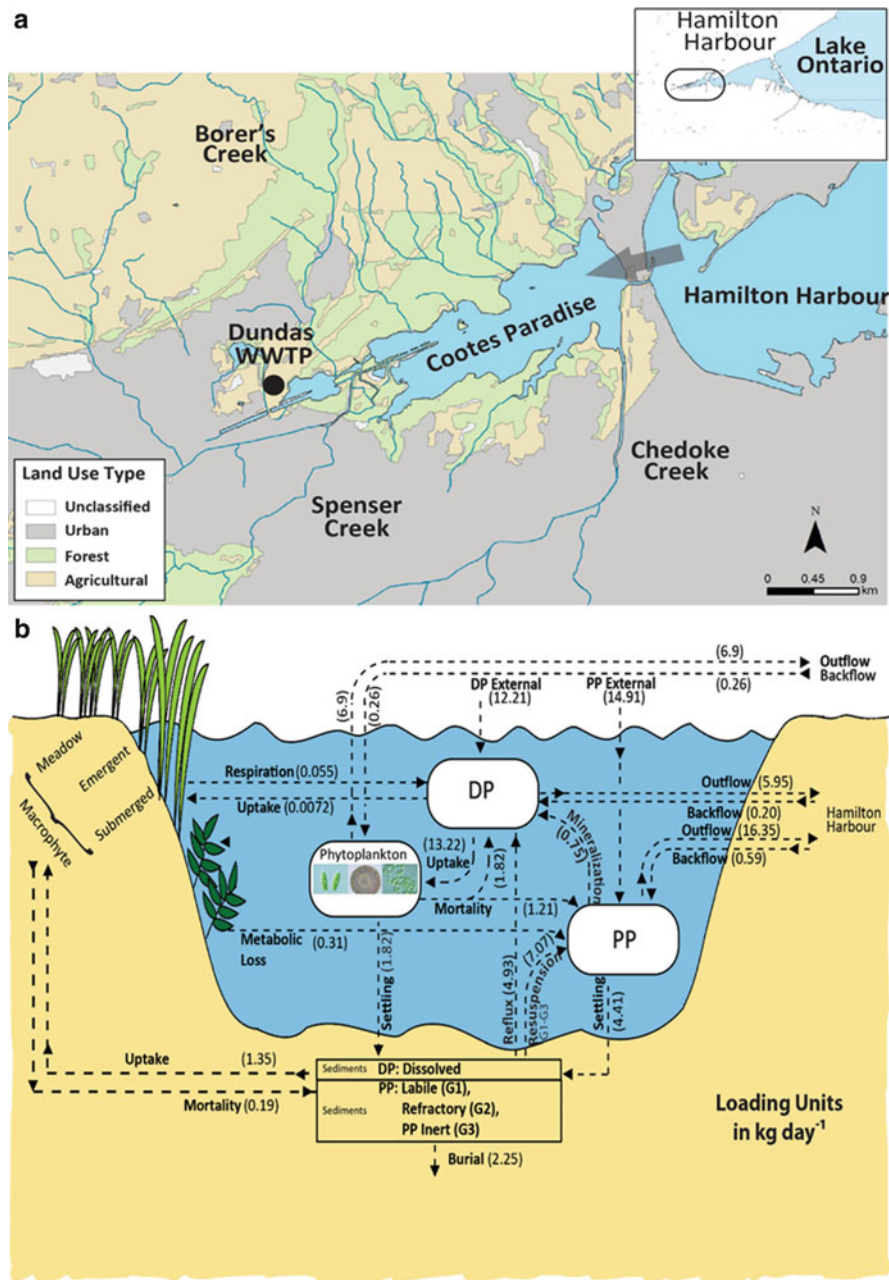
$$\frac{\partial C}{\partial t} = -\frac{\partial Cu}{\partial x} + D_x \frac{\partial^2 C}{\partial x^2} - kC$$

where  $C$  the chemical concentration in fixed element of space,  $x$  (distance) the direction of the flow,  $u$  (distance/time) is velocity for advective transport,  $D_x$  (distance<sup>2</sup>/time) is a diffusion coefficient for diffusive transport, and  $k$  (inverse time) is a rate constant for a first-order reaction. One-dimensional representations can also be used to simulate the vertical stratification of a deep lake without significant variability in the horizontal plane (Fig. 10.1b). Two or three-dimensional segmentations will be more appropriate for larger systems (estuaries, large lakes with complex morphology, fragmented landscapes) characterized by significant variability of their properties in both horizontal and vertical directions (Fig. 10.1c, d). In this chapter, we offer a series of case studies that illustrate ecological models with different spatial configurations. Given that each segmentation depicts different trade-offs among model complexity, information gained, and predictive uncertainty, our objective is to draw parallels and ultimately pinpoint the strengths and weaknesses of each strategy.

## 10.2 Zero- and One-Dimensional Lake Models

### 10.2.1 Zero-Dimensional Model for the Phosphorus Cycle in a Hypereutrophic Wetland

Cootes Paradise is a large marsh in western Lake Ontario that is hydraulically connected to Hamilton Harbour by a man-made channel (Fig. 10.2). It is characterized by hypereutrophic conditions, stemming from the agricultural and urban



**Fig. 10.2** Map of Cootes Paradise and land use classification of the surrounding watershed (a). Average daily phosphorus fluxes (kg day<sup>-1</sup>) corresponding to each simulated process during the growing season (May–October) (b)



development of the (previously forested) watershed along with the sewage effluent discharged into the marsh for over nine decades (Thomasen and Chow-Fraser 2012). The vegetation cover in Cootes Paradise had receded to less than 15% by the 1990s, relative to >90% cover with very high plant diversity at the turn of the twentieth century (Chow-Fraser 2005). Coinciding with the vegetation decline, the fishery shifted from a desirable warm water fishery of northern pike and largemouth bass to one dominated by planktivorous and benthivorous species, such as bullheads, invasive common carp, and alewife. In particular, common carp, an exotic species introduced into Lake Ontario at the end of the nineteenth century, accounted for up to 45% of the overall water turbidity (Lougheed et al. 2004). High turbidity had many detrimental effects across the entire food web, such as reducing light penetration to a level that was insufficient for submersed aquatic vegetation/periphyton growth, clogging filter-feeding structures of invertebrates, and affecting the behaviour and survival of visually hunting predators and mating fish (Thomasen and Chow-Fraser 2012). To ameliorate the prevailing adverse ecological conditions in the wetland, a number of restoration strategies have been implemented, such as carp exclusion, nutrient loading reduction, and macrophyte planting (Lougheed et al. 2004).

In this context, Kim et al. (2016) presented a modelling exercise aimed at understanding the primary drivers of eutrophication in Cootes Paradise by elucidating the interplay between various phosphorus-loading sources, internal flux rates, phytoplankton activity, and the potential of macrophytes to become an integral part of the bottom-up forcing into the system. Cootes Paradise marsh is approximately 4 km long, with a maximum width of 1 km, and a mean depth of 0.7 m. Because of its small size, the surface area and volume of the marsh can vary significantly according to water level fluctuations, reaching a maximum of 2.5 km<sup>2</sup> and 3.6 × 10<sup>6</sup> m<sup>3</sup>, respectively (Mayer et al. 2005). Thus, Kim et al. (2016) adopted a zero-dimensional approach representing the Cootes Paradise as a spatially homogeneous system with a hydraulic connection to Hamilton Harbour. The focal points of the model calibration were the reproduction of the water level variability and the realistic characterization of processes such as phosphorus release via reflux/diffusion and resuspension from the sediments. The role of macrophytes in the phosphorus cycle was accounted for by the dry-mass biomass submodel presented by Asaeda et al. (2000), and modified by Kim et al. (2013), by differentiating among three macrophyte functional groups: emergent, meadow, and submerged. Each equation considers macrophyte growth through uptake of dissolved inorganic phosphorus from the interstitial water, respiration releasing phosphorus back to the water column, and mortality depositing phosphorus to the sediment pool.

After the model calibration against a 17-year (1996–2012) time-series of water quality data, the Cootes Paradise model provided internal loading estimates ( $\approx 12.01 \text{ kg day}^{-1}$ ) that were substantially lower than sediment reflux rates reported in previous modelling work from the 1990s (Prescott and Tsanis 1997). This discrepancy was attributed to the sediment resuspension induced by carp bioturbation, which ceased after the construction of a barrier (or fishway) at the outlet of Cootes Paradise. The fishway became operational during the winter of 1997 and



used 5-cm wide grating to physically exclude large fish, targeting carp, from the marsh (Lougheed et al. 2004). This biomanipulation practice effectively prevented large carp (>40 cm) from entering the marsh after February 1997. According to the Cootes Paradise marsh model projections (Fig. 10.2), the phosphorus contribution of internal sources (reflux, resuspension, macrophyte respiration) and sinks (sedimentation) appears to be significantly lower relative to the external sources (exogenous inflows) and sinks (outflows to Hamilton Harbour). Release of phosphorus from actively growing submerged and emergent macrophytes is typically considered minimal, whereas decaying macrophytes may act as an internal phosphorus source adding considerable quantities of phosphorus into the water (Granéli and Solander 1988; Asaeda et al. 2000). Nonetheless, the Cootes Paradise model demonstrated that macrophytes play a minimal role in the phosphorus budget of the marsh, reflecting the fact that their abundance (e.g., biomass and density) is fairly low in its current state.

Kim et al. (2016) identified the water level fluctuations as another critical factor that can profoundly modulate the interplay among physical, chemical, and biological components of the Cootes Paradise ecosystem. Lower water levels (and thus smaller water volumes) imply lower dilution and higher nutrient concentrations; a pattern consistent with Kim et al.'s (2016) predictions of higher ambient *TP* values towards the end of the summer-early fall, when the lower water levels in the marsh occur. Further, with lower water levels, wind energy is more easily transmitted to the bottom sediments that, in turn, would accentuate the release of phosphorus due to stirring and mixing (Prescott and Tsanis 1997; Chow-Fraser 2005). The same mechanisms also appear to be the main drivers of the spatiotemporal variability of water turbidity, thereby influencing the illumination of the water column; especially, the light environment near the sediment surface in open-water sites, which currently does not favour submerged macrophyte growth (Chow-Fraser 2005). In the same context, two threshold water levels have been proposed for evaluating the resilience of submerged macrophytes; a maximum threshold, above which light availability becomes limiting, and a minimum threshold, below which conditions are excessively dry (Harwell and Havens 2003). On a final note, the simplified segmentation of the Cootes Paradise model did not allow researchers to reproduce the water quality gradients occasionally established between western and eastern ends of the marsh. This weakness was highlighted by Kim et al. (2016) as a key missing point that may not allow delineation of the role of a wastewater treatment plant located in the innermost section of Cootes Paradise.

### ***10.2.2 One-Dimensional Model for Nutrient Cycles and Plankton Dynamics in Lakes and Reservoirs***

The model SALMO (Benndorf and Recknagel 1982; Recknagel and Benndorf 1982) is a process-based one-dimensional lake model that simulates concentrations of the state variables  $\text{PO}_4\text{-P}$ ,  $\text{NO}_3\text{-N}$ , DO, detritus, chlorophyta, bacillariophyta,

cyanophyta and cladocera (Fig. 10.3) at daily time steps for the mixed total water body, and epi- and hypolimnion during thermal stratification (Fig. 10.4).

The mass balances for the nutrients and detritus are determined by transport processes such as import, export, sedimentation and exchange between epi- and hypolimnion, as well as consumption by phytoplankton, microbial recycling of detritus and resuspension from anaerobic sediments.

Detritus is also subject to grazing by zooplankton. The mass balances for the phytoplankton phyla chlorophyta, bacillariophyta and cyanophyta, and for cladocera include transport processes by sedimentation, import and export, but are predominantly determined by photosynthesis, respiration and grazing, as well as assimilation, respiration and mortality, respectively. The zooplankton mortality includes predation by planktivorous fish represented by parameters reflecting an annually constant stock size. The DO budget is determined by O<sub>2</sub> solubility (Henry's Law), plankton photosynthesis, and respiration. SALMO requires daily input data for volumes, mean and maximum depths of mixed and stratified water bodies, loadings of PO<sub>4</sub>-P, NO<sub>3</sub>-N, and detritus by the inflowing water, incident solar radiation and water temperature.

Model inputs (see Fig. 10.3) characterise lake specific nutrient loadings, climate conditions, seasonal circulation types, and morphometry by routinely measured variables. Model parameters reflect lake specific underwater light transmission, temperature, light and nutrient limitations of phyla-specific phytoplankton growth, temperature and food limitation of cladocera growth, stock size of planktivorous fish, specific sinking velocities and grazing preferences for phytoplankton phyla and detritus. The model can therefore easily be implemented and validated for different lakes and drinking water reservoirs based on routine limnological measurements, and can serve as a flexible tool for scenario analysis of eutrophication management options such as control of external and internal nutrient loadings (e.g. Recknagel et al. 1995; Chen et al. 2014) and climate change, artificial destratification and aeration, food web manipulation by carnivorous fish, hypolimnetic withdrawal and partial drawdown (see Chap. 16).

Figures 10.5 and 10.6 display examples for applications of SALMO to a variety of lakes and reservoirs with different circulation types and trophic states. The Saldenbach Reservoir (Germany) had been a key water body during the development and testing phase of SALMO (e.g. Recknagel and Benndorf 1982). As illustrated in Fig. 10.5, the one-dimensional mode of SALMO simulated concentrations of PO<sub>4</sub>-P, phyto- and zooplankton separately for epi- and hypolimnion for summer months while the zero-dimensional mode simulated the remaining months of the year 1975. In the case of Lake Taihu (China), the zero-dimensional mode of SALMO simulated successfully extreme conditions of that shallow hypertrophic water body as prerequisite for a scenario analysis on management options (Chen et al. 2014). Again different conditions had to be matched by SALMO when applied to the two warm-monomictic Millbrook and Mt Bold Reservoirs in Australia (Fig. 10.6). The Millbrook Reservoir is equipped with an aerator that artificially destratifies the water body during summer being simulated by the zero-dimensional SALMO. The Millbrook Reservoir will be revisited in Chap. 16 when the model

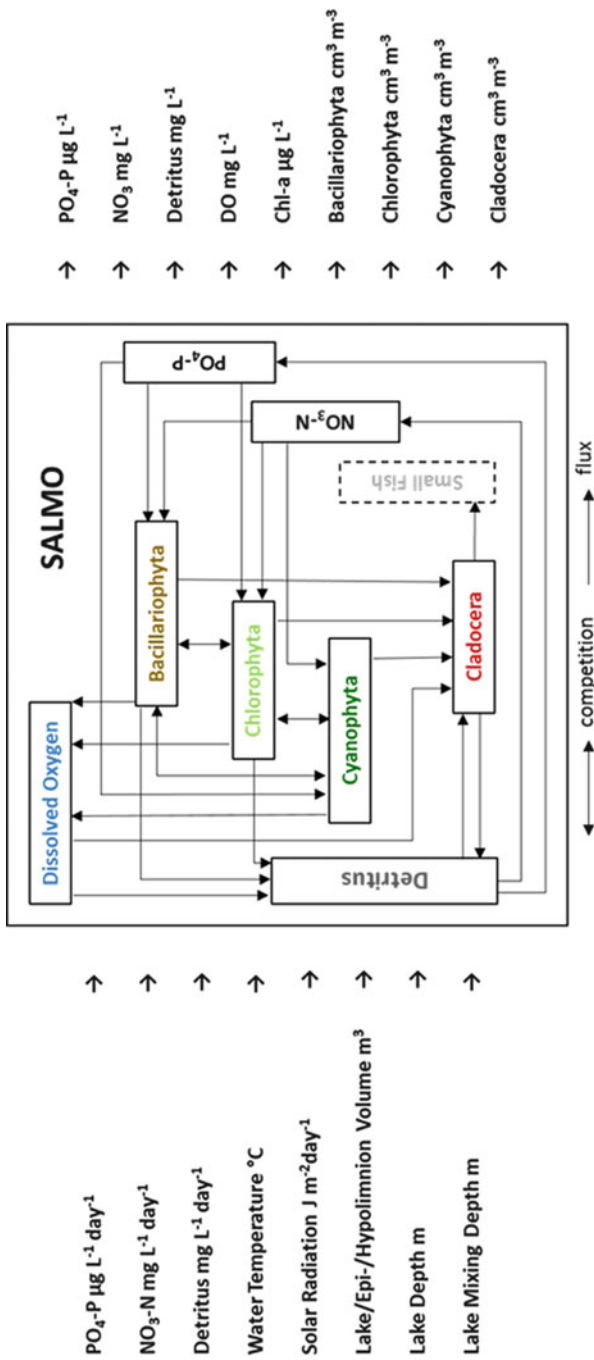


Fig. 10.3 Basic drivers, state variables and process pathways of the lake model SALMO

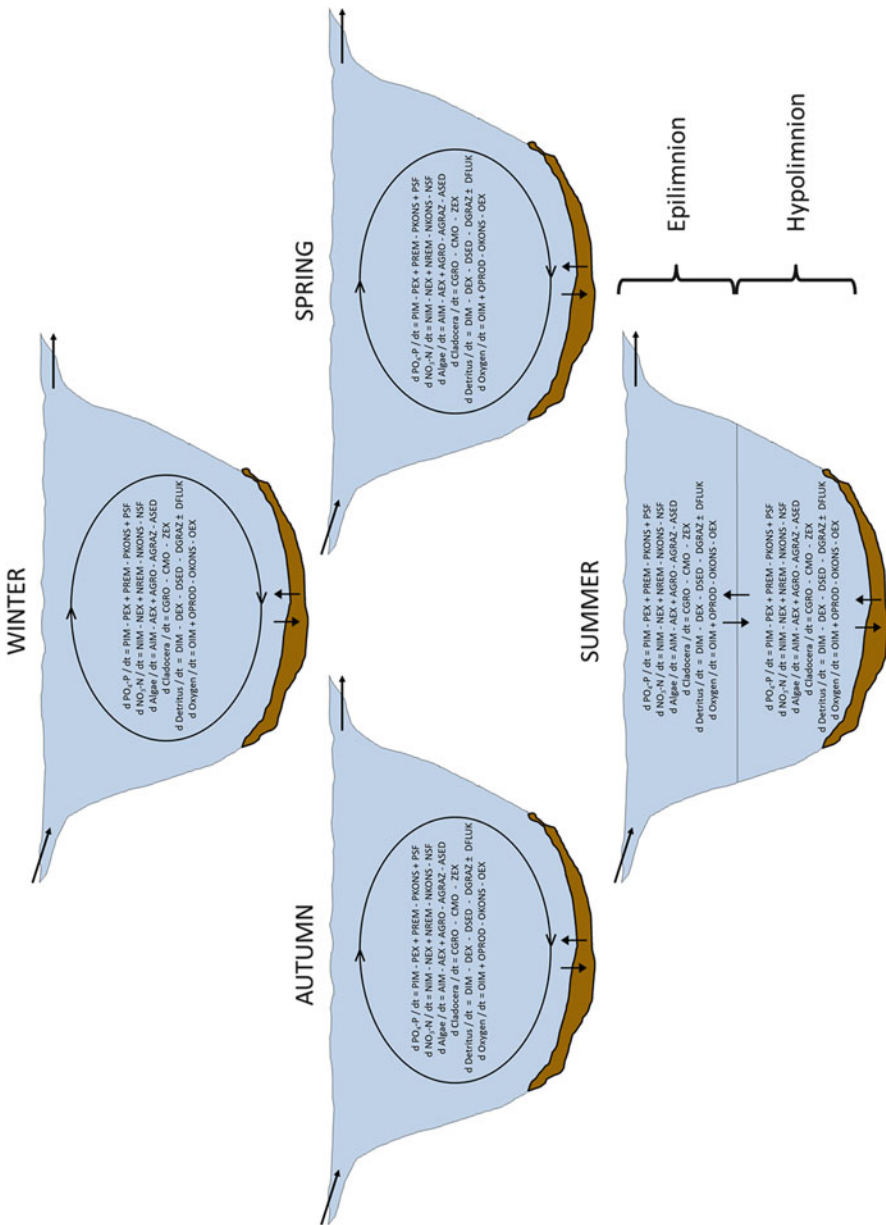
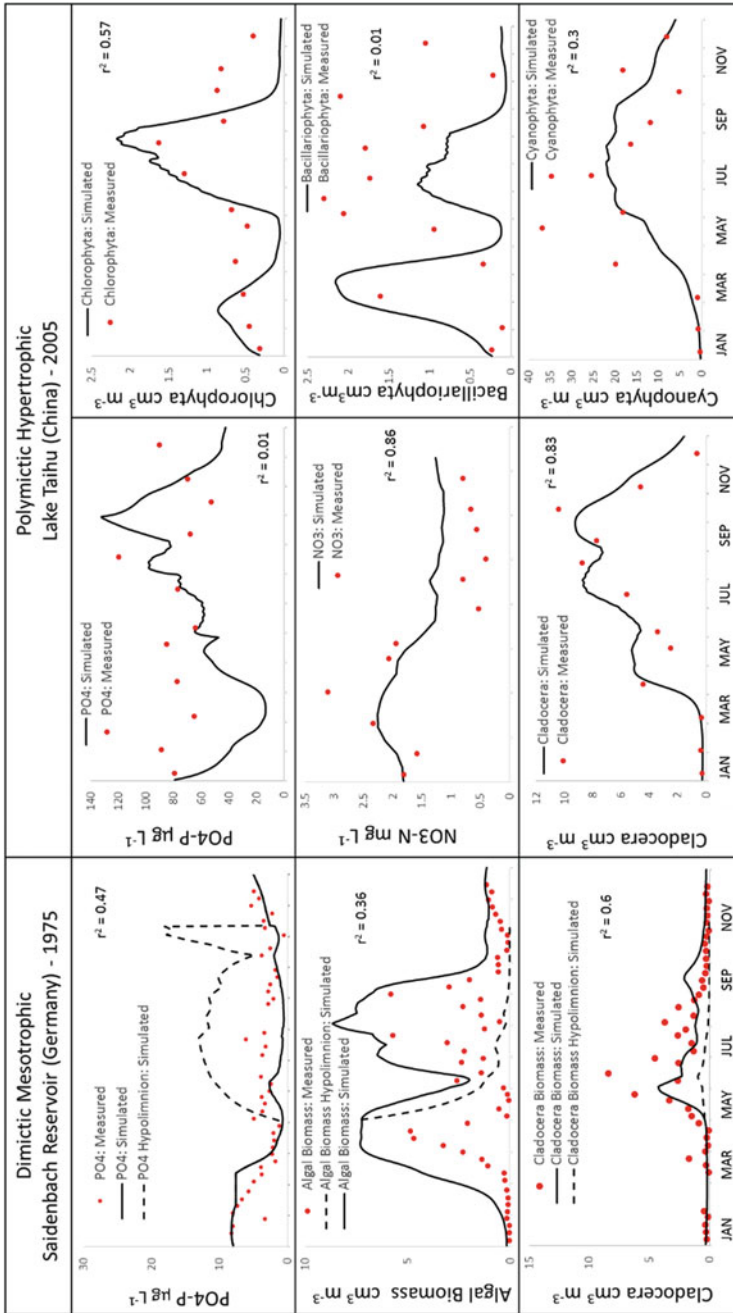
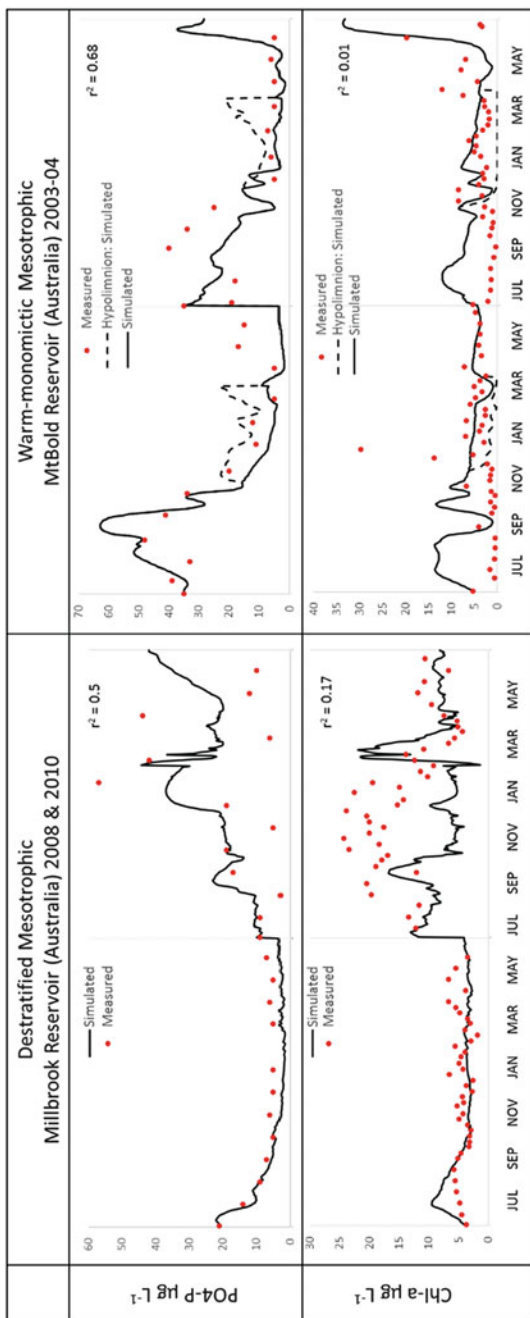


Fig. 10.4 Conceptual scheme for SALMO simulations of warm-monomictic lakes by seasonally alternating between a zero- and a one-dimensional model



**Fig. 10.5** Validation of SALMO for the dimictic mesotrophic Saldenbach Reservoir (*left column*) and the polymictic hypertrophic Lake Taihu (*middle and right columns*)



**Fig. 10.6** Validation of SALMO for the de-stratified mesotrophic Millbrook Reservoir from 07/2008 to 06/2009 and 07/2010 to 06/2011 (*left column*) and the warm-monomictic mesotrophic Mt Bold Reservoir from 07/2003 to 06/2005 (*right column*)

ensemble SWAT-SALMO is applied to the Millbrook catchment-reservoir system. The Mt Bold Reservoir is thermally stratified during summer.

The performance of SALMO for different water bodies depends firstly on having accurate data for the key driving variables reflecting depths and volume fluctuations, nutrient loadings, light and temperature dynamics, secondly on having accurate measurements of key state variables for validation, and thirdly on correct calibration of phyto- and zooplankton related growth parameters reflecting nutrient, food, light and temperature limitations. The calibration of SALMO for specific water bodies focuses on key parameters determined by sensitivity analysis (e.g. Recknagel 1984) and their multi-objective optimization by evolutionary algorithms (Cao et al. 2008; Cao and Recknagel 2009; Chen et al. 2014) within the range of their standard error and against measured state variables.

### 10.3 Multi-dimensional Lake Models

The transition from 0-dimensional to higher dimensional models makes it necessary to include processes related to internal lake dynamics as well as more complex boundary conditions. Biogeochemical processes essentially stay the same, but we have to add different transport mechanisms driven by external and internal forces. While in the previous chapters, we focussed on the biogeochemical processes, here we examine physical processes changing the natural environment of a lake.

In a natural lake environment processes will be affected by driving forces at the surface—and to a lesser extend at the bottom—which may vary in time and space generating complex flow patterns in a lake. Driving forces such as weather or inflow will change the balance of heat and momentum in a lake, which can be described as an analogue to the previously mentioned mass balance. Physically speaking, to solve for processes in a real lake environment, we need to solve additional differential equations for velocity (momentum balance) at each point in the domain as well as for temperature (heat balance) or salinity (constituent mass balance), see e.g., Hutter and Jöhnk (2004). The resulting system of equations, Navier-Stokes equation and heat balance using Fourier's Law of heat conduction, is usually too complex to be solved directly and has to be adapted for specific situations. Knowing transport properties, advective and diffusive, throughout the lake then allows for the simulation of spread and distribution of constituents like nutrients, particulate matter or algae.

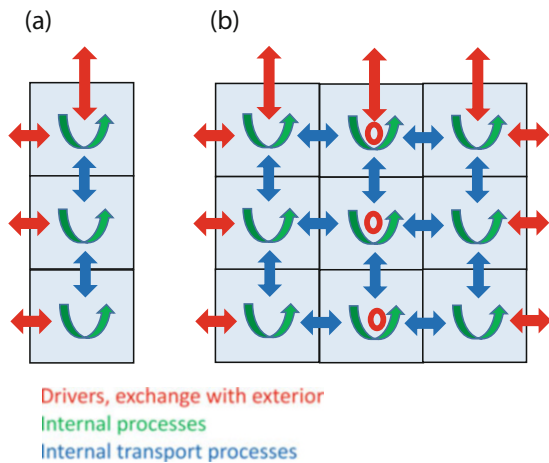
One simplification step is the separation of large from small-scale processes. This then leads to simplified equations describing the general flow or transport in a lake and a parameterization of small-scale process describing turbulence, i.e. diffusional transport. The latter can either be a set of extra differential equation like the k- $\epsilon$  turbulence model (e.g. Joehnke and Umlauf 2001), or a simplified version describing diffusional processes via a functional dependence on the vertical density gradient in a lake (or in physical terms better expressed as the buoyancy frequency). Another simplification, which can be well observed in oceans and

lakes, comes from the fact that these systems are to a large extent laterally homogeneous, i.e., in a lake changes mainly happen in the vertical as long as one is far away from a boundary. However, the smaller the system is, the more influence from the boundary needs to be taken into account, and the larger a lake or the more complex its bathymetry is the more likely three-dimensional flows will form.

In the previous examples of 1D lake models, it was assumed that lake stratification could be described by a two-layered system. This is a valid assumption, as long as the dynamic changes during build-up or breakdown of stratification are of no interest. Specific processes at the interface between these two layers are also neglected by such a simplifying assumption, e.g., a metalimnetic oxygen minimum (Joehnk and Umlauf 2001) where the system is determined by enhanced microbiological decomposition through accumulation in the metalimnion due to an increase in density.

In Fig. 10.7, different strategies for describing the spatial variability of a lake are depicted. In each compartment of such geometry, nutrient cycling and food web processes can be described individually. However, the more spatial complexity one needs to take into account, the more physical, transport processes have to be accounted for. In a zero dimensional system (Fig. 10.1a), e.g., a shallow well-mixed pond, no specific physical processes have to be looked at as long as the time scales of biogeochemical processes are larger than the time scales of turbulent processes mixing the system. A one-dimensional horizontal system (Fig. 10.1b), e.g. a channel type lake, might have well mixed conditions along its axis, horizontal transport processes have to be taken into account for this case. A classical example is the longitudinal decrease of dissolved oxygen in a river by degradation of biochemical oxygen demand (BOD) described by the Streeter-Phelps equation accounting for horizontal flow (Streeter and Phelps 1925). The one-dimensional vertical case (Fig. 10.1b) is an adequate description of a lake when neglecting horizontal gradients generated by boundary effects. Here, changes in the vertical

**Fig. 10.7** Different process types in a one- (a) and a two-dimensional (b) discretization of a lake

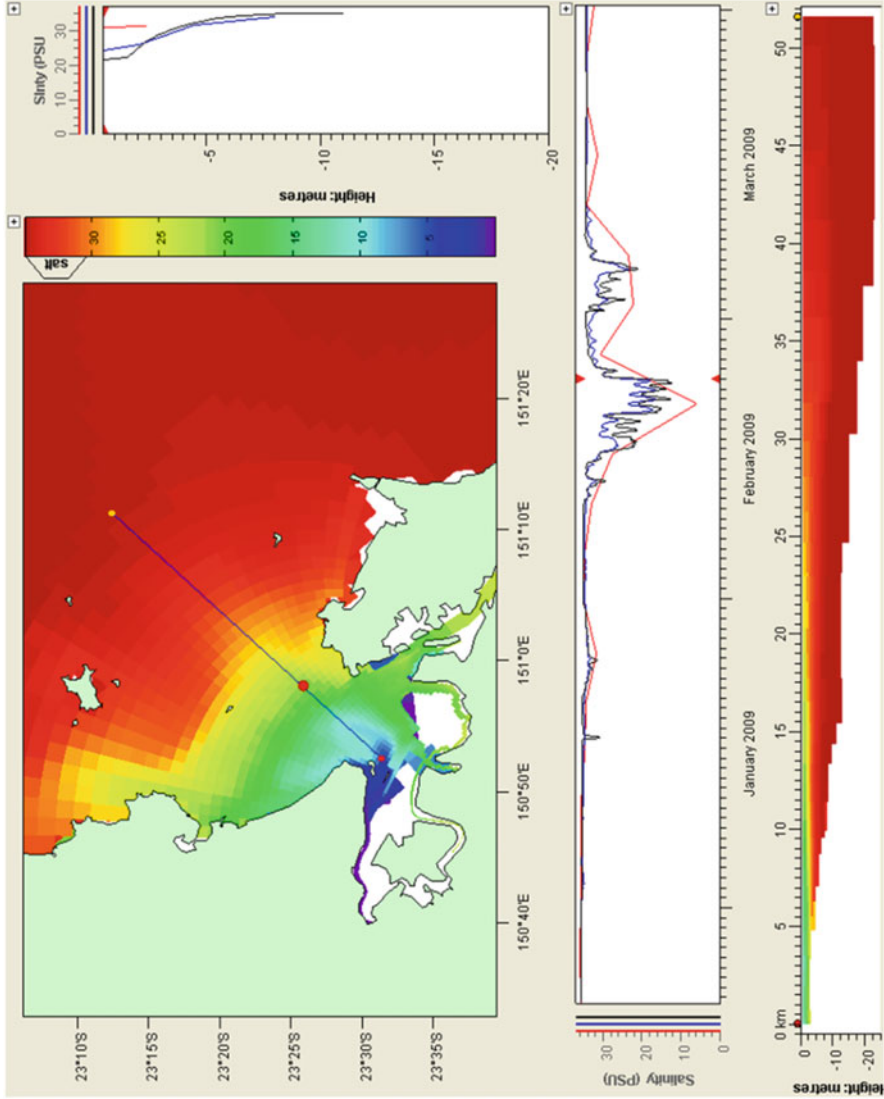




are generated by turbulent and convective mixing, which are attributed to wind stress at its surface or heat loss from the surface. This is a widely used approach in describing lake dynamics. Numerous models exist to describe the hydrodynamics and thermal characteristics of this type of lake approximation (e.g. Stepanenko et al. 2013, 2014) including complex biogeochemical processes (see Janssen et al. 2015 for an overview).

The more small scale processes of a water body need to be included, the more knowledge on interacting bio-physical processes as well as on its geometry is required. For a 1-dimensional case it is assumed that the water body can be idealized by a set of serially connected “grid cells” (Fig. 10.7a). In a 2-dimensional case grids are connected across two dimensions—attaching a set of 1D cell structures (Fig. 10.7b), and finally a three-dimensionally resolved lake would consist of slices of 2-dimensional grids fitting the shape of lake morphometry and communicating across cell boundaries. While cell-internal biogeochemical processes are described as in the examples above, the exchange between cells and the prescription of driving forces at the cells’ external boundaries have to be defined based on physical principles (Fig. 10.7). This increase in geometric complexity also accompanies a higher complexity in physical processes (Wüest and Lorke 2003), which in most cases makes it necessary to significantly lower the time step (down to minutes or even seconds) of numerical solvers to resolve the various time scales of physical processes. For 1- and 2-dimensional systems, the computational overhead of running hydrodynamic and food web models in parallel is not restricted by current computing technology. However, for large 3D models (or for lower dimensional models running a multitude of scenarios) it may be necessary to decouple hydrodynamics from biological dynamics and using averaged (e.g. hourly or daily) physical quantities for the biological model parts (e.g. Skerratt et al. 2013). The necessary higher time resolution and spatial knowledge of drivers and boundary conditions for 2D- or 3D- is often not met with actual lake monitoring, in which case a reduction of geometric complexity is more adequate to describe the problem.

Large lakes and estuaries often have a very complex shape, which introduces a further complication in higher dimensional modelling of such systems. To adequately describe the geometry of e.g. small bays or river channels attached to the larger water body, it is necessary to either substantially reduce the grid spacing in these sub-systems, which would increase the CPU time, or to implement sub-gridding. The latter uses a higher resolved grid only for the specific region and communicates with the low-resolution grid via boundary conditions at its matching side, i.e., prescribed fluxes, water level, etc. This allows for fast calculation of the large-scale transport mechanisms and high resolution in the sub-grid region. In the Fitzroy Estuary attached to the Great Barrier Reef, Australia, such sub-gridding is used to simulate the spread of freshwater plumes from rivers during flood events. While the large scale grid size in this coastal system is of the order of 4 km, the sub-gridded region has resolutions down to 200 m. Figure 10.8 shows a snapshot of the salinity distribution in the estuary for the sub-gridded region in comparison to the large scale grid solution (red lines).



**Fig. 10.8** Salinity simulation of river inflow into the Fitzroy Estuary, Australia, for a subgridded region (blue/black lines) in comparison to results for a large-scale ocean model (red lines). Profiles and time series are for the location in the middle of the estuary (red dot)

### ***10.3.1 Horizontal and Vertical Transport of Nutrients and Organisms***

Using simple connected regions for a lake, like the partitioning into epilimnion/hypolimnion for a food web model does not allow for a description of transport. Instead an exchange rate between compartments has to be described. This can be achieved by balancing the amount of constituents in the compartments over time and deriving the relative quantity of a constituent transferring over a time period from one into the adjacent compartment, e.g., particles sinking out of the epilimnion. For higher dimensional lake models with better spatial resolution, this process will be substituted with one based on physical transport mechanisms, i.e. advective transport of a quantity—temperature, particles, etc.—with a local velocity and re-distribution due to turbulent diffusion. The advective transport or velocity is a direct result of solving the hydrodynamic equations of motion in a lake. They describe a general flow pattern usually driven by wind action or inflow. Turbulent diffusion summarizes the small-scale processes generated by shearing in a fluid and due to local density instabilities. The latter is usually described as thermal instabilities when looking at freshwater systems, which are generated by surface cooling through heat loss usually during night-time. While the flow patterns act on time scales of hours to days or longer, the diffusional process describe fast processes with time scales of minutes or smaller. These diffusional processes are the drivers of constituent re-distribution in the water column. As such, the strength of turbulent diffusion will determine the amount of light a phytoplankton will be able to harvest while it is stochastically moved through the water column. Describing this motion through “diffusion” is again an approximation of the true process; it describes the mean distribution of a large amount of particles, but is not capable to follow the path of a single particle.

### ***10.3.2 Multi-segment Lake Model for Studying Dreissenids and Macrophytes***

The invasion of dreissenid mussels has been responsible for a major restructuring of the biophysical environment in many parts of the Laurentian Great Lakes, with profound alterations on the nutrient dynamics in the littoral zone (Coleman and Williams 2002). The nearshore shunt (sensu Hecky et al. 2004) has been hypothesized to impact the fate and transport of particulate matter, and subsequently alter the relative productivity of inshore sites and their interactions with the offshore areas. An important implication of the causal linkage between dreissenids and nutrient variability in the littoral zone is the weakening of the external loading signal, which led Hecky et al. (2004) to question whether conventional TP mass-balance models developed during the pre-dreissenid period in the Great Lakes were structurally adequate during the post-dreissenid era. In this context, Gudimov et al.

(2015) presented a mechanistic model designed to examine the role of macrophyte dynamics, to explicitly represent the impact of dreissenids in lakes, and to sensibly portray the interplay between water column and sediments.

In Lake Simcoe, Ontario, Canada, dreissenid mussel distribution is determined by a complex interplay among lake depth, substrate availability and exposure to wave disturbance (Ozersky et al. 2011; Evans et al. 2011). Specifically, the highest dreissenid biomass is typically found at areas of intermediate depth, where water movement is high enough to ensure that the lake bottom is dominated by rocky substrate but not excessively high to cause catastrophic disturbances to the dreissenid community. Gudimov et al. (2015) used their phosphorus mass-balance model to test the hypothesis that the spatial and temporal variability of P in Lake Simcoe was predominantly driven by internal mechanisms following the establishment of dreissenids. Because of the large size and complex shape of Lake Simcoe, a zero-dimensional spatial configuration would have been inadequate as the fundamental assumption that the lake is thoroughly mixed with uniform concentrations throughout is profoundly violated. On the other hand, there was not sufficient information (water levels, circulation patterns) to support the implementation of an explicit 2D or 3D hydrodynamic model. As an optimal compromise between the two strategies, the horizontal variability of Lake Simcoe was accommodated with four completely-mixed compartments, while the stratification patterns typically shaping the water quality in Kempenfelt Bay, Cook's Bay and the main basin were reproduced by the addition of three hypolimnetic compartments (Fig. 10.9, left panel). According to the Gudimov et al. (2015) model, the Lake Simcoe segmentation resembles Nicholls' (1997) conceptualization, in that the two embayments (Kempenfelt Bay and Cook's Bay) along with the shallow littoral zone at the east end (East Basin) are separated from the main basin (Fig. 10.9, left panel). The epilimnetic segments were interconnected through bi-directional hydraulic exchanges to account for wind-driven flows and tributary discharges from adjacent watersheds.

The Lake Simcoe model was designed to improve the fidelity of epilimnetic TP simulations through detailed specification of internal P recycling pathways (Fig. 10.9, left panel), such as the macrophyte dynamics and dreissenid activity as well as the fate and transport of P in the sediments, including the sediment resuspension, sorption/desorption in the sediment particles, and organic matter decomposition. Thus, the ordinary differential equations describing the dynamics of P in the water column consider all the external inputs, advective horizontal mass exchanges between adjacent segments, macrophyte uptake, macrophyte P release through respiration, dreissenid filtration, dreissenid excretion and pseudofeces egestion, vertical diffusive exchanges when stratification is established, and refluxes from the bottom sediments.

After the model calibration against the observed patterns in Lake Simcoe during the 1999–2007 period, Gudimov et al. (2015) first attempted to shed light on the role of the phosphorus fluxes associated with the dreissenid mussels. It was predicted that dreissenids filter a considerable amount of particulate P from the water column ( $6.2\text{--}238$  tonnes P year<sup>-1</sup>), but the effective clearance rate is

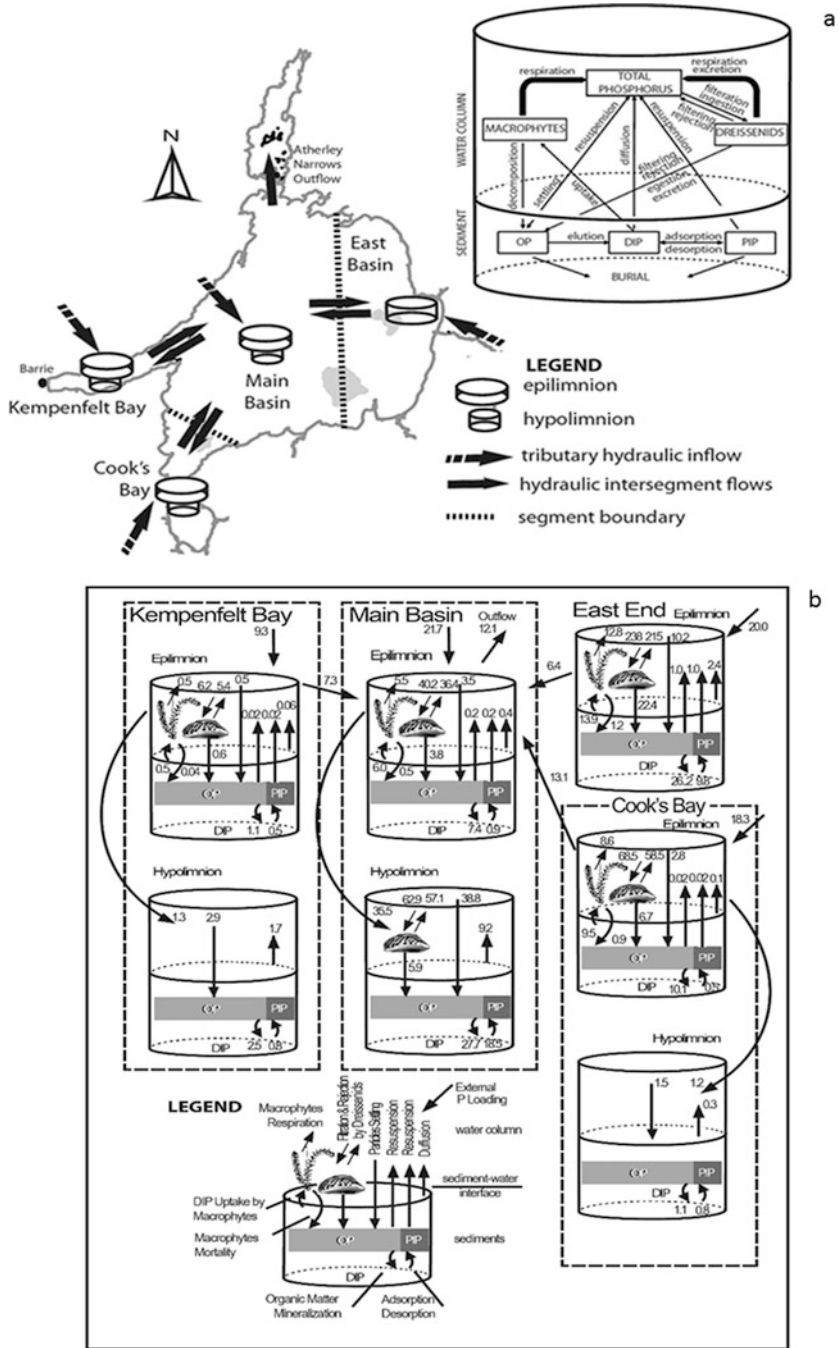


Fig. 10.9 Spatial segmentation and conceptual diagram of phosphorus pathways in the process-based model of Lake Simcoe (a). Simulated phosphorus fluxes (tonnes P year<sup>-1</sup>) in water column and sediment layer in the spatial segments of Lake Simcoe (b)

significantly lower (0.8–22.8 tonnes P year<sup>-1</sup>) with a substantial amount of the filtered particles (>85%) returned into the water column as feces, pseudofeces or other metabolic excreta. The latter finding is not surprising as the ratio between zebra mussel filtration and effective clearance rate can vary between 3.4 and 6.9 (Yu and Culver 1999). In particular, the Gudimov et al. (2015) study highlighted the critical role of dreissenids in the shallow eastern end of Lake Simcoe, where they filter 238.5 tonnes P year<sup>-1</sup> from the water column and subsequently egest 215.0 tonnes P year<sup>-1</sup>, while an additional 22.4 tonnes P year<sup>-1</sup> of metabolic excreta are deposited onto the sediments. Because of its shallow morphometry, a large portion of the eastern area is located within the euphotic and well-mixed zone, and therefore the elevated benthic photosynthesis and access of the dreissenids to sestonic algae create favourable conditions for biodeposition and nutrient recycling (Ozersky et al. 2013). Importantly, the large fetch of Lake Simcoe, the relatively deep epilimnion, and the fairly rapid horizontal mixing often induce hydrodynamic conditions that may allow the localized impacts of dreissenids to shape ecosystem-scale patterns (Schwalb et al. 2013).

Consistent with empirical evidence from the system, the Lake Simcoe model predicted that macrophyte intake was responsible for a significant loss of P from the interstitial waters, thereby providing a significant pathway for the rapid transport of the nutrients assimilated from the sediments into the water column. P diffusive fluxes from the sediments accounted for about 30–35% of the exogenous P loading in Lake Simcoe. The retention capacity in Cook's Bay was estimated to be about 28%, which is distinctly lower than estimates from the 1980s. Thus, the colonization of the embayment by dreissenids and the recent proliferation of macrophytes appear to have decreased the P retention in Cook's Bay, where the predominant fraction of TP is carbonate-bound P (apatite-P) mainly due to the accelerated erosion in the catchment (Dittrich et al. 2012). The sediments in the main basin are mostly driven by fast diagenetic processes of settling organic matter from the epilimnion, resulting in internal P loading of 9.2 tonnes P year<sup>-1</sup>. In a similar manner, the hypolimnetic sediments in Kempenfelt Bay are responsible for a fairly high diffusive P flux into the water column ( $\approx 1.7$  tonnes P year<sup>-1</sup>), presumably reflecting the highest proportion of the redox-sensitive P sediment pool compared to other lake segments (Dittrich et al. 2012).

## 10.4 Concluding Remarks

As knowledge regarding the complex components of environmental systems continues to grow, there is a demand for increasing the articulation level of our mathematical models. Generally, the premise for constructing complex models is to mirror the complexity of natural systems and consider ecological processes that can become important in future states and are driven by significantly different conditions. Modelers essentially believe in the myth that if they can include 'all' the significant processes in the mathematical equations, then the model will closely

mimic the 'real system' and thus will have increased predictive ability under a wide range of environmental conditions. However, there is always a trade-off between model complexity, transparency, uncertainty and validity, as well as data obtainability. Increasing computational potential is tempting to solve biogeochemical models in a 2- or 3-dimensional manner to cope with lateral changes in aquatic systems as well as including more complex physical transport phenomena (MacKay et al. 2009). However, this requires an adequate level of in-lake monitoring but also access to large scale resolved meteorological data as drivers of physical processes. Assimilation of remote sensing data with hydrodynamic modelling (e.g. Pinardi et al. 2015) may further improve predictive abilities of models (see also Chap. 15).

In the context of aquatic biogeochemical modeling, there is increasing pressure to explicitly treat multiple biogeochemical cycles, to increase the functional diversity of biotic communities, and to refine the mathematical description of the higher trophic levels (Arhonditsis and Brett 2004; Anderson 2005; Fennel 2008). In particular, there are views in the literature suggesting the inclusion of multiple nutrients along with the finer representation of plankton communities, as necessary model augmentations for disentangling critical aspects of aquatic ecosystem dynamics, e.g., species populations are more sensitive to external perturbations (nutrient enrichment, episodic meteorological events), and key biogeochemical processes are tightly linked to specific plankton functional groups (Flynn 2005). Nonetheless, the derivation of distinct functional groups from fairly heterogeneous planktonic assemblages poses challenging problems. Because of the still poorly understood ecology, we do not have robust group-specific parameterizations that can support predictions in a wide array of spatiotemporal domains (Anderson 2005).

Preliminary efforts to incorporate plankton functional types into global biogeochemical models were based on speculative parameterization and, not surprisingly, resulted in unreliable predictions (Anderson 2005). In the same context, a recent meta-analysis evaluated the ability of 124 aquatic biogeochemical models to reproduce the dynamics of phytoplankton functional groups (Shimoda and Arhonditsis 2016). Most notably, moderate fit statistics were found for diatoms (median  $r^2 = 0.31$ , RE = 70%) and cyanobacteria (median  $r^2 = 0.36$ , RE = 65%), and even worse performance was recorded for cryptophytes (median  $r^2 = 0.39$ , RE = 79%), flagellates (median  $r^2 = 0.07$ , RE = 78%) and haptophytes (median  $r^2 = 0.39$ , RE = 41%), which likely reflects our limited knowledge of their ecophysiological parameters compared to other well-studied functional groups. Significant variability also exists with respect to the mathematical representation of key physiological processes (e.g. growth strategies, nutrient kinetics, settling velocities) as well as group-specific characterizations typically considered in the pertinent literature. Furthermore, recent attempts to integrate biogeochemistry with fish production underscore the uncertainty arising from the mismatch between the operating time scales of planktonic processes and fish life cycles as well as the need to consolidate the mechanistic description and parameterization of several critical processes, such as the reproduction and mortality of the adult stages (Fennel 2008). Despite repeated efforts to increase model complexity, we still have not gone



beyond the phase of identifying the unforeseeable ramifications and the challenges that we need to confront so as to improve the predictive power of our models. Until we have the knowledge to mathematically depict the interplay among physical, chemical, and biological processes with greater fidelity and less uncertainty, the gradual incorporation of model complexity, where possible and relevant, is the most prudent strategy. The Bayesian analysis of model uncertainty will be addressed in detail in Chap. 11.

**Acknowledgements** The Cootes Paradise modeling project has received funding support from the Ontario Ministry of the Environment (Canada-Ontario Grant Agreement 120808). The Lake Simcoe modeling project was undertaken with the financial support of the Government of Canada provided through the Department of the Environment. We also thank the South Australian Water Corporation for financial and logistic support for studying the Millbrook and Mt Bold reservoirs.

## References

- Anderson TR (2005) Plankton functional type modelling: running before we can walk? *J Plankton Res* 27:1073–1081
- Arhonditsis GB, Brett MT (2004) Evaluation of the current state of mechanistic aquatic biogeochemical modeling. *Mar Ecol Prog Ser* 271:13–26
- Asaeda T, Trung VK, Manatunge J (2000) Modeling the effects of macrophyte growth and decomposition on the nutrient budget in Shallow Lakes. *Aquat Bot* 68(3):217–237
- Benndorf J, Recknagel F (1982) Problems of application of the ecological model SALMO to lakes and reservoirs having various trophic states. *Ecol Mod* 17:129–145
- Cao H, Recknagel F (2009) Hybridisation of process-based ecosystem models with evolutionary algorithms: multi-objective optimisation of process representations and parameters of the lake simulation library SALMO-OO. In: Jørgensen SE, Chon TS, Recknagel F (eds) *Handbook of ecological modelling and informatics*. WIT Press, Southampton, pp 169–185
- Cao H, Recknagel F, Cetin L, Zhang B (2008) Process-based simulation library SALMO-OO for lake ecosystems. Part 2: multi-objective parameter optimisation by evolutionary algorithms. *Ecol Inform* 3:181–190
- Chen Q, Zhang Z, Recknagel F et al (2014) Adaptation and multiple parameter optimization of the simulation model SALMO as prerequisite for scenario analysis on a shallow eutrophic lake. *Ecol Model* 273:109–116
- Chow-Fraser P (2005) Ecosystem response to changes in water level of Lake Ontario marshes: lessons from the restoration of Cootes Paradise Marsh. *Hydrobiologia* 539(1):189–204
- Coleman FC, Williams SL (2002) Overexploiting marine ecosystem engineers: potential consequences for biodiversity. *Trends Ecol Evol* 17(1):40–44
- Dittrich M, Chesnyuk A, Gudimov A et al (2012) Phosphorus retention in a mesotrophic lake under transient loading conditions: insights from a sediment phosphorus binding form study. *Water Res* 47(3):1433–1447
- Evans DO, Skinner AJ, Allen R, McMurtry MJ (2011) Invasion of zebra mussel, “*Dreissena polymorpha*”, in Lake Simcoe. *J Great Lakes Res* 37:36–45
- Fennel W (2008) Towards bridging biogeochemical and fish-production models. *J Mar Syst* 71:171–194
- Flynn KJ (2005) Castles built on sand: dysfunctionality in plankton models and the inadequacy of dialogue between biologists and modellers. *J Plankton Res* 27:1205–1210
- Granéli W, Solander D (1988) Influence of aquatic macrophytes on phosphorus cycling in lakes. *Hydrobiologia* 170(1):245–266



- Gudimov A, Kim D-K, Young JD et al (2015) Examination of the role of dreissenids and macrophytes in the phosphorus dynamics of Lake Simcoe, Ontario, Canada. *Ecol Inform* 26:36–53
- Harwell MC, Havens KE (2003) Experimental studies on the recovery potential of submerged aquatic vegetation after flooding and desiccation in a large subtropical lake. *Aquat Bot* 77 (2):135–151
- Hecky R, Smith RE, Barton D et al (2004) The nearshore phosphorus shunt: a consequence of ecosystem engineering by dreissenids in the Laurentian Great Lakes. *Can J Fish Aquat Sci* 61 (7):1285–1293
- Hutter K, Jöhnk KD (2004) Continuum methods of physical modeling – continuum mechanics, dimensional analysis, turbulence. Springer, Berlin
- Janssen ABG, Arhonditsis GB, Beusen A et al (2015) Exploring, exploiting and evolving diversity of aquatic ecosystem models: a community perspective. *Aquat Ecol* 49:513–548. doi:[10.1007/s10452-015-9544-1](https://doi.org/10.1007/s10452-015-9544-1)
- Joehnk KD, Umlauf L (2001) Modelling the metalimnetic oxygen minimum in a medium sized alpine lake. *Ecol Model* 136(1):67–80. doi:[10.1016/S0304-3800\(00\)00381-1](https://doi.org/10.1016/S0304-3800(00)00381-1)
- Kim D-K, Zhang W, Rao Y et al (2013) Improving the representation of internal nutrient recycling with phosphorus mass balance models: a case study in the Bay of Quinte, Ontario, Canada. *Ecol Model* 256:53–68
- Kim DK, Peller T, Gozum Z et al (2016) Modelling phosphorus dynamics in Cootes Paradise marsh: uncertainty assessment and implications for eutrophication management. *Aquat Ecosyst Health Manag* 19(1):1–18
- Lougheed VL, Theysmeyer T, Smith T, Chow-Fraser P (2004) Carp exclusion, food-web interactions, and the restoration of Cootes Paradise marsh. *J Great Lakes Res* 30(1):44–57
- MacKay MD, Neale PJ, Arp CD et al (2009) Modeling lakes and reservoirs in the climate system. *Limnol Oceanogr* 54(6):2315–2329
- Mayer T, Rosa F, Charlton M (2005) Effect of sediment geochemistry on the nutrient release rates in Cootes Paradise Marsh, Ontario, Canada. *Aquat Ecosyst Health Manag* 8(2):133–145
- Nicholls KH (1997) A limnological basis for a Lake Simcoe phosphorus loading objective. *Lake Reser Manage* 13(3):189–198
- Ozersky T, Barton DR, Depew DC et al (2011) Effects of water movement on the distribution of invasive dreissenid mussels in Lake Simcoe, Ontario. *J Great Lakes Res* 37:46–54
- Ozersky T, Barton DR, Hecky RE, Guildford SJ (2013) Dreissenid mussels enhance nutrient efflux, periphyton quantity and production in the shallow littoral zone of a large lake. *Biol Invasions* 15(12):2799–2810
- Pinardi M, Fenocchi A, Giardino C et al (2015) Assessing potential algal blooms in a shallow fluvial lake by combining hydrodynamic modelling and remote-sensed images. *Water* 7:1921–1942
- Prescott KL, Tsanis IK (1997) Mass balance modelling and wetland restoration. *Ecol Eng* 9 (1-2):1–18
- Recknagel F (1984) A comprehensive sensitivity analysis for an ecological simulation model. *Ecol Model* 26:77–96
- Recknagel F, Benndorf J (1982) Validation of the ecological simulation model SALMO. *Int Rev Hydrobiol* 67(1):113–125
- Recknagel F, Hosomi M, Fukushima T, Kong D-S (1995) Short- and long-term control of external and internal phosphorus loads in lakes – a scenario analysis. *Water Res* 29(7):1767–1779
- Schwalb A, Bouffard D, Ozersky T et al (2013) Impacts of hydrodynamics and benthic communities on phytoplankton distributions in a large, dreissenid-colonized lake (Lake Simcoe, Ontario, Canada). *Inland Waters* 3(2):269–284
- Shimoda Y, Arhonditsis GB (2016) Phytoplankton functional type modelling: running before we can walk? A critical evaluation of the current state of knowledge. *Ecol Model* 320:29–43
- Skerratt J, Wild-Allen K, Rizwi F et al (2013) Use of a high resolution 3D fully coupled hydrodynamic, sediment and biogeochemical model to understand estuarine nutrient dynamics

- under various water quality scenarios. *Ocean Coast Manage* 83:52–66. doi:[10.1016/j.ocecoaman.2013.05.005](https://doi.org/10.1016/j.ocecoaman.2013.05.005)
- Spear RC (1997) Large simulation models: Calibration, uniqueness and goodness of fit. *Environ Model Softw* 12:219–228
- Stepanenko VM, Martynov A, Jöhnk KD et al (2013) A one-dimensional model intercomparison study of thermal regime of a shallow, turbid midlatitude lake. *Geosci Model Dev* 6(4):1337–1352. doi:[10.5194/gmd-6-1337-2013](https://doi.org/10.5194/gmd-6-1337-2013)
- Stepanenko V, Jöhnk KD, Machulskaya E et al (2014) Simulation of surface energy fluxes and stratification of a small boreal lake by a set of one-dimensional models. *Tellus A* 66. doi:[10.3402/tellusa.v66.21389](https://doi.org/10.3402/tellusa.v66.21389)
- Streeter HW, Phelps EB (1925) A study of the pollution and natural purification of the Ohio river. III. Factors concerned in the phenomena of oxidation and reaeration, Public Health Bulletin no. 146, Reprinted by US Department of Health, Education and Welfare, Public Health Service, 1958, ISBN B001BP4GZI
- Thomasen S, Chow-Fraser P (2012) Detecting changes in ecosystem quality following long-term restoration efforts in Cootes Paradise Marsh. *Ecol Indic* 13(1):82–92
- Wüest A, Lorke A (2003) Small-scale hydrodynamics in lakes. *Annu Rev Fluid Mech* 35:373–412
- Yu N, Culver DA (1999) Estimating the effective clearance rate and refiltration by zebra mussels, *Dreissena polymorpha*, in a stratified reservoir. *Freshw Biol* 41(3):481–492

# Chapter 11

## Uncertainty Analysis by Bayesian Inference

George Arhonditsis, Dong-Kyun Kim, Noreen Kelly, Alex Neumann,  
and Aisha Javed

**Abstract** The scientific methodology of mathematical models and their credibility to form the basis of public policy decisions have been frequently challenged. The development of novel methods for rigorously assessing the uncertainty underlying model predictions is one of the priorities of the modeling community. Striving for novel uncertainty analysis tools, we present the Bayesian calibration of process-based models as a methodological advancement that warrants consideration in ecosystem analysis and biogeochemical research. This modeling framework combines the advantageous features of both process-based and statistical approaches; that is, mechanistic understanding that remains within the bounds of data-based parameter estimation. The incorporation of mechanisms improves the confidence in predictions made for a variety of conditions, whereas the statistical methods provide an empirical basis for parameter value selection and allow for realistic estimates of predictive uncertainty. Other advantages of the Bayesian approach include the ability to sequentially update beliefs as new knowledge is available, the rigorous assessment of the expected consequences of different management actions, the optimization of the sampling design of monitoring programs, and the consistency with the scientific process of progressive learning and the policy practice of adaptive management. We illustrate some of the anticipated benefits from the Bayesian calibration framework, well suited for stakeholders and policy makers when making environmental management decisions, using the Hamilton Harbour and the Bay of Quinte—two eutrophic systems in Ontario, Canada—as case studies.

### 11.1 Does Uncertainty Really Matter?

In the context of environmental management, the central objectives of policy analysis and decision-making are to identify the important drivers of ecological degradation, to pinpoint the sources of controversy, and to help anticipate the unexpected. The explicit consideration of uncertainty enables one to think more

---

G. Arhonditsis (✉) • D.-K. Kim • N. Kelly • A. Neumann • A. Javed  
University of Toronto Scarborough, Scarborough, ON, Canada  
e-mail: [georgea@utsc.utoronto.ca](mailto:georgea@utsc.utoronto.ca)

carefully about these matters, to elucidate the relative role of different causal factors, and to delineate contingency plans (Dawes 1988). Environmental problems have a way of resurfacing themselves and are rarely (if ever) solved completely. Nonetheless, even if some facets may change overtime, the core problems often remain the same. Thus, having a framework that rigorously evaluates the underlying uncertainty makes it much easier to distinguish between valid assumptions and erroneous actions and, thus, maximize the efficiency of adaptive management strategies (Morgan et al. 1992).

The concepts of “uncertainty” and “risk” are understood in a variety of different ways by scientists, stakeholders, policy makers, and the public in ecology/environmental science. Uncertainty is a generic term comprising many concepts (Pappenberger and Beven 2006). No direct measurement of an empirical quantity can be absolutely exact and, therefore, uncertainty arises from *random error* in direct measurements. In addition, biases are often introduced through the measuring apparatus and/or experimental protocols. This experimental procedure typically reflects the *systematic error* associated with the difference between the true value of the quantity of interest and the value to which the mean of the measurements converges as more measurements are taken. Another source of uncertainty lies in the *subjective judgments* used to overcome knowledge gaps and lack of empirical measurements related to the major ecological mechanisms and/or variables underlying the environmental problem at hand. *Inherent randomness* is often perceived as a distinctly different type of uncertainty in that it is in principle irreducible. Nonetheless, this indeterminacy is not considered a matter of principle in environmental science, but rather the product of our incomplete knowledge of the world. It is argued that once we shed light on unknown causal variables and important ecological processes, we should be able to reduce the apparent uncertainty. In cases of environmental policy analysis, where there is no clear empirical evidence and scientific support in favor of a certain management option, significant uncertainty arises from potential *disagreements* among decision makers and stakeholders, reflecting their different perspectives and conscious (or unconscious) biases. Perhaps, the most familiar source of uncertainty is the *variability* that environmental quantities demonstrate over time and space. While these quantities can be effectively described by frequency distributions, what we typically fail to acknowledge and effectively communicate is the degree of confidence about the parameters (mean, median, standard deviation or various percentiles) of these distributions given the available information in a certain location or time period.

Along the same line of thinking, all mathematical models are simplistic representations of natural ecosystems and, therefore, their application in an environmental policy analysis context introduces the so-called *approximation uncertainty* (Arhonditsis et al. 2007). This uncertainty stems from the assumptions made and imperfect knowledge used to determine model structure and inputs (Beck 1987; Reichert and Omlin 1997). Model input error mainly stems from the uncertainty underlying the values of model parameters, initial conditions, and forcing functions as well as the realization that all models are drastic simplifications of reality that approximate the actual processes, i.e., essentially, all parameters are effective

(e.g., spatially and temporally averaged) values unlikely to be represented by fixed constants (Arhonditsis et al. 2006). Model structure error arises from (1) the selection of the appropriate state variables (model endpoints) to reproduce ecosystem functioning, given the environment management problem at hand; (2) the selection of the suitable equations among a variety of mathematical formulations for describing the ecosystem processes, e.g., linear, quadratic, sigmoidal, and hyperbolic functional forms to reproduce fish predation on zooplankton (Edwards and Yool 2000); and (3) the fact that our models are based on relationships which are derived individually in controlled laboratory environments but may not collectively yield an accurate picture of the natural ecosystem dynamics (Arhonditsis et al. 2006).

The general premise for constructing mathematical models is to mirror the complexity of natural systems and account for all the ecological processes that can potentially become important in future hypothesized ecosystem states, and thus increase our predictive ability. Nonetheless, by striving for increased model complexity, and thereby (implicitly or explicitly) embracing a reductionist description of natural system dynamics, we accentuate the disparity between what we want to tease out from a mathematical model and what can realistically be observed given the available technology, staffing, and resources to study the natural system. In doing so, it often becomes impossible to impose quantitative (or even qualitative) constraints on what should be considered “acceptable” model performance (Beven 2006). This problem profoundly undermines the very basic application of mathematical models as inverse analysis tools, i.e., any information on the levels and the variability of the state (or dependent) variables is used through the model calibration exercise to infer the most likely values of independent variables (model parameters) typically representing ecological rates and functional properties of the abiotic environment and/or the biotic communities. Instead, what modelers encounter is a situation in which several distinct choices of model inputs lead to the same model output, i.e., many sets of parameters fit the data about equally well. This non-uniqueness of the model solutions is known in the modeling literature as *equifinality* (Beven 1993). In recognition of the uncertainty and equifinality problems, it is suggested that the model calibration practice should change from seeking a single “optimal” value for each model parameter, to seeking a distribution of parameter sets that all meet a pre-defined fitting criterion (Stow et al. 2007; Arhonditsis et al. 2007). These acceptable parameter sets may then provide the basis for estimating prediction error associated with the model parameters.

Model uncertainty analysis is an attempt to formulate the joint probability distribution of model inputs and then update our knowledge about this distribution after the consideration of the calibration dataset. In this regard, Bayesian inference represents a suitable means to combine existing information (prior) with current observations (likelihood) for projecting the future. Several recent studies illustrate how Bayesian inference techniques can be used to quantify the information that data contain about model inputs, to offer insights into the covariance structure among parameter estimates, and to obtain predictions along with uncertainty bounds for model outputs (Bayarri et al. 2007; Arhonditsis et al. 2007, 2008a, b).

Specifically, Bayesian calibration schemes have been introduced with simple mathematical models and statistical formulations that explicitly accommodate measurement error, parameter uncertainty, and model structure error. Nonetheless, the emergence of the holistic management paradigm has increased the demand for even more complex biogeochemical models with considerably greater uncertainty (Zhang and Arhonditsis 2008; Ramin et al. 2011; Reichert and Schuwirth 2012). In particular, there is increasing pressure for the development of integrated water quality models that effectively connect the watershed with downstream biogeochemical processes. This need stems from the emerging management questions related to contemporary climate and land use changes that should be connected with the receiving water bodies (Rode et al. 2010). In this context, significant progress has been made in regards to the computational demands and error propagation control through complex model structures (Dietzel and Reichert 2012; Kim et al. 2014).

In this chapter, we present two case studies that illustrate how the assessment of uncertainty can assist in developing integrated environmental modeling systems, overcoming the conceptual or scale misalignment between processes of interest and supporting information, and exploiting disparate sources of data that differ with regards to their quality and resolution. The two systems are the Hamilton Harbour and Bay of Quinte, Ontario, Canada. There is a great deal of modeling work that has been done toward establishing realistic eutrophication goals and impartially evaluating the likelihood of delisting the two systems as Areas of Concerns (AOCs). Existing watershed, eutrophication, and food web models shed light on different facets of the ecosystem functioning. Here, we address several critical questions that have emerged from these models: To what extent do the models coalesce with respect to their assumptions and inference drawn? What are the major sources of uncertainty that will ultimately determine the attainment of the existing delisting goals? Our aim is to highlight the major lessons learned about the watershed dynamics, the eutrophication phenomena, and the broader implications for food web integrity. We also place special emphasis on the knowledge gaps of our current understanding of the two systems. Our thesis is that the uncertainty stemming from several “ecological unknowns” can offer critical planning information to determine the optimal management actions in the two areas.

## **11.2 Hamilton Harbour**

### ***11.2.1 Introduction***

Located at the western end of Lake Ontario, Hamilton Harbour is a large 2150 ha embayment surrounded by a watershed of approximately 500 km<sup>2</sup> (HH RAP 2003). The harbour has a roughly triangular shape with a length of 8 km along its main axis and a maximum width of 6 km along its eastern shoreline. It has a maximum depth

of 23 m, an average depth of 13 m, a surface area of 21.5 km<sup>2</sup>, and a volume of  $2.8 \times 10^8$  m<sup>3</sup>. The harbour exchanges water with western Lake Ontario via the Burlington Ship Canal, which is a man-made canal, 836 m long, 89 m wide and 9.5 m deep. The residence time of the harbour is significantly reduced by these exchange flows, which have a large influence on water quality and hypolimnetic dissolved oxygen concentrations (Yerubandi et al. 2016). The majority of the loads of inorganic nutrients and organic matter entering Hamilton Harbour originate from the Woodward and Skyway wastewater treatment plants (WWTPs), combined sewer overflows (CSOs), and ArcelorMittal Dofasco and Stelco steel mills (Hiriart-Baer et al. 2009). Other significant loads are delivered by three main tributaries that feed into the Harbour: Grindstone Creek, Red Hill Creek, and Spencer Creek, which reaches the harbour through a 250 ha shallow area of both marsh and open water called Cootes Paradise (HH RAP 2003). While the Redhill Creek watershed is ~80% urbanized, much of Grindstone and Spencer Creeks remain undeveloped as less than 20% of their watershed areas has been developed (HH RAP 2003). As a consequence of the excessive loading of nutrients and other pollutants, the harbour experiences serious water quality problems, such as algal blooms, low water transparency, predominance of toxic cyanobacteria, and low hypolimnetic oxygen concentrations often beginning in early summer.

Hamilton Harbour has long been considered one of the most degraded sites in the Great Lakes, and was listed as one of the 43 Areas of Concern (AOCs)<sup>1</sup> in the mid-1980s by the Water Quality Board of the International Joint Commission (Hall and O'Connor 2016). Since then, the Hamilton Harbour Remedial Action Plan (RAP) has assembled a variety of government, private sector, and community participants to decide on actions to restore the harbour environment. To this end, the RAP identified a number of beneficial use impairments<sup>2</sup> (BUIs), including the beneficial use *Eutrophication or Undesirable Algae* (HH RAP 2003). The foundation of the remedial measures and setting of water quality goals for the restoration of the harbour was based on the premise that reducing ambient phosphorus concentrations could control the chlorophyll *a* concentrations and water clarity. Using a framework that involved data analysis, expert judgment, and modeling along with consideration of what was deemed desirable and achievable for the harbour (Hall et al. 2006), critical thresholds for the TP concentration were set at 17  $\mu\text{g L}^{-1}$ , chlorophyll *a* concentration at 10  $\mu\text{g L}^{-1}$ , Secchi disc depth at 3.0 m, while the

---

<sup>1</sup>Great Lakes Areas of Concern are designated geographic areas within the Great Lakes Basin that show severe environmental degradation.

<sup>2</sup>An impairment of beneficial uses means a change in the chemical, physical or biological integrity of the Great Lakes system sufficient to cause any of the following: Restrictions on Fish and Wildlife Consumption; Tainting of Fish and Wildlife Flavor; Degraded Fish and Wildlife Populations; Fish Tumors or Other Deformities; Bird or Animal Deformities or Reproductive Problems; Degradation of Benthos; Restrictions on Dredging Activities; Eutrophication or Undesirable Algae; Restrictions on Drinking Water Consumption or Taste and Odor Problems; Beach Closings; Degradation of Aesthetics; Added Costs to Agriculture or Industry; Degradation of Phytoplankton and Zooplankton Populations; Loss of Fish and Wildlife Habitat.

maximum allowable exogenous TP loadings in the harbour were set at  $142 \text{ kg day}^{-1}$  (Charlton 2001). Reductions of external TP loading into the harbour led to water quality improvement and resurgence of aquatic macrophytes, but the system still receives substantial loads of phosphorus, ammonia, and suspended solids from the WWTPs, as well as from non-point loading sources and, therefore, only moderate improvements in TP, chlorophyll *a* and total ammonia concentrations have been observed since the mid-1990s (Hiriart-Baer et al. 2009, 2016).

Environmental modeling has been an indispensable tool of the Hamilton Harbour restoration efforts and a variety of data-oriented and process-based models are in place to determine realistic water quality goals. However, none of the existing modeling efforts in the Hamilton Harbour had rigorously assessed the effects of the uncertainty underlying model predictions (parametric and structural error, misspecified boundary conditions) on the projected system responses, nor have models to address percentile-based standards been used (Zhang and Arhonditsis 2008). Given the substantial social and economic implications of management decisions, it is important to implement modeling practices accommodating the type of probabilistic standards that seem to be more appropriate for complex environmental systems, such as the Hamilton Harbour (Ramin et al. 2011). In the following sections, we review the modeling efforts conducted to date in order to quantitatively assess the uncertainty in implementing management actions, and to highlight the applicability of percentile-based standards for setting water quality targets in the Hamilton Harbour and its watershed.

### ***11.2.2 Eutrophication Modeling to Elucidate the Role of Lower Food Web***

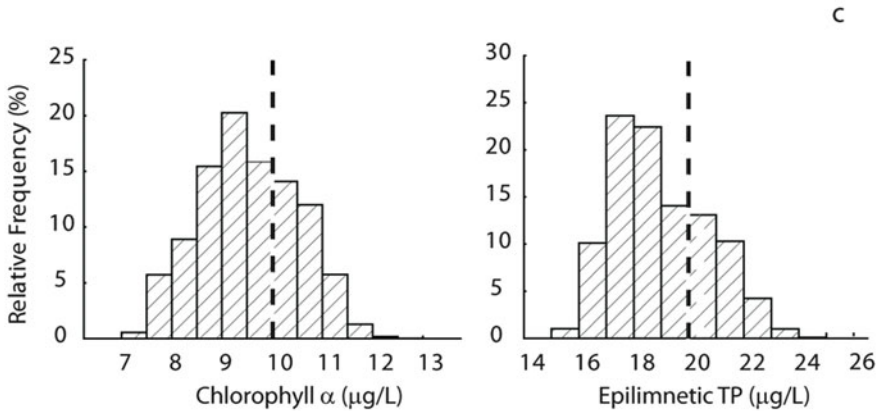
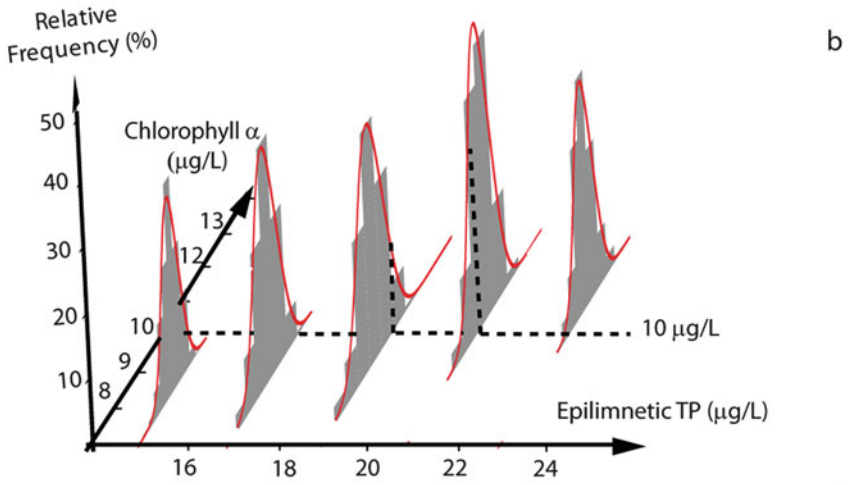
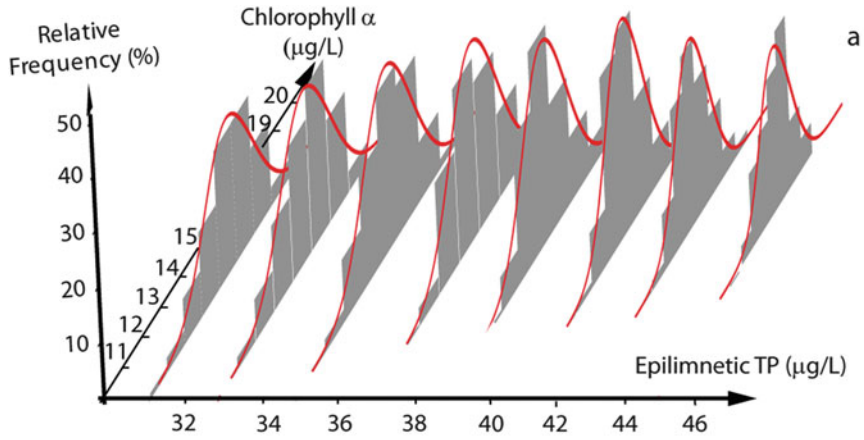
A series of process-based eutrophication models were built to depict the interplay among the different ecological mechanisms underlying the eutrophication problems, and to guide a water quality criteria-setting process that explicitly acknowledges the likelihood of standards violations in Hamilton Harbour (Gudimov et al. 2010, 2011; Ramin et al. 2011, 2012). As a starting point, Ramin et al. (2011) developed an ecological model that considered the interactions among eight state variables: nitrate, ammonium, phosphate, generic phytoplankton, cyanobacteria, zooplankton, organic nitrogen, and organic phosphorus. The model was based on a two-compartment vertical segmentation representing the epilimnion and hypolimnion of the harbour. The planktonic food web model was subsequently calibrated with Bayesian inference techniques founded upon a statistical formulation that explicitly accommodated measurement error, parameter uncertainty, and model structure imperfection. Concurrently with the Ramin et al. (2011) study, Gudimov et al. (2010) conducted a second (independent) modeling exercise with an upgraded model structure that utilized a three-compartment vertical segmentation representing the epilimnion, metalimnion, and hypolimnion, included three



phytoplankton functional groups to more realistically depict the continuum between diatom and cyanobacteria-dominated communities, and two zooplankton functional groups to account for the role of herbivorous and omnivorous zooplankton in the system. With these approaches, both Ramin et al. (2011) and Gudimov et al. (2010) provided a good representation of the seasonal variability of the prevailing water quality conditions and accurately reproduced the major cause-effect relationships underlying the harbour dynamics. Using the upgraded model structure, Gudimov et al. (2011) revisited several of the critical assumptions made in the previous two studies, and further explored the general uncertainty involved in their assumptions of ecosystem functioning. Building from these models, Ramin et al. (2012) used Bayesian averaging techniques to synthesize the forecasts from models of differing complexity to examine the robustness of earlier predictions regarding the harbour's response to nutrient loading scenarios (see Chap. 16).

These models collectively addressed two critical questions regarding the present status and future response of the Hamilton Harbour system: Is it possible to meet the eutrophication delisting goals of the AOC, if the RAP's proposed nutrient loading reduction targets are actually implemented? How frequently would these water quality goals be violated? The adoption of a water quality criterion that permits a pre-specified level of violations in space and time offers a more realistic assessment of the anticipated water quality conditions as it accommodates both natural variability and sampling error. Overall, similar projections were achieved by Ramin et al. (2011) and by Gudimov et al. (2010), projecting that the  $17 \mu\text{g TP L}^{-1}$  target would likely be met if the RAP phosphorus-loading target of  $142 \text{ kg day}^{-1}$  were achieved. However, by using a more representative summer epilimnetic TP dataset to calibrate the eutrophication model, Gudimov et al. (2011) demonstrated that the latter water quality target was too stringent, and most likely unattainable (Fig. 11.1). As corroborated by Ramin et al. (2012), a more pragmatic goal of  $20 \mu\text{g TP L}^{-1}$  would permit an acceptable frequency level of violations, e.g.,  $<10\%$  of the weekly samples during the stratified period (Fig. 11.1).

In contrast to the TP criterion, and depending on the assumptions made about the strength of the top-down control, as well as the importance of the internal nutrient sources (e.g., phosphorus release from the sediments, nutrient mineralization), Ramin et al. (2011) and Gudimov et al. (2010) provided evidence that the mean chlorophyll *a* target was achievable, although their projections had  $>50\%$  probability of exceeding the  $10 \mu\text{g L}^{-1}$  threshold level, even under the most drastic external nutrient loading reduction scenarios. In a follow-up study, Gudimov et al. (2011) revisited the ecological parameterization of the previous two models in order to test whether the chlorophyll *a* criterion could be achieved with a lower frequency of violations. With this analysis, two critical "ecological unknowns" were identified to influence the model's capacity to assess compliance with the chlorophyll *a* criterion; namely, the importance of the epilimnetic nutrient regeneration mediated by the microbial food web, and the likelihood of a structural shift in the lower food web towards a zooplankton community dominated by large-sized and fast-growing herbivores (e.g., *Daphnia*) (Gudimov et al. 2011). Given these uncertainties, Ramin et al. (2012) emphasized that the criteria setting process



**Fig. 11.1** Chlorophyll  $a$  predictive distributions for different levels of TP concentrations under (a) the present and (b) the Hamilton Harbour RAP loading targets (see text). Panel (c) illustrates the

should allow for a realistic percentage of violations of the target, such that exceedances of <10–15% of the weekly samples collected during the stratified period should still be considered as compliance, in order to explicitly accommodate the natural variability or inherent unpredictability of the system response.

In the same context, the uncertain role of planktivory and sediment diagenesis in the system emerged as two additional important ecological mechanisms for achieving the water quality targets in Hamilton Harbour. Gudimov et al. (2010) provided evidence that the anticipated structural shifts of the zooplankton community could determine the restoration rate, as well as the stability of the new trophic state in the harbour. Larger zooplankton taxa are particularly efficient in suppressing the standing phytoplankton biomass, but are also preferentially consumed by fish, and therefore the level of planktivory may shape the response rate to the nutrient loading reductions (Gudimov et al. 2010). Further, Gudimov et al. (2011) demonstrated that the epilimnetic TP concentrations were highly sensitive to the internal phosphorus loading assumptions, as a nearly two-fold increase of the sediment fluxes dramatically increased the number of violations of the TP delisting target. Thus, the internal nutrient loading from the sediments may be an important regulatory factor of the harbour.

The accuracy of the predictions made by the eutrophication model is conditional upon the credibility of the nutrient loading estimates to the harbour, which were highly uncertain and inadequately accounted for the contribution of non-point sources, episodic meteorological events (e.g., spring thaw, intense summer storms), and short-term variability at the local WWTPs (Gudimov et al. 2010, 2011). These uncertainties could potentially influence the exceedance frequency and the confidence of compliance with the water quality standards, particularly during the summer-stratified period (Gudimov et al. 2010). Given the pivotal role played by ambient phosphorus in the ecology of this system, there is a clear need to improve the tributary loading estimates in the area.

### 11.2.3 Nutrient Export Modeling for the Hamilton Harbour Watershed

The identification of the major nutrient source areas in the Hamilton Harbour watershed is of great management interest, as subwatersheds characterized by both high total delivery and high delivery per area are priority areas for management intervention. However, considerable knowledge gaps exist regarding the

---

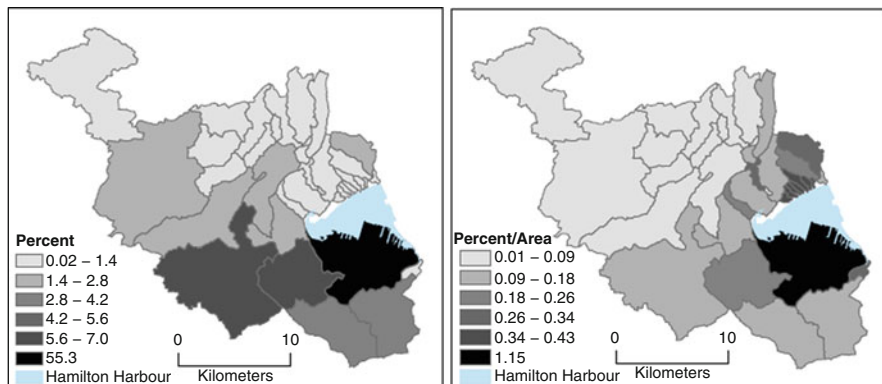
**Fig. 11.1** (continued) predictive distributions of chlorophyll *a* and epilimnetic TP concentrations examined to accommodate the inter- and intra-annual variability. *Vertical dashed lines* indicate the water quality targets of  $10 \mu\text{g}\cdot\text{L}^{-1}$  chl *a* and  $20 \mu\text{g}\cdot\text{L}^{-1}$  epilimnetic TP [Reproduced from Gudimov et al. (2011)]

complex interplay among hydrological factors, geological features, land uses, and spatial patterns of the built environment that modulates the attenuation rates of nutrient and contaminants. Following the development of the eutrophication models, Wellen et al. (2012, 2014a, b, c) employed two different watershed models to advance our understanding of how urban sites cycle nutrients and contaminants, so planning decisions that least impact Hamilton Harbour can be better informed.

Wellen et al. (2012, 2014a) implemented Bayesian inference techniques to parameterize the SPARROW (SPATIally Referenced Regressions On Watershed attributes) non-linear regression model in the Hamilton Harbour watershed. SPARROW is a spatially distributed, hybrid empirical/process-based model that estimates the relation between in-stream measurements of nutrient fluxes and the sources and sinks of nutrients within watersheds over annual timescales (McMahon et al. 2003). Source processes are described with export coefficients that predict TP mobilization, while the sink processes are represented by delivery factors, predicting how landscape attributes modulate the delivery of mobilized TP to streams, and attenuation coefficients, predicting the amount of the delivered TP remaining in transit per length of stream or per reservoir. With the SPARROW strategy, a two-level hierarchical structure is implemented, where watersheds are first divided into subwatersheds that each drain to a water-quality monitoring station, then each subwatershed is further divided into reach catchments draining to a particular stream segment (Schwarz et al. 2006).

Using data from Ontario's Provincial Water Quality Monitoring Network (PWQMN), Wellen et al. (2012, 2014a) offered the first estimates of export coefficients and delivery rates from the different subcatchments and generated testable hypotheses regarding the nutrient export "hot spots" in the studied watershed. The derived total phosphorus export estimates suggest that urban land uses may export more phosphorus per area than agricultural lands. This finding was somewhat contrary to the popular notion that the rates of nutrient export from urban lands are lower than those of agricultural lands due to lower nutrient subsidies. Wellen et al. (2014a) was able to show that subwatersheds which are both large and in close proximity to Hamilton Harbour have the highest nutrient delivery values per area, as the attenuation of their loads en route to the system is very low and the urban developments are more concentrated along the shore (Fig. 11.2).

The same modeling work has demonstrated that stream attenuation coefficients are quite variable in time (Fig. 11.3). The mechanisms that modulate the variability of nutrient attenuation across stream size are fairly well established in the literature. They generally refer to the tighter coupling of smaller streams with their streambeds, whereby biological and chemical removal processes in the sediments have greater access to nutrients in the water column (Alexander et al. 2004). The longer hydraulic residence time of smaller streams allows these processes to operate for longer times. Recent work suggests that stream stage explains the inter-annual variation of nutrient attenuation at a particular site over time, implying that the coupling between streambed and water column changes from year to year (Basu et al. 2011). Consistent with these findings, Wellen et al. (2012) showed that the inter-annual variability of the average discharge, a function of stream stage, can

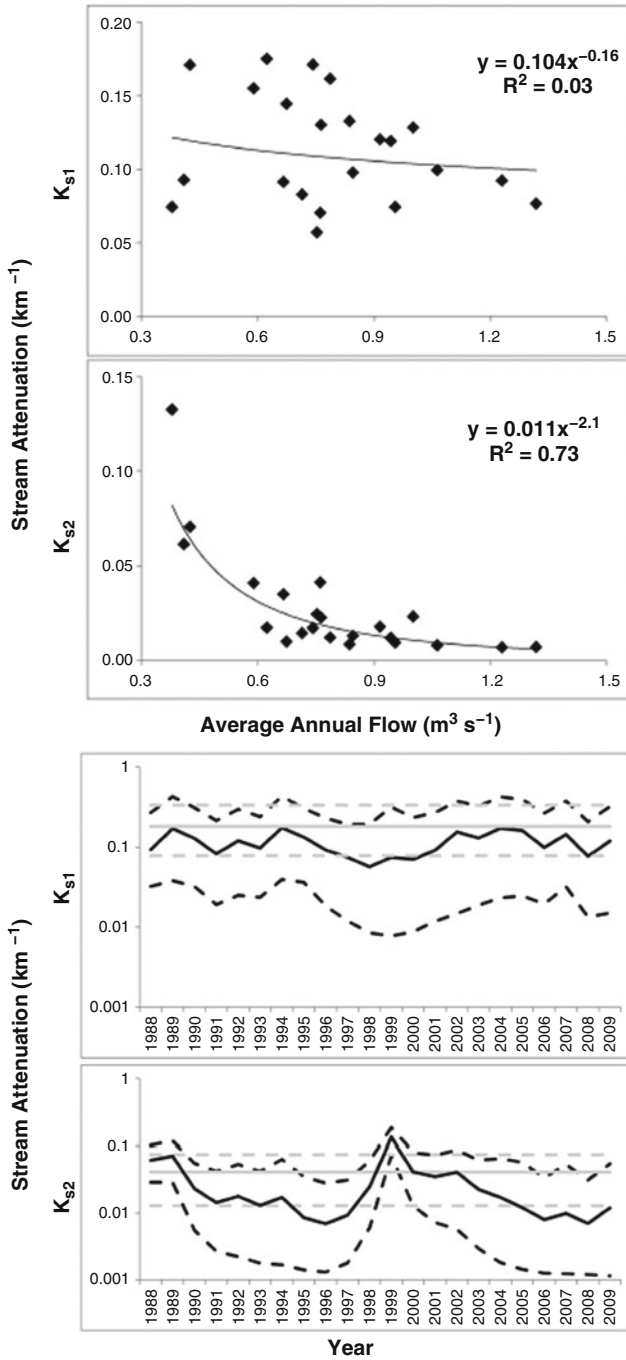


**Fig. 11.2** Estimated contribution of each subwatershed to the total phosphorus loading in Hamilton Harbour. The map on the *left* expresses the load of each subwatershed as a percentage of the total phosphorus load, including the combined sewer overflows and taking into account attenuation en route to Hamilton Harbour. The map on the *right* normalizes the percentage contribution by the corresponding subwatershed areas [Reproduced from Wellen et al. (2014a)]

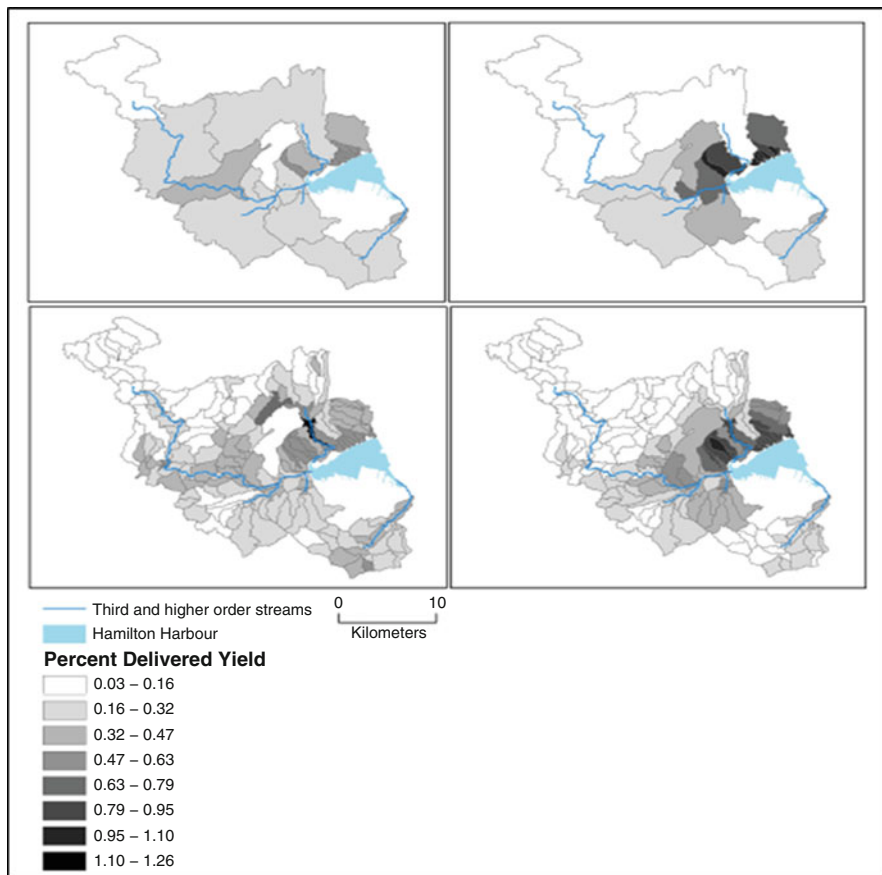
explain more than half of the variability of stream attenuation estimates from the SPARROW model in higher-order streams.

An interesting implication of the Hamilton Harbour’s SPARROW modeling is that the year-to-year variability of the contribution of phosphorus source areas may be strongly affected by the capacity of stream reaches to attenuate nutrient loads (Fig. 11.4). Empirical studies of nutrient uptake in rivers indicate significant variability of nutrient attenuation rates at annual timescales for phosphorus (Doyle et al. 2003) and nitrogen (Claessens et al. 2009). Donner et al. (2004) found that nutrient attenuation rates varied nearly two-fold between wet and dry years in the Mississippi River, with wet years exhibiting lower attenuation. Basu et al. (2011) also showed an inverse relationship between stream stage and nutrient attenuation that was consistently manifested across spatial and temporal scales. This finding implies that fluctuations in stage (and discharge) may indeed affect the spatial location of significant nutrient source areas at various scales. While previous research has documented the variability of in-stream attenuation at annual time-scales, the Hamilton Harbour modeling work allowed estimating how this variability impacts basin-scale nutrient source areas.

Wellen et al. (2014a) applied the SPARROW model to evaluate the potential improvement of parameter estimates (and the decrease of predictive uncertainty) if the precision of the currently available nutrient loading estimates in Hamilton Harbour is increased. Parameter identification was overwhelmingly improved with an increase in the spatial intensity of sampling stations, while an increase in the credibility of the measured nutrient loads significantly reduced the uncertainty of the model predictions, even when the number of stations monitored was halved (Wellen et al. 2014a). When a higher quality dataset was used to parameterize the model, the subwatersheds that displayed the greatest contraction in their 95%



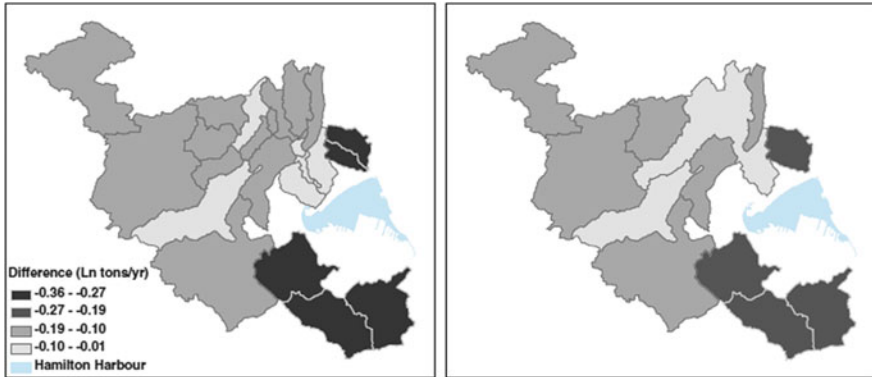
**Fig. 11.3** (Upper panels) Scatterplots of yearly total phosphorus stream attenuation rates ( $k_{s1}$  refers to attenuation in first- and second-order streams,  $k_{s2}$  to attenuation in third- and



**Fig. 11.4** Spatio-temporal variability of total phosphorus delivered yield at the watershed (*top panels*) and reach (*bottom panels*) scales. (*Left panels*) The percent contribution of total load into the Hamilton Harbour per square kilometer for 2006, the year with the lowest value of  $k_{s,2}$ . (*Right panels*) The percent contribution of total load to the Harbour per square kilometer for 1999, the year with the highest value of  $k_{s,2}$  [Reproduced from Wellen et al. (2012)]

**Fig. 11.3** (continued) higher-order streams) against annual average streamflow. (*Bottom panels*) Time series plots of the two attenuation coefficients over a 22-year study period (1988–2009). *Dashed black lines* indicate upper and lower limits of the 95% credible intervals (In Bayesian statistics, a **credible interval** is an interval in the domain of a posterior probability distribution used for interval estimation. Credible intervals are analogous to confidence intervals in frequentist statistics, but differ on a philosophical basis; Bayesian intervals treat their bounds as fixed and the estimated parameter as a random variable, whereas frequentist confidence intervals treat their bounds as random variables and the parameter as a fixed value.); *solid black lines* indicate the medians of the posterior distributions of the two coefficients. *Grey lines* depict the attenuation rate values typically reported in the literature [Reproduced from Wellen et al. (2012)]



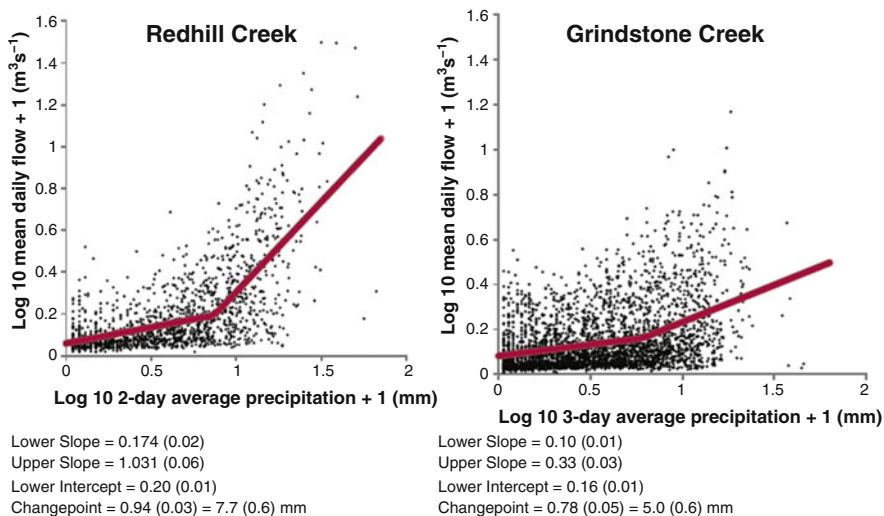


**Fig. 11.5** Value of information of additional monitoring in the Hamilton Harbour watershed. Maps show the difference between the width of the 95% credible intervals of the posterior loading estimates derived from the high and the current precision scenarios for sampling with all 24 stations originally used to calibrate the SPARROW model (*right*) and sampling with a subset of 12 stations (*left*) [Reproduced from Wellen et al. (2014a)]

credible intervals were the headwater streams as well as locations closest to the harbour characterized by high delivery rates and urban land uses (Fig. 11.5). Using the uncertainty patterns provided by the SPARROW model predictions, Wellen et al. (2014a) proposed that additional water-quality data-collection efforts in the watershed should be focused on “hot spots” sites characterized by: (1) a mid-range likelihood of impairment (*i.e.* the probability of exceeding a threshold level lying within the 25–75% range); (2) model predictions of unacceptably high variance; (3) locations where data uncertainty drives the model residuals; and/or (4) locations where modeled loads showed the greatest reduction in the width of their 95% credible intervals when higher quality dataset are obtained.

Even though the SPARROW modeling exercise has gained considerable insights, the annual resolution of the latter model, along with the fact that the PWQMN program collects monthly samples primarily during baseflow conditions, impedes the accurate characterization of TP dynamics during high flow conditions. In particular, examination of the daily flows of Redhill and Grindstone Creeks supports the idea of a single threshold separating two states of response of the two Creeks to precipitation (Wellen et al. 2014b). Figure 11.6 shows scatterplots of  $\log_{10}$  transformed daily flows and averages of the previous 2 or 3 days of precipitation along with the fitted piecewise regressions. These periods were chosen to implicitly include the effect of antecedent moisture. The data used are from the period 1988–2009, representing the months from May through November. Redhill Creek’s threshold was estimated at a 2-day average of 7.7 mm, and would be reached by one day with 15.2 mm of precipitation or 2 days of 7.7 mm. Grindstone Creek’s threshold was estimated to be a 3-day average of 5.0 mm. It was hypothesized that the watershed response to precipitation occurs in distinct states, such that precipitation depth above these thresholds triggers an



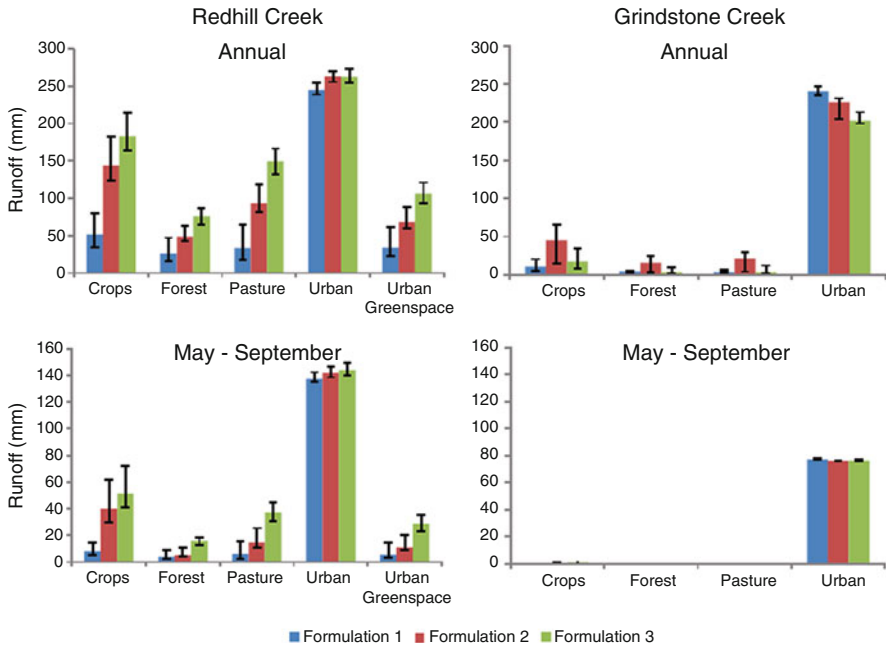


**Fig. 11.6** Piecewise regression graphs relating the 2- or 3-day average precipitation to the daily streamflow measured from 1988 to 2009. Only data from the months May–November are plotted. Statistics below graphs show the means and, *in parentheses*, standard deviations of the parameters of the regressions [Reproduced from Wellen et al. (2014b)]

extreme state, which is characterized by a qualitatively different response of the watershed to precipitation.

To solidify this working hypothesis, Long et al. (2014, 2015) collected 87 24-h level-weighted composite samples from a variety of catchment states (rain, snowmelt, baseflow) from all four major tributaries to Hamilton Harbour between July 2010 and May 2012. The key findings from this research were as follows: (1) daily TP loads varied by three orders of magnitude between wet and dry conditions, with storm events and spring freshets driving peak daily loads in urban and agricultural watersheds, respectively; (2) areal TP loads were significantly higher from the urban relative to the agricultural watersheds; and (3) the characterization of TP concentrations during high flow conditions was essential in establishing accurate concentration versus flow relationships and subsequently nutrient load estimates. The brief but intense events that occurred less than 10% of the time were found to be responsible for 50–90% of TP loads delivered from local tributaries.

Capitalizing upon this high-resolution dataset, a SWAT model was used to simulate the water cycle and sediment export in the area (Wellen et al. 2014b, c). Surface runoff is the primary pathway through which many pollutants (including phosphorus) enter waterways, and so identifying sources of surface runoff can aid in locating possible pollutant source areas (McDowell and Srinivasan 2009). In Fig. 11.7, estimates of surface runoff generation are presented for the different land uses in Redhill and Grindstone Creeks across three formulations (i.e., different statistical configurations of the Bayesian calibration framework; see Wellen et al. 2014b). Runoff generated during the entire year was distinguished from runoff

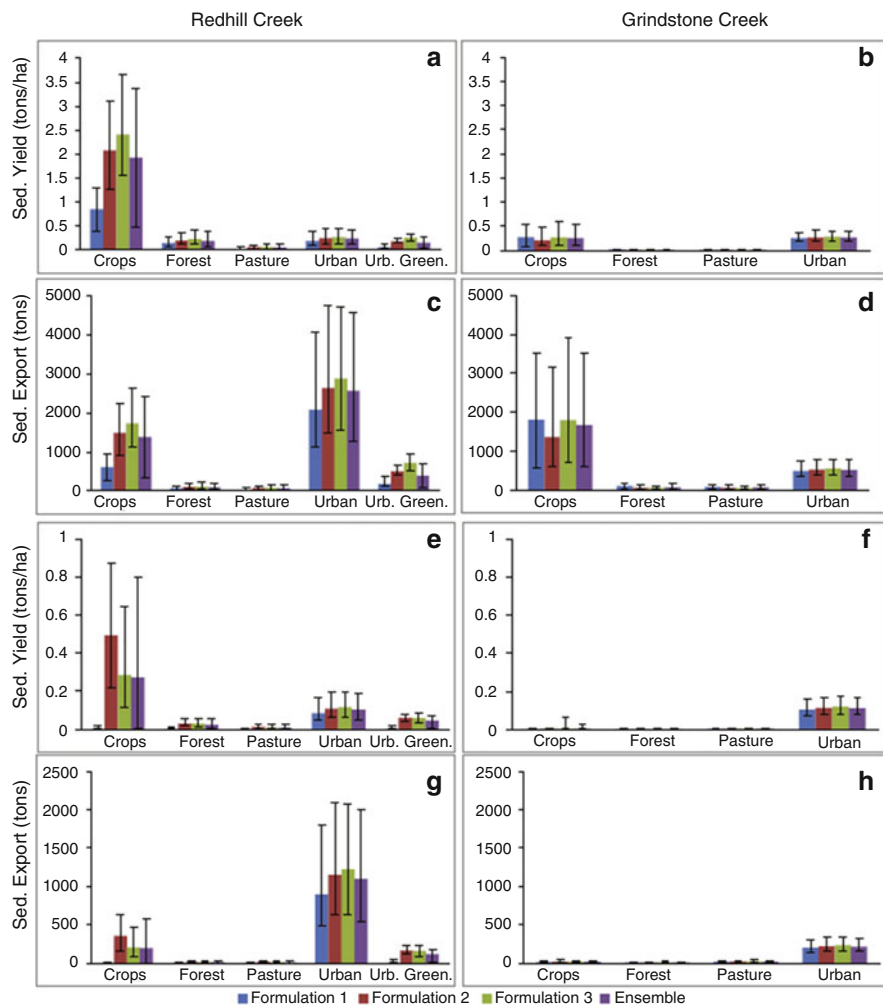


**Fig. 11.7** SWAT predictions of surface runoff depths generated in Redhill and Grindstone Creeks by different land uses. Formulations 1–3 correspond to different statistical configurations of the Bayesian calibration framework as presented in Wellen et al. (2014b). The error bars indicate 95% credible intervals of the predictions

generated during the growing season (May–September), as this is the period when the receiving water body is most sensitive to eutrophication. In both Creeks, urban land use generated the greatest depth of runoff; 245–262 mm for Redhill Creek and 202–240 mm for Grindstone Creek. For Redhill Creek, this compares to 51–183 mm for crops, 26–76 mm for forest, 34–149 mm for pasture, and 34–106 mm for urban green space. For Grindstone Creek, the urban runoff estimate compares to 11–45 mm for crops, 3–16 mm for forest, and 3–21 mm for pasture. During the growing season, this disparity became more acute, particularly in Grindstone Creek. Between May and September, runoff generation in Redhill Creek ranged from 8–51 mm for crops, 4–16 mm for forest, 6–37 mm for pasture, and 6–29 mm for urban green space. For Grindstone Creek, this compares to 1 mm for crops, <1 mm for forest, and <1 mm for pasture. Urban areas effectively by-pass catchment storage, as nearly all the precipitation falling on them becomes surface runoff and reaches the stream in less than one day, leaving little time for evapotranspiration. While the importance of urban areas as a surface runoff source increased slightly during the growing season in Redhill Creek, the model surprisingly predicts that almost no surface runoff reaches the stream from any of the pervious surfaces in Grindstone Creek from May to September. While it is likely that the contribution of runoff for Grindstone Creek is somewhat underestimated,

there seem to be important differences in soil type and/or vegetation cover between the two catchments which may be responsible for generating the markedly different amounts of runoff during the growing season.

Despite the small aerial coverage of the agricultural areas in Redhill Creek (5%) and the urban areas in Grindstone Creek (9%), these areas were responsible for a disproportionate amount of overland sediment export to streams (Fig. 11.8). Cropland was estimated to contribute between 20% and 30% of Redhill Creek’s



**Fig. 11.8** SWAT predictions of sediment yield and export by land use for the entire annual cycle (2010–2012; a–d) and for the growing season (May–September, 2010–2012; e–h). Ensemble refers to the averaged predictions of the three statistical configurations of the Bayesian calibration framework presented in Wellen et al. (2014c). The *error bars* indicate 95% credible intervals of the predictions

total sediment export to streams (720–3299 tons), while urban areas were estimated to contribute between 17% and 36% of Grindstone Creek’s total sediment export (410–1830 tons). During the growing season, urban residential areas are the main sources of sediment export to both streams, comprising 70–99% of all sediment exported to streams in Redhill Creek (217–1143 tons) and 60–81% of all estimated sediment exported to Grindstone Creek (74–214 tons).

During the calibration of the sediment routing submodel, reliable data were not available on stream bankwidth and depth. In order to draw reliable inferences on the sediment yield and streambed sediment storage status for Redhill and Grindstone Creeks at the sub-basin scale, Wellen et al. (2014c) used the entire predictive range of sediment storage for each subbasin (bed storage = upstream sediment in + erosional sediment in – downstream sediment out) (Fig. 11.9). It was assumed that if the 95% credible interval of the bed storage distribution was non-overlapping

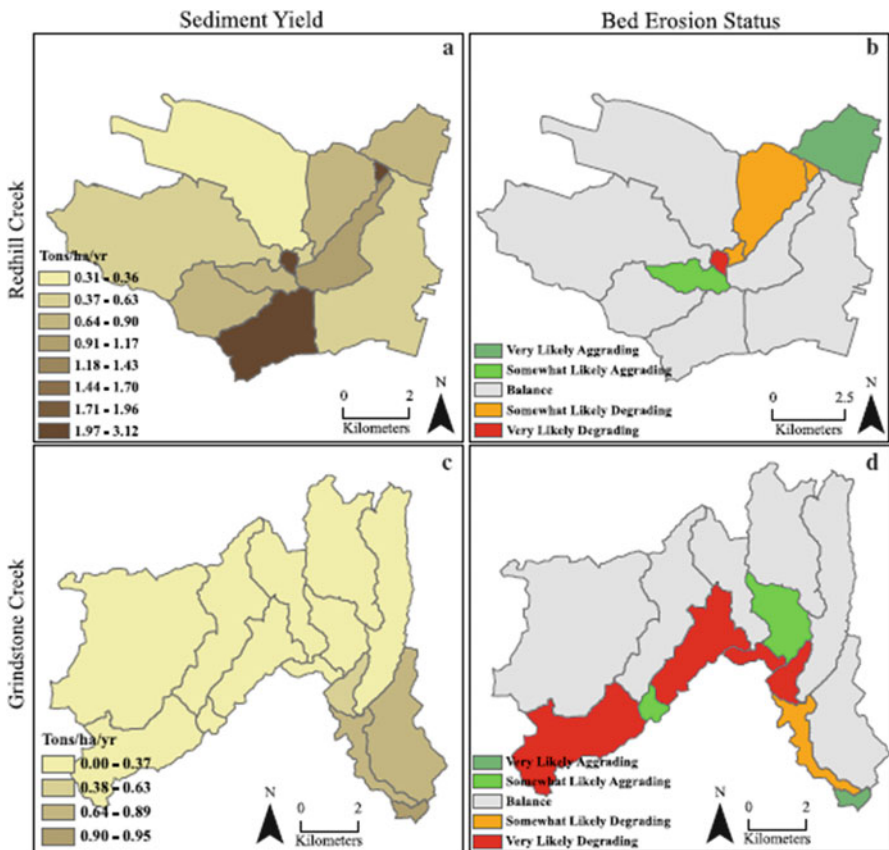


Fig. 11.9 Estimated sediment yield and bed erosion status for Redhill and Grindstone Creeks [Reproduced from Wellen et al. (2014c)]

with zero, reliable statements could be made about whether the reach was gaining or losing sediment during the period 2010–2012. If the bed storage was positive, the reach was categorized as very likely aggrading, while if the bed storage was negative, the reach was categorized as very likely degrading. If there was overlap with zero, the reach was categorized as likely aggrading or degrading, depending on which side of zero the median of the distribution laid. Some reaches categorized as balanced, as their credible intervals of absolute bed storage were less than 1 ton per year. The headwater areas of both Creeks were classified as balanced, while all the reaches losing sediment from their bed are located along the main channel. The final downstream reach was characterized as gaining sediment in both Creeks, reflecting the wider streams and gentler slopes. Notably, the sub-basin characterized as having the highest class of sediment yield in Redhill Creek's southern end was in balance, indicating that the substantial agricultural sediment mass estimated to be added to the streams in that reach was largely propagated downstream. In Grindstone Creek, there are few reaches that are storing sediment. In particular, the reaches containing most of the urban area towards the mouth of the basin are either at balance or likely degrading, implying that much of the urban sediment added to Grindstone Creek is exported downstream.

## 11.3 Bay of Quinte

### 11.3.1 Introduction

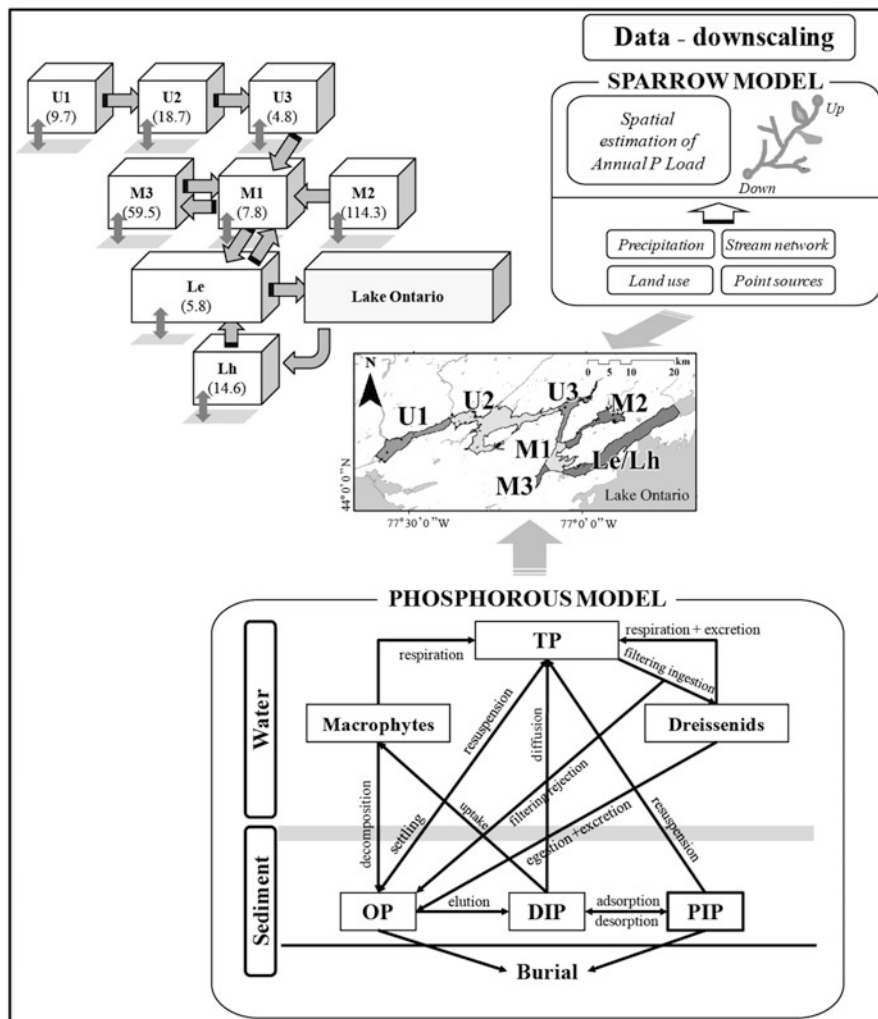
The Bay of Quinte, a Z-shaped embayment at the northern end of Lake Ontario, has experienced a long history of eutrophication problems, characterized by frequent and spatially extensive algal blooms, predominance of toxic cyanobacteria, dominance (or invasion) of undesirable fish species, and destruction of wildlife habitats (Arhonditsis et al. 2016, Shimoda et al. 2016). Because of these ecological degradation problems, the Great Lakes Water Quality Agreement between the United States and Canada established a number of objectives, guidelines, and initiatives to restore and maintain physicochemical and biological integrity. The Bay of Quinte was designated as one of the 43 Areas of Concern around the Great Lakes by the International Joint Commission (IJC) in 1986, whereby the Canadian government made a commitment to introduce a comprehensive action plan that primarily aimed to control nutrient loading from municipal sewage treatment plants. Phosphorus reduction in detergents along with upgrades at the WWTPs resulted in a dramatic reduction (>95%) of the phosphorus discharges from the 1960s, 215 kg day<sup>-1</sup>, to the 2000s, <10 kg day<sup>-1</sup> (Kinstler and Morley 2011).

Despite the substantial improvement of the ambient water quality conditions, high P concentrations and summer cyanobacteria blooms remain a central issue in

the bay (Watson et al. 2011). Invasions of zebra (*Dreissena polymorpha*) and quagga (*Dreissena bugensis*) mussels have further complicated ecosystem structure and functioning since the mid-1990s (Dermott and Bonnell 2011). In the post-dreissenid era, total phosphorus concentrations demonstrate significant within-year variability, characterized by relatively low spring and fall levels, 10–15  $\mu\text{g TP L}^{-1}$ , and high summer concentrations,  $> 50 \mu\text{g TP L}^{-1}$  (Shimoda et al. 2016). This ambient TP variability may also stem from the biological nutrient regeneration and sediment diagenesis processes, reflecting the impact of the memory of the system (Kim et al. 2013).

Existing empirical evidence suggests that the presence of dreissenids may have led to structural changes that could ultimately be translated into an ecosystem regime shift (deYoung et al. 2008). Namely, in the Bay of Quinte, increased light penetration resulting from dreissenid filtration of suspended solids stimulated the growth of submerged macrophytes that rapidly proliferated into deeper waters (Leisti et al. 2012). Regarding the phytoplankton community, the dreissenid invasion could cause shifts of the algal assemblage stemming directly from their feeding selectivity or indirectly from an increase in water column transparency, although the role of the feedback loop associated with their nutrient recycling activity could not be ruled out (Arhonditsis et al. 2016). Specifically, the arrival of dreissenid mussels coincided with both desirable (e.g., *Aphanizomenon* and *Oscillatoria* decline) and undesirable (e.g., *Microcystis* increase) shifts in the phytoplankton community composition (Shimoda et al. 2016). The increased frequency of harmful algal blooms in the post-dreissenid period has profound ramifications for several beneficial use impairments in the Bay of Quinte, such as *Eutrophication or undesirable algae*, *Restrictions on drinking water or taste and odor problems*, and *Degradation of aesthetics*.

Environmental modeling has been an indispensable tool of the Bay of Quinte restoration efforts and a variety of data-oriented and process-based models have been used for elucidating ecosystem dynamics and evaluating the likelihood of delisting the system as an AOC. Quite recently, a network of models was developed to connect the watershed processes with the dynamics of the Bay of Quinte (Zhang et al. 2013; Arhonditsis et al. 2016; Kim et al. 2013, 2016, 2017). This integrated watershed-receiving water body modeling framework has been used to evaluate management scenarios that would lead to significant reduction of phosphorus export from the Bay of Quinte watershed and to quantify the overall uncertainty associated with the severity of the eutrophication phenomena in the area (Fig. 11.10).



**Fig. 11.10** Conceptual diagram of the integrated phosphorus-modeling framework for the Bay of Quinte. The spatial segmentation of the model for the receiving water body consists of the following compartments: (*U1*) the segment that extends from the mouth of Trent River until the city of Belleville; (*U2*) the segment that begins from the mouth of Moira River and comprises the Big Bay, Muscote Bay, and North Point Bay; and (*U3*) the area influenced by the inflows of Napanee River, extending until the outlet of Hay Bay. In the middle Bay, there are three segments corresponding to the main stem (*M1*) and the two adjacent embayments: Hay Bay (*M2*), and Picton Bay (*M3*). The lower segment of the Bay, representing the transitional area to Lake Ontario, was separated into the epilimnetic (*Le*) and hypolimnetic (*Lh*) compartments. *Numbers in parentheses* correspond to the average flushing rate of each segment [Reproduced from Arhonditsis et al. (2016)]



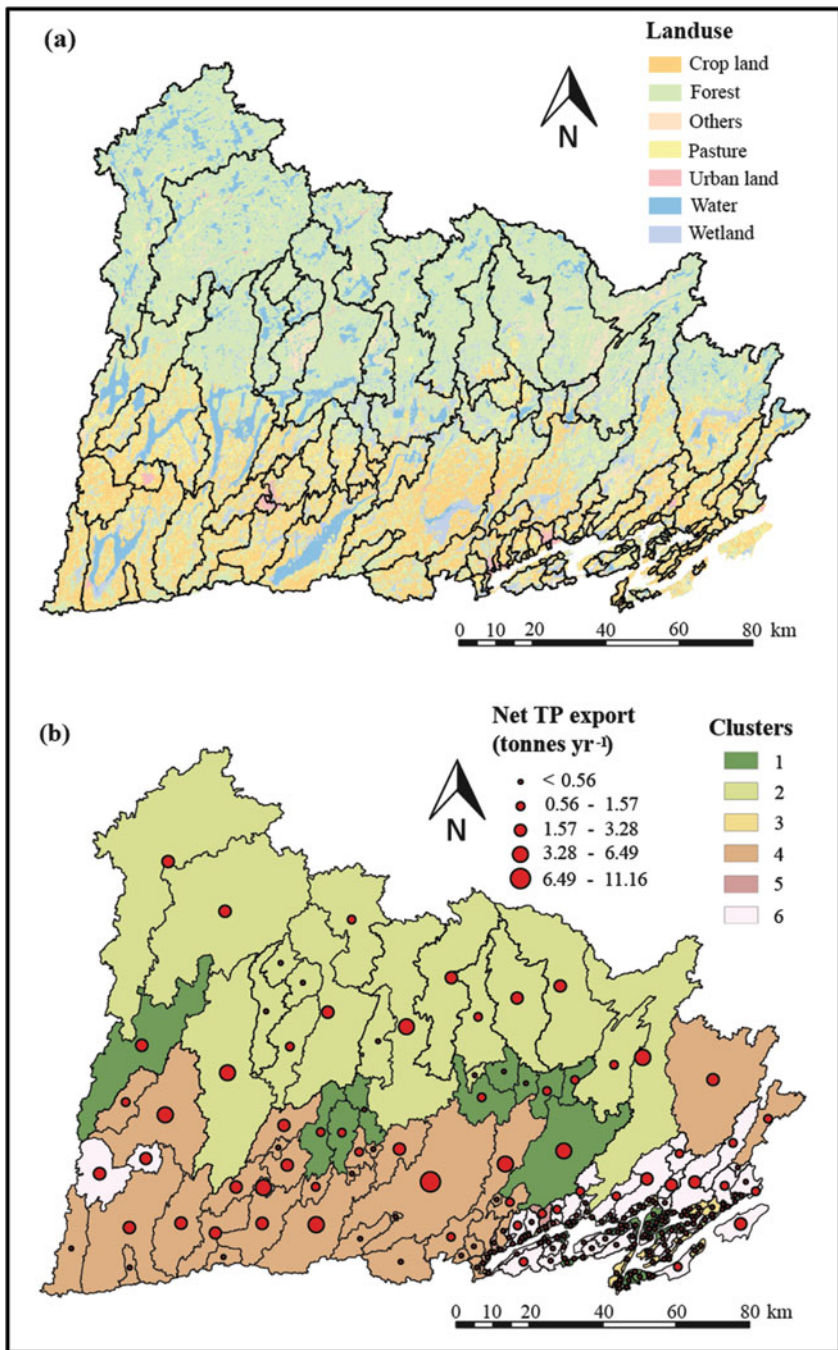
### ***11.3.2 Modeling the Relationship Among Watershed Physiography, Land Use Patterns, and Phosphorus Loading***

One of the emerging imperatives of eutrophication management is the advancement of our understanding of the relationships among land use, agricultural activities, hydrological processes, and water quality (Wellen et al. 2015). Prior to the watershed modeling exercise, Kim et al. (2016) implemented Self-Organizing Maps (SOM) to gain insights into the physiographical features and land-use patterns in the Bay of Quinte watershed, and to subsequently associate them with the phosphorus non-point source loading. In this application, eighteen classification variables were used, such as the landscape slope, saturated soil hydraulic conductivity, soil bulk density, and areal fractions for different land use types (lakes, ponds, alvars, bogs, coniferous swamps, deciduous swamps, fens, marshes, deciduous forests, coniferous forests, cutovers, mining areas, urban lands, pastures, and croplands) in 73 gauged and 137 ungauged subwatersheds. Thus, a total of 210 spatial units were distributed on 2-dimensional hexagonal maps, and then clustered in different groups according to their similarities.

Based on the spatial heterogeneity of these classification variables, SOM delineated six spatial clusters in the Bay of Quinte watershed with fairly distinct land-use patterns (Fig. 11.11). Coniferous and deciduous coverage along with pastures and croplands dominate the landscape in cluster 1. Different types of wetlands, such as fen ( $\approx 10\%$ ), coniferous swamp ( $\approx 8\%$ ), and alvar ( $\approx 0.4\%$ ) have also their highest areal fraction values in the same cluster. In cluster 2, the average landscape slope is steep and the soil bulk density is high. The areal fractions of forests as well as mining and logging sites are also high. In cluster 3, most of the subwatersheds are located in the vicinity of the Bay of Quinte, where crops occupy  $\approx 75\%$  of the area. Not surprisingly, the annual TP yield per area and average TP concentrations are the highest ( $528 \text{ kg km}^{-2} \text{ year}^{-1}$  and  $103 \mu\text{g L}^{-1}$ ) in these same regions. In cluster 4, soil hydraulic conductivity is significantly higher, deciduous swamp are more abundant relative to the rest of the watershed, cropland coverage is the second highest ( $\approx 41\%$ ), and thus the net TP export is high. In cluster 5, urban land represents  $\approx 74\%$  of the land-use coverage and net TP export and yield are the second highest ( $3.72 \text{ tonnes year}^{-1}$  and  $209 \text{ kg km}^{-2} \text{ year}^{-1}$ ), which is further accentuated by the increased point source loading ( $2.44 \text{ tonnes year}^{-1}$ ). In cluster 6, pasture and cropland approximately correspond to 60% of the area, and these subwatersheds are mainly located adjacent to the Bay of Quinte.

Nutrient loads, yields, and deliveries at landscape and regional scales were estimated using the SPARROW model (Kim et al. 2017). The goodness-of-fit between observed and predicted TP loading values from the SPARROW model was excellent in the logarithmic scale ( $r^2 > 0.95$ ), although there were four sites with errors greater than  $10 \text{ tonnes year}^{-1}$  when the SPARROW predictions were back-transformed to the original scale. The posterior parameter values offered insights into the patterns of phosphorus export and delivery in the Bay of Quinte

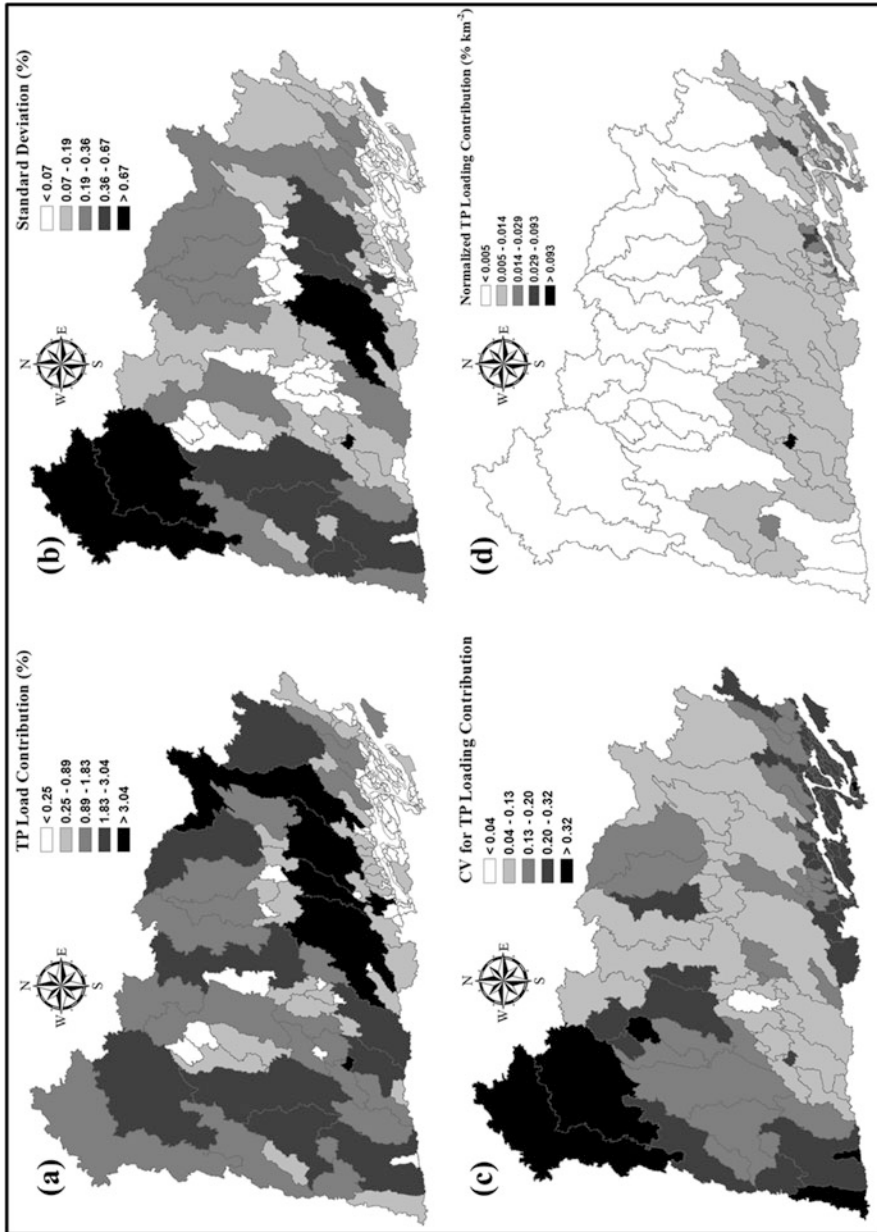




**Fig. 11.11** Map of Bay of Quinte watershed: (a) land use types, and (b) classification based on artificial neural networks and associated phosphorus export per subwatershed [Reproduced from Kim et al. (2016)]

watershed. The main findings from the SPARROW modeling exercise were as follows: (1) urban areas are characterized by a fairly high areal phosphorus export with a mean estimate of 126 kg of TP per km<sup>2</sup> on an annual basis; (2) the contribution of phosphorus from agricultural land uses can vary considerably among the various crop types (30–127 TP kg per km<sup>2</sup>), but is generally lower than the impact of urban sites. Similar to the Hamilton Harbour, this finding contradicts the popular notion that rates of nutrient export from urban lands are below those of agricultural lands due to lower anthropogenic nutrient subsidies, such as fertilizer implementation (Moore et al. 2004; Soldat et al. 2009). Nonetheless, other studies in the region of Southern Ontario have found urban total phosphorus export rates to be comparable (or even higher) than agricultural total phosphorus export rates (Winter and Duthie 2000); (3) the crop-specific export coefficient values were on par with those typically reported in the literature (Harmel et al. 2008); (4) the attenuation rate in low flow streams (3.7% of TP per kilometer) appears to be distinctly greater than in those with high flow (1.1% of TP per kilometer); and (5) fallow areas are responsible for approximately 70 kg of TP per km<sup>2</sup> on an annual basis.

In the context of watershed management, the spatial distribution of net (instead of the cumulative) TP loading that ultimately inflows into the receiving waterbody was used to identify the most influential subwatersheds (Fig. 11.12). The percentage of net loading was mostly greater in the downstream catchment of the major tributaries. By contrast, the relative contribution of the ungauged watersheds close to the bay was significantly lower primarily due to their small areal extent (Fig. 11.12a). On the other hand, the error associated with the estimates of the relative contribution of the different subwatersheds was higher in the Trent River basin (SE > 67%) than the rest of the tributaries. Interestingly, the Trent River's upper catchment also exhibited high variability in the percentage net TP loads (Fig. 11.12b). The coefficient of variation (CV) values of the relative contributions along with the net contributions normalized by the corresponding subwatershed areas were also used to delineate the hot-spots in the Bay of Quinte watershed. The highest CVs (>32%) were found in the upper catchment of Trent River (Fig. 11.12c). Counter to the error estimates, however, the ungauged watershed close to the bay was characterized by fairly high CVs (Fig. 11.12c). This trend was more pronounced when the normalized percentage TP loads were considered (Fig. 11.12d). Unlike the CV values, the normalized percentage TP loads were low in the upper catchment of Trent River, but were distinctly higher in the lower part of the watershed, especially near the bay (Fig. 11.12d). Overall, this strategy pinpointed many locations close to the water body that may be responsible for significant nutrient fluxes, due to their landscape attributes and soil characteristics (Kim et al. 2017).

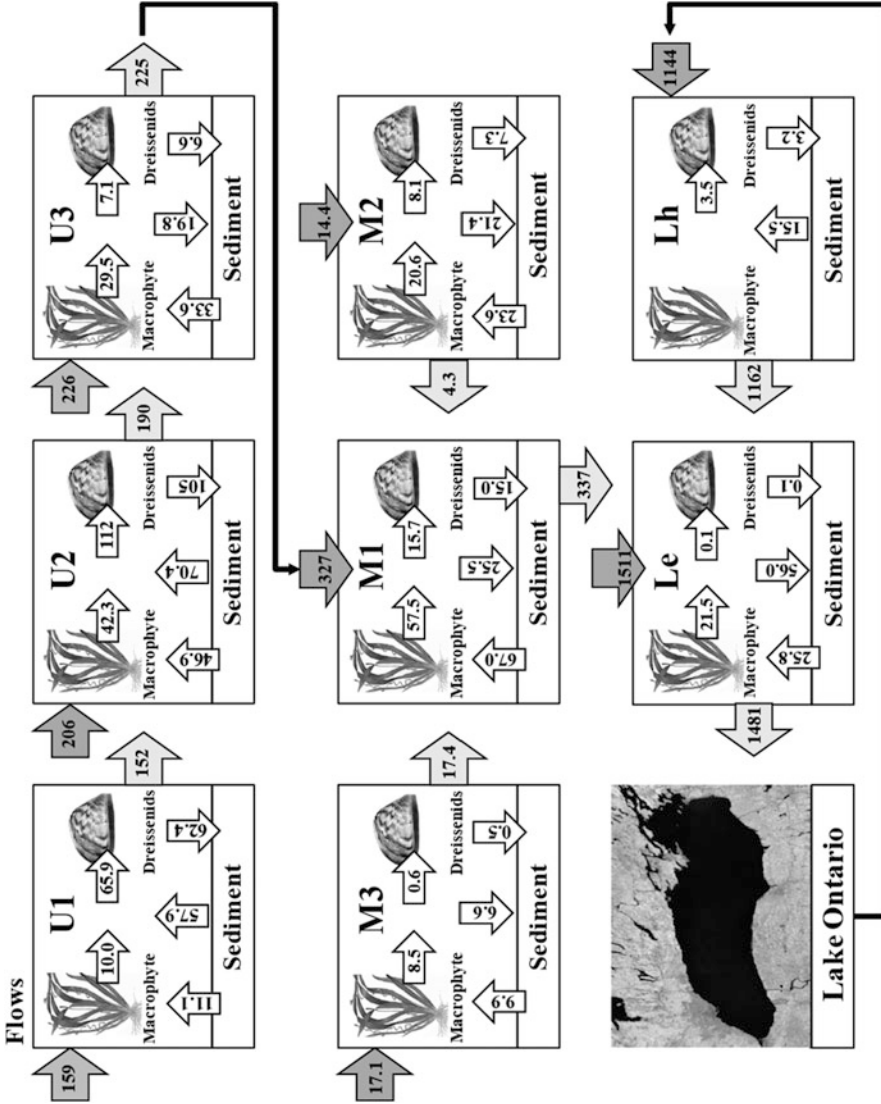


**Fig. 11.12** Percentage contribution of the annual net TP loads to the Bay of Quinte: (a) average prediction, (b) standard error (SE) and (c) coefficient of variation of the corresponding predictions, and (d) average prediction normalized by the subwatershed areas [Reproduced from Kim et al. (2017)]

### ***11.3.3 Eutrophication Risk Assessment with Process-Based Modeling and Determination of Water Quality Criteria***

The basis of the eutrophication risk assessment analysis was the mechanistic model presented by Kim et al. (2013), which introduced several novel mathematical formulations regarding the representation of macrophyte dynamics; the role of dreissenids in the system; several processes related to the fate and transport of phosphorus in the sediments along with the interplay between water column and sediments, such as particulate sedimentation being dependent upon the standing algal biomass, sediment resuspension, sorption/desorption in the sediment particles, and organic matter decomposition. The model was then calibrated to match the measured TP concentrations in the upper, middle, and lower segments of the Bay during the 2002–2009 period (Kim et al. 2013; Arhonditsis et al. 2016). The model demonstrated satisfactory ability to fit the monthly TP levels in the Bay of Quinte, and was able to reproduce the end-of-summer increase of the ambient TP levels in the upper segment, even in years (e.g., 2005) when the corresponding concentrations were greater than  $60 \mu\text{g L}^{-1}$ . The model also faithfully depicted the spatial gradients in the system, with distinctly higher TP levels in the upper segment relative to those experienced in the middle/lower Bay (Kim et al. 2013).

The model was then used to draw inferences on the spatial variability of the various external and internal TP flux rates in the Bay of Quinte (Fig. 11.13). The net TP contributions (sources or sinks) represent the mass of phosphorus associated with the various compartments (water column, sediments, macrophytes, dreissenids) throughout the growing season (May–October) averaged over the 2002–2009 period. In the U1 segment, the phosphorus budget is predominantly driven by the external sources (phosphorus loading:  $159 \text{ kg day}^{-1}$ ) and sinks (outflows:  $152 \text{ kg day}^{-1}$ ). The sediments (resuspension and diffusion from the sediments to water column minus particle settling) act as a net source of phosphorus in this segment ( $57.9 \text{ kg day}^{-1}$ ). Dreissenids subtract approximately  $65.9 \text{ kg day}^{-1}$  from the water column (particle filtration minus respiration) and subsequently deposit  $62.4 \text{ kg day}^{-1}$  via their excretion and particle rejection. In a similar manner, the U2 segment receives  $206 \text{ kg day}^{-1}$  from exogenous sources, including the upstream inflows, and transports downstream  $190 \text{ kg day}^{-1}$ . The net contribution of the sediments accounted for  $70.4 \text{ kg day}^{-1}$ , while dreissenids on average reduce the ambient TP levels by  $112 \text{ kg day}^{-1}$ . The main differences between the two segments in the upper Bay are the TP fluxes related to macrophyte P intake from the sediments and respiration that can reach the levels of  $46.9$  and  $42.3 \text{ kg day}^{-1}$  relative to the fluxes of  $11.1$  and  $10.0 \text{ kg day}^{-1}$  in the U1 segment. Likewise, the macrophyte intake from the sediments minus the amount of P regenerated from the decomposition of the dead plant tissues varies between  $35$  and  $65 \text{ kg day}^{-1}$  in segments U3 and M1, while the subsequent release of their metabolic by-products is approximately responsible for  $19$ – $26 \text{ kg day}^{-1}$ . The settling of particulate P dominates over the resuspension and diffusion from the sediments to the water column



**Fig. 11.13** Spatial variability of the various external and internal TP flux rates ( $\text{kg day}^{-1}$ ) in the Bay of Quinte. *Arrow directions* indicate the net contribution (sources or sinks) of the various compartments (water column, sediments, macrophytes, dreissenids). *Dark gray arrows* show the TP inflows in a spatial segment, while the *light-gray arrows* depict the corresponding outflows [Reproduced from Arhonditsis et al. (2016)]

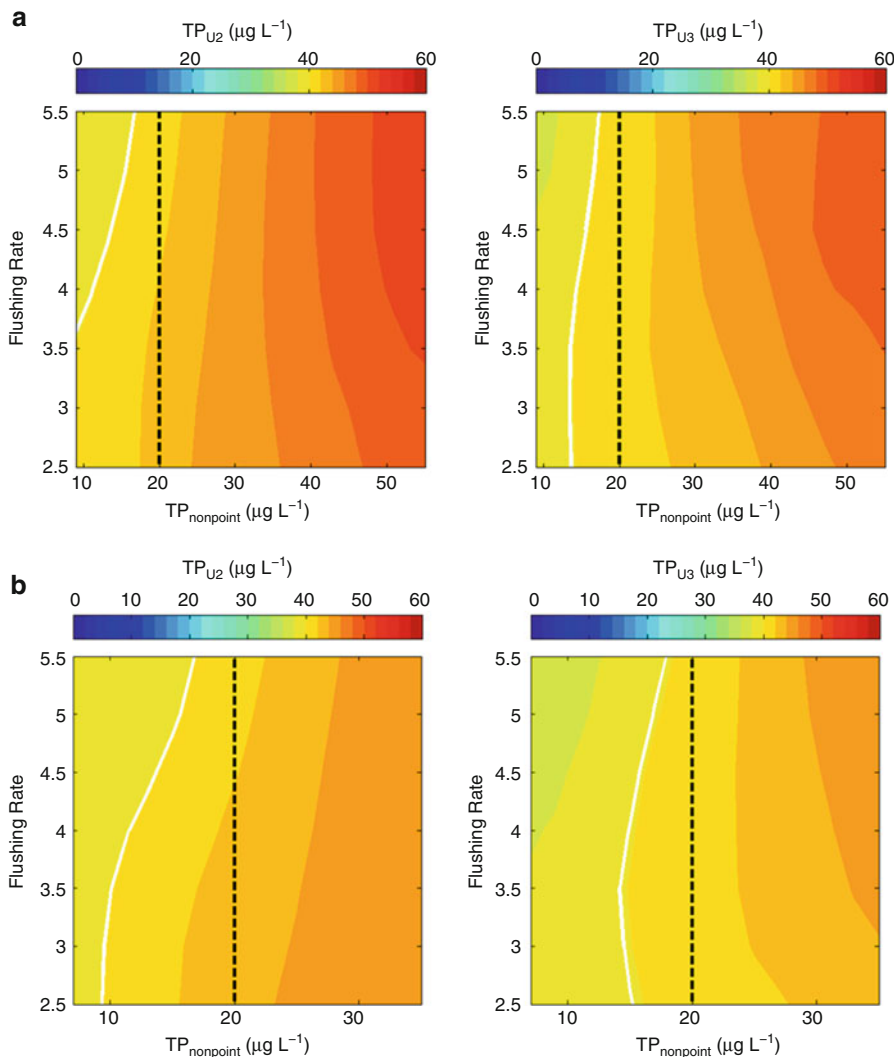


with the corresponding net fluxes ranging between 25 and 35 kg day<sup>-1</sup>. In Hay Bay (M2), the fluxes mediated by the macrophytes and dreissenids primarily modulate the TP dynamics and the same pattern appears to hold true in Picton Bay (M3). In the lower Bay of Quinte (Le and Lh), the model postulates a significant pathway (>1100 kg P day<sup>-1</sup>) through which the inflowing water masses from Lake Ontario well up from the hypolimnion to the epilimnion and are subsequently exported from the system. In the same area, the internal biotic sources (macrophytes) similarly represent an important vector of phosphorus transport.

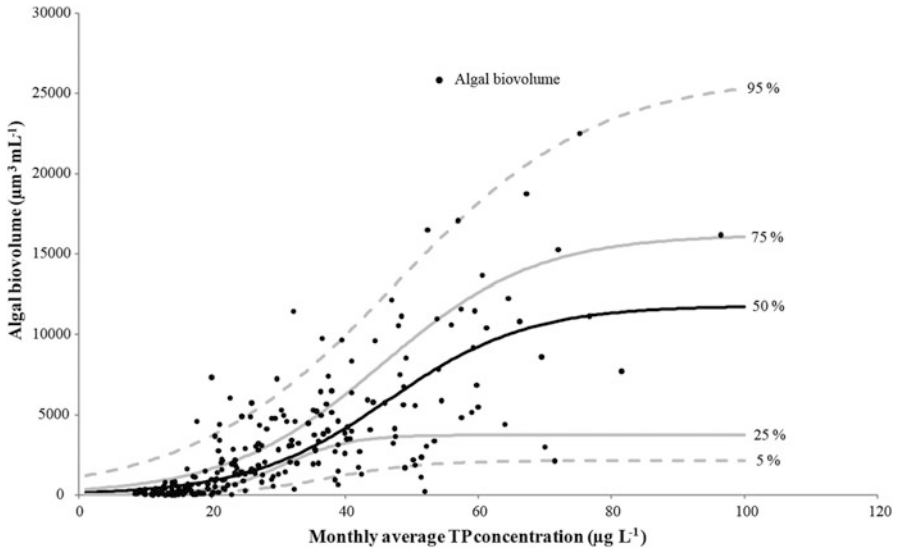
In general, the Bay of Quinte modeling work highlights the internal recycling as one of the key drivers of phosphorus dynamics. The flow from the Trent River is the predominant driver of the dynamics in the upper segment until the main stem of the middle area. However, the sediments in the same segment release a significant amount of phosphorus and the corresponding fluxes are likely amplified by the macrophyte and dreissenid activity. From a management standpoint, the presence of a significant positive feedback loop in the upper Bay of Quinte suggests that the anticipated benefits of additional reductions of the exogenous point and non-point loading may not be realized within a reasonable time frame, i.e., 5–10 years (Kim et al. 2013). Analysis of nutrient loading scenarios showed that the restoration pace of the Bay could be slow, even if the riverine total phosphorus concentrations reach levels significantly lower than their contemporary values, <25 µg TP L<sup>-1</sup> (Fig. 11.14; see also Kim et al. 2013; Arhonditsis et al. 2016).

Bearing in mind that the TP targeted levels merely represent a “means to an end” and not “the end itself”, the actual question that the stakeholders in the area ponder is to what extent the anticipated benefits from a more efficient external phosphorus loading control could also be capitalized as a significant decrease of the algal bloom frequency? With respect to the total phytoplankton biovolume, Nicholls et al. (2002) showed that it declined after the control of phosphorus in the 1970s, but did not change significantly after the establishment of dreissenids in the system. As previously mentioned, Nicholls and Carney (2011) showed that the arrival of dreissenid mussels may be associated with positive (e.g., *Aphanizomenon* and *Anabaena* decline) effects on the integrity of the Bay of Quinte ecosystem. However, the recent increase of the cyanophyte *Microcystis* has had significant implications for the aesthetics and other beneficial uses of the Bay of Quinte, through the formation of “scums” on the water surface as well as the fact that some strains of *Microcystis* are toxin producers. These structural shifts in the phytoplankton community composition could stem directly from the feeding selectivity of dreissenids or indirectly from the improvements in the transparency of the water column (Blukacz-Richards and Koops 2012), but the role of the feedback loop associated with their nutrient recycling activity could conceivably be another important factor.

According to the predictions of a non-linear quantile regression model (Shimoda et al. 2016), the current average TP concentrations (30–40 µg L<sup>-1</sup>) represent the area where the algal biovolume vs TP relationship is characterized by a steep slope and thus any further improvements in the ambient nutrient levels are likely to induce more favorable quantitative and qualitative changes in phytoplankton (Fig. 11.15). Nonetheless, existing empirical evidence from the system is indicative



**Fig. 11.14** Simulated maximum TP concentrations during the growing season (May–October) in the Bay of Quinte. *Upper panels (a)* refer to the predictions associated with the reference environmental conditions; and *lower panels (b)* represent the predictions of a TP loading reduction scenario (60% point sources, 20% non-point sources, and 50% urban storm water). The first eleven years (2002–2012) were based on real meteorological and nutrient loading conditions, while the final (12th) year was forced with a wide range of combinations of TP riverine concentrations and flows that were generated from the mean ( $\pm$  error) predictions of the SPARROW model. The *white contour line* corresponds to the proposed targeted level of  $40 \mu\text{g TP L}^{-1}$ . The flushing rates express the frequency (number of times) of water renewal in the upper Bay during the growing season. The *black dotted line* represents a threshold level of  $20 \mu\text{g L}^{-1}$  for the flow-weighted TP concentration in all the major tributaries in the upper Bay of Quinte [Reproduced from Arhonditsis et al. (2016)]



**Fig. 11.15** Quantile regression model for total phytoplankton biovolume against monthly average TP concentration in the Bay of Quinte (Arhonditsis et al. 2016)

of a weak correlation between chlorophyll *a* and cyanobacteria toxin concentrations (Watson et al. 2011), suggesting that a complex interplay among physical, chemical, and biological factors may drive the spatiotemporal abundance and composition patterns of the algal assemblages in the Bay of Quinte (Nicholls et al. 2002). In a system like the Bay of Quinte, where both external and internal loading drives the severity of eutrophication phenomena, there will inevitably be some uncertainty in the overall eutrophication risk assessment.

There are several compelling reasons (knowledge gaps, natural variability, complex interactions among a suite of ecological mechanisms) to avoid overly confident statements about the future response of this impaired system, and thus the most prudent strategy is to explicitly recognize an acceptable level of violations of the delisting goals. Specifically, Kim et al. (2013) challenged the usefulness of the historical delisting criterion of a seasonal average TP concentration lower than  $30 \mu\text{g L}^{-1}$ , as it is neither a reflection of the considerable intra-annual variability in the upper Bay nor representative of the water quality conditions in near shore areas of high public exposure (e.g., beaches). It would seem very unlikely that a single-value water quality standard monitored in a few offshore sampling stations can capture the entire range of dynamics in the system (e.g., the extremes seen in the near shore sites) or the magnitude of the end-of-summer TP peaks. Kim et al. (2013) instead advocated the pragmatic stance that the delisting objectives should revolve around extreme (and not average) values of variables of management interest and must explicitly accommodate all the sources of uncertainty (insufficient information, lack of knowledge, and natural variability) by permitting a realistic frequency of standard violations. Namely, the critical threshold level should be set at a value



of  $40 \mu\text{g TP L}^{-1}$ , which cannot be exceeded more than 10–15% in both time and space. Under the assumption that the TP concentrations in the Bay of Quinte follow a log-normal distribution and that TP values  $<15 \mu\text{g L}^{-1}$  are likely to occur only 10% of the time during the growing season, then 10–15% exceedances of the  $40 \mu\text{g TP L}^{-1}$  level are approximately equivalent to a targeted seasonal average of 25–28  $\mu\text{g TP L}^{-1}$ . Thus, the replacement of the historical paradigm (binary assessment) with a probabilistic approach to water quality criteria does not intend to make the delisting of AOCs easier, but rather to offer a more comprehensive method for tracking the prevailing conditions in the Bay.

## 11.4 Concluding Remarks

We have demonstrated some of the benefits for environmental management when identifying the uncertainties and knowledge gaps of the natural environment, differentiating between predictable and unpredictable patterns, and critically evaluating model outputs. The presentation of the model outputs as a probabilistic assessment of environmental conditions makes the model results more credible for local decision makers and stakeholders. The often-misleading deterministic statements are avoided and environmental goals are set by explicitly acknowledging an inevitable risk of not achieving 100% compliance in time and space. The acceptable level of violations is then subject to decisions that reflect different socioeconomic values and environmental priorities.

The Bayesian (iterative) nature of the presented modeling networks is conceptually similar to the policy practice of adaptive management, i.e., an iterative implementation strategy that is recommended to address the often-substantial uncertainty associated with water quality model forecasts and avoid the implementation of inefficient and flawed management plans. The use of Bayesian inference techniques is also consistent with the scientific process of progressive learning and offers a natural mechanism for sequentially updating our knowledge on model inputs and structure every time new data are collected from the system. Thus, modeling tools can be iteratively updated to accommodate the significant year-to-year variability associated with the external nutrient loading or the weather conditions, thereby serving as a reliable long-term management tool for policy analysis. Importantly, the probabilistic statements provided from the Bayesian calibration can also indicate where the limited monitoring resources should be focused (Zhang and Arhonditsis 2008). In particular, additional data collection efforts should target hot spots, where the model predictive distribution indicates a high probability of non-attaining water quality goals or, alternatively, an *unacceptably high* variance. Thus, we can assess the value of information (value of additional monitoring; “Where should additional data collection efforts be focused?”) and subsequently optimize the sampling design for environmental monitoring. In other words, uncertainty does matter and its quantification is not an excuse to avoid providing answers

to pressing environmental problems, but rather a prudent strategy to improve the rigor of model-based management of our natural resources!

**Acknowledgements** George Arhonditsis wishes to acknowledge the continuous support of his work on model uncertainty analysis from the National Sciences and Engineering Research Council of Canada (Discovery Grants). The Hamilton Harbour modeling project has received funding support from the Ontario Ministry of the Environment (Canada-Ontario Grant Agreement 120808). The Bay of Quinte modeling project was undertaken with the financial support of the Lower Trent Region Conservation Authority provided through the Bay of Quinte Remedial Action Plan Restoration Council.

## References

- Alexander RB, Smith RA, Schwarz GE (2004) Estimates of diffuse phosphorus sources in surface waters of the United States using a spatially referenced watershed model. *Water Sci Technol* 49:1–10
- Arhonditsis GB, Adams-Van Harn BA, Nielsen L et al (2006) Evaluation of the current state of mechanistic aquatic biogeochemical modeling: citation analysis and future perspectives. *Environ Sci Technol* 40:6547–6554
- Arhonditsis GB, Qian SS, Stow CA et al (2007) Eutrophication risk assessment using Bayesian calibration of process-based models: application to a mesotrophic lake. *Ecol Model* 208:215–229
- Arhonditsis GB, Papantou D, Zhang W et al (2008a) Bayesian calibration of mechanistic aquatic biogeochemical models and benefits for environmental management. *J Marine Syst* 73:8–30
- Arhonditsis GB, Perhar G, Zhang W et al (2008b) Addressing equifinality and uncertainty in eutrophication models. *Water Resour Res* 44:W01420
- Arhonditsis GB, Kim D-K, Shimoda Y et al (2016) Integration of best management practices in the Bay of Quinte watershed with the phosphorus dynamics in the receiving water body: What do the models predict? *Aquat Ecosyst Health Manage* 19:1–18
- Basu NB, Rao PSC, Thompson SE et al (2011) Spatiotemporal averaging of in-stream solute removal dynamics. *Water Resour Res* 47:W00J06
- Bayarri MJ, Berger JO, Cafeo J et al (2007) Computer model validation with functional output. *Ann Stat* 35:1874–1906
- Beck ME (1987) Tectonic rotations on the leading edge of South America: the Bolivian orocline revisited. *Geology* 15:806–808
- Beven K (1993) Prophecy, reality and uncertainty in distributed hydrological modelling. *Adv Water Resour* 16:41–51
- Beven K (2006) A manifesto for the equifinality thesis. *J Hydrol* 320:18–36
- Blukacz-Richards EA, Koops MA (2012) An integrated approach to identifying ecosystem recovery targets: application to the Bay of Quinte. *Aquat Ecosyst Health Manage* 15:464–472
- Charlton MN (2001) The Hamilton Harbour remedial action plan: eutrophication. *Verh Internat Verein Limnol* 27:4069–4072
- Claessens L, Tague CL, Band LE et al (2009) Hydro-ecological linkages in urbanizing watersheds: an empirical assessment of in-stream nitrate loss and evidence of saturation kinetics. *J Geophys Res Biogeosci* 114:G04016
- Dawes RM (1988) Rational choice in an uncertain world. Harcourt Brace Jovanovich, San Diego
- Dermott R, Bonnell R (2011) Benthic fauna in the Bay of Quinte. Bay of Quinte remedial action plan: Monitoring Report #20, Kingston, ON, pp 51–71
- deYoung B, Barange M, Beaugrand G et al (2008) Regime shifts in marine ecosystems: detection, prediction and management. *Trends Ecol Evol* 23:402–409

- Dietzel A, Reichert P (2012) Calibration of computationally demanding and structurally uncertain models with an application to a lake water quality model. *Environ Modell Softw* 38:129–146
- Donner SD, Kucharik CJ, Oppenheimer M (2004) The influence of climate on in-stream removal of nitrogen. *Geophys Res Lett* 31:L20509
- Doyle MW, Stanley EH, Harbor JM (2003) Hydrogeomorphic controls on phosphorus retention in streams. *Water Resour Res* 39:1147
- Edwards AM, Yool A (2000) The role of higher predation in plankton population models. *J Plankton Res* 22:1085–1112
- Gudimov A, Stremilov S, Ramin M, Arhonditsis GB (2010) Eutrophication risk assessment in Hamilton Harbour: system analysis and evaluation of nutrient loading scenarios. *J Great Lakes Res* 36:520–539
- Gudimov A, Ramin M, Labencki T et al (2011) Predicting the response of Hamilton Harbour to the nutrient loading reductions: a modeling analysis of the “ecological unknowns”. *J Great Lakes Res* 37:494–506
- HH RAP (2003) Hamilton Harbour Remedial Action Plan, Report Stage 2 Update. Hamilton Harbour Technical Team. Burlington, ON
- Hall JD, O’Connor K, Ranieri J (2006) Progress toward delisting a Great Lakes Area of Concern: the role of integrated research and monitoring in the Hamilton Harbour Remedial Action Plan. *Environ Monit Assess* 113:227–243
- Hall JD, O’Connor KM (2016) Hamilton Harbour remedial action plan process: connecting science to management decisions. *Aquat Ecosyst Health Manage* 19:107–113
- Harmel D, Qian S, Reckhow K, Casebolt P (2008) The MANAGE database: nutrient load and site characteristic updates and runoff concentration data. *J Environ Qual* 37:2403–2406
- Hiriart-Baer VP, Milne J, Charlton MN (2009) Water quality trends in Hamilton Harbour: two decades of change in nutrients and chlorophyll a. *J Great Lakes Res* 35:293–301
- Hiriart-Baer VP, Boyd D, Long T et al (2016) Hamilton Harbour over the last 25 years: insights from a long-term comprehensive water quality monitoring program. *Aquat Ecosyst Health Manage* 19:124–133
- Kim D-K, Zhang W, Rao Y et al (2013) Improving the representation of internal nutrient recycling with phosphorus mass balance models: a case study in the Bay of Quinte, Ontario, Canada. *Ecol Model* 256:53–68
- Kim D-K, Zhang W, Watson S, Arhonditsis GB (2014) A commentary on the modelling of the causal linkages among nutrient loading, harmful algal blooms, and hypoxia patterns in Lake Erie. *J Great Lakes Res* 40:117–129
- Kim D-K, Kaluskar S, Mugalingam S, Arhonditsis GB (2016) Evaluating the relationships between watershed physiography, land use patterns, and phosphorus loading in the Bay of Quinte, Ontario, Canada. *J Great Lakes Res* 42:972–984
- Kim D-K, Kaluskar S, Mugalingam S et al (2017) A Bayesian approach for estimating phosphorus export and delivery rates with the SPATIally Referenced Regression On Watershed attributes (SPARROW) model. *Ecol Inform* 37:77–91
- Kinstler P, Morley A (2011) Point source phosphorus loadings 1965 to 2009. Bay of Quinte remedial action plan: monitoring report #20. Kingston, ON, pp 15–17
- Leisti KE, Doka SE, Minns CK (2012) Submerged aquatic vegetation in the Bay of Quinte: Response to decreased phosphorous loading and Zebra Mussel invasion. *Aquat Ecosyst Health Manage* 15:442–452
- Long T, Wellen C, Arhonditsis G, Boyd D (2014) Evaluation of stormwater and snowmelt inputs, land use and seasonality on nutrient dynamics in the watersheds of Hamilton Harbour, Ontario, Canada. *J Great Lakes Res* 40:964–979
- Long T, Wellen C, Arhonditsis G et al (2015) Estimation of tributary total phosphorus loads to Hamilton Harbour, Ontario, Canada, using a series of regression equations. *J Great Lakes Res* 41:780–793

- McDowell RW, Srinivasan MS (2009) Identifying critical source areas for water quality: 2. Validating the approach for phosphorus and sediment losses in grazed headwater catchments. *J Hydrol* 379:68–80
- McMahon G, Alexander RB, Qian S (2003) Support of total maximum daily load programs using spatially referenced regression models. *J Water Res* 129:315–329
- Moore RB, Johnson CM, Robinson KW, Deacon JR (2004) Estimation of total nitrogen and phosphorus in New England streams using spatially referenced regression models. US Department of the Interior, US Geological Survey, New Hampshire, p 42
- Morgan MG, Henrion M, Small M (1992) *Uncertainty: a guide to dealing with uncertainty in quantitative risk and policy analysis*. Cambridge University Press, New York
- Nicholls KH, Heintsch L, Carney E (2002) Univariate step-trend and multivariate assessments of the apparent effects of P loading reductions and zebra mussels on the phytoplankton of the Bay of Quinte, Lake Ontario. *J Great Lakes Res* 28:15–31
- Nicholls KH, Carney EC (2011) The phytoplankton of the Bay of Quinte, 1972–2008: point-source phosphorus loading control, dreissenid mussel establishment, and a proposed community reference. *Aquat Ecosyst Health Manage* 14:33–43
- Pappenberger F, Beven KJ (2006) Ignorance is bliss: or seven reasons not to use uncertainty analysis. *Water Resour Res* 42:W05302
- Ramin M, Stremilov S, Labencki T et al (2011) Integration of numerical modeling and Bayesian analysis for setting water quality criteria in Hamilton Harbour, Ontario, Canada. *Environ Modell Softw* 26:337–353
- Ramin M, Labencki T, Boyd D et al (2012) A Bayesian synthesis of predictions from different models for setting water quality criteria. *Ecol Model* 242:127–145
- Reichert P, Omlin M (1997) On the usefulness of overparameterized ecological models. *Ecol Model* 95:289–299
- Reichert P, Schuwirth N (2012) Linking statistical description of bias to multi-objective model calibration. *Water Resour Res* 48:W09543
- Rode M, Arhonditsis G, Balin D (2010) New challenges in integrated water quality modelling. *Hydrol Process* 24:3447–3461
- Schwarz GE, Hoos AB, Alexander RB, Smith RA (2006) *The SPARROW surface water-quality model: theory, application and user documentation*. U.S. Geological Survey Techniques and Methods Report, Book 6, Chapter B3; USGS [https://pubs.usgs.gov/tm/2006/tm6b3/PDF/tm6b3\\_part1a.pdf](https://pubs.usgs.gov/tm/2006/tm6b3/PDF/tm6b3_part1a.pdf)
- Shimoda Y, Watson S, Palmer ME (2016) Delineation of the role of nutrient variability and dreissenids (Mollusca, Bivalvia) on phytoplankton dynamics in the Bay of Quinte, Ontario, Canada. *Harmful Algae* 55:121–136
- Soldat DJ, Petrovic AM, Ketterings QM (2009) Effect of soil phosphorus levels on phosphorus runoff concentrations from turfgrass. *Water Air Soil Pollut* 199:33–44
- Stow CA, Reckhow KH, Qian SS (2007) Approaches to evaluate water quality model parameter uncertainty for adaptive TMDL implementation. *JAWRA* 43:1499–1507
- Watson SB, Borisko J, Lalor J (2011) Bay of Quinte harmful algal bloom programme phase I – 2009. Bay of Quinte remedial action plan: monitoring report #20. Kingston, ON, pp 27–50
- Wellen C, Arhonditsis GB, Labencki T, Boyd D (2012) A Bayesian methodological framework or accommodating interannual variability of nutrient loading with the SPARROW model. *Water Resour Res* 48:W10505
- Wellen C, Arhonditsis GB, Labencki T, Boyd D (2014a) Application of the SPARROW model in watersheds with limited information: a Bayesian assessment of the model uncertainty and the value of additional monitoring. *Hydrol Process* 28:1260–1283
- Wellen C, Arhonditsis GB, Long T, Boyd D (2014b) Accommodating environmental thresholds and extreme events in hydrological models: a Bayesian approach. *J Great Lakes Res* 40:102–116

- Wellen C, Arhonditsis GB, Long T, Boyd D (2014c) Quantifying the uncertainty of nonpoint source attribution in distributed water quality models: a Bayesian assessment of SWAT's sediment export predictions. *J Hydrol* 519:3353–3368
- Wellen C, Kamran-Disfani A-R, Arhonditsis GB (2015) Evaluation of the current state of distributed watershed nutrient water quality modeling. *Environ Sci Technol* 49:3278–3290
- Winter JG, Duthie HC (2000) Export coefficient modeling to assess phosphorus loading in an urban watershed. *JAWRA* 36:1053–1061
- Yerubandi RR, Boegman L, Bolkhari H, Hiriart-Baer V (2016) Physical processes affecting water quality in Hamilton Harbour. *Aquat Ecosyst Health Manage* 19:114–123
- Zhang W, Arhonditsis GB (2008) Predicting the frequency of water quality standard violations using Bayesian calibration of eutrophication models. *J Great Lakes Res* 34:698–720
- Zhang W, Kim D-K, Rao Y et al (2013) Can simple phosphorus mass balance models guide management decision? A case study in the Bay of Quinte, Ontario, Canada. *Ecol Model* 257:66–79

# Chapter 12

## Multivariate Data Analysis by Means of Self-Organizing Maps

Young-Seuk Park, Tae-Soo Chon, Mi-Jung Bae, Dong-Hwan Kim,  
and Sovan Lek

**Abstract** Ecological data range widely in variability, showing non-linear and complex relationships among variables. Although conventional multivariate analyses are useful tools to explore ecological data, data mining by non-linear methods is preferred because a high degree of complexity resides in ecological phenomena. One of these methods is artificial neural networks in machine learning based on biologically inspired learning algorithms. Self-organizing map (SOM) is one of the most popular unsupervised artificial neural networks and are commonly used to seek patterns and clusters in ecological data. SOMs are versatile in analysing non-linear and complex data, which are observed frequently in ecological systems. In this paper, we explain the theory of SOMs and their application in ecological modelling, with a focus on learning processes, visualization, preprocessing of input data, and interpretation of results. We also discuss the advantages and disadvantages of SOM approaches.

### 12.1 Introduction

Ecological data are complex both spatially and temporally. They range widely in variability, showing non-linear and complex relationships between explanatory and response variables, mixed with noise, redundancy, and outliers (Gauch 1982;

---

Y.-S. Park (✉) • D.-H. Kim  
Kyung Hee University, Seoul, Republic of Korea  
e-mail: [parkys@khu.ac.kr](mailto:parkys@khu.ac.kr)

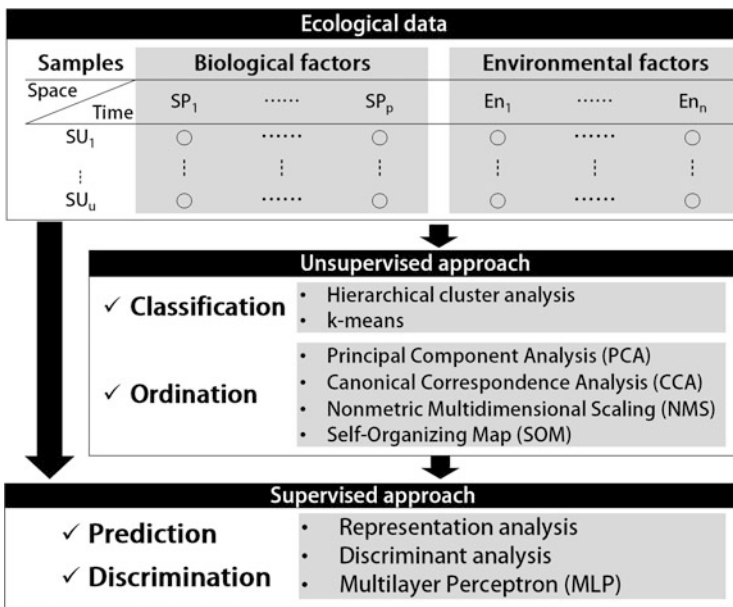
T.-S. Chon  
Pusan National University, Busan, Republic of Korea  
e-mail: [tschon.chon@gmail.com](mailto:tschon.chon@gmail.com)

M.-J. Bae  
Nakdonggang National Institute of Biological Resources, Gyeongsangbuk-do, Republic of Korea

S. Lek  
Université Paul Sabatier, Toulouse, France  
e-mail: [sovanarath.lek@univ-tlse3.fr](mailto:sovanarath.lek@univ-tlse3.fr)

Jongman et al. 1995; Park et al. 2003). Explaining variation in complex ecological data through either statistical analyses or machine learning can be considered in two steps: unsupervised and supervised approaches (Fig. 12.1). Although various ways exist to extract information from datasets, unsupervised approaches are used conventionally to summarize the variability in the data as a first step using statistical methods, including classification (i.e. hierarchical cluster analysis, k-means, etc.) or ordination (i.e. principal component analysis (PCA), nonmetric multidimensional scaling (NMDS), detrended correspondence analysis (DCA), isometric feature mapping (Isomap), etc.) as well as machine learning algorithms, including a self-organizing map (SOM) (Park et al. 2003). Subsequently, this data analysis is followed by supervised approaches in either statistical (e.g. regression analysis and discriminant analysis) or learning algorithms (e.g. a multilayer perceptron) (Fig. 12.1). The supervised approaches are helpful in investigating more specific questions in the later phases of data analysis after information is initially extracted from the original data.

PCA is an indirect gradient analysis method, seeking the strongest linear correlation structure among variables (Legendre and Legendre 1998). It reduces multi-dimensional data to lower dimensions that keep the characteristics of the raw data as much as possible. Eigen values, which explain a portion of the original total variance, are calculated, and then eigenvectors, which contain the coefficients of



**Fig. 12.1** Schematic diagram of the modelling procedure. Unsupervised approaches are used to summarize the properties of given data sets in the first step, and then supervised approaches are used for the prediction and discrimination of variables in revealing input–output relationships. The arrows represent a direct relationship between modelling steps

the linear equation for a given axis, are founded. Finally, each axis score using the eigenvector is shown in an ordination space (Bae et al. 2008).

DCA was developed to correct the distortions that occur in correspondence analysis (Hill and Gauch 1980). The first dimension is split into several intervals and the second axis scores are adjusted in order to make mean score within each segment zero. As individual segments of each axis are expanded or contracted, the within-sample variation of species scores is equalized (McCune and Grace 2002; Bae et al. 2008).

NMDS maintains the rank ordering of the distances in a low dimensional space, expressed as a monotonic function (Shepard 1962; Borg and Groenen 1997; Mahechaa et al. 2007). It calculates the best position of the data on reduced dimensions through an iterative search that minimizes the stress of the reduced dimensions. “Stress” is a measure of departure from monotonicity in the relationship between the dissimilarity distance in the original dimensional space and distance in the reduced dimensional ordination space. The value of stress based on Kruskal’s rules of thumb is between 0 and 100 (Daniel and Scott 2007). If the value is close to 0, we can conclude NMDS result is appropriate to use (Bae et al. 2008).

Isomap (Tenenbaum et al. 2000) is an algorithm for the ordination, combining the classical techniques of PCA and NMDS to a class of nonlinear manifolds (Mahechaa et al. 2007). The algorithm is based on a nonlinear geodesic inter-point distance matrix. Isomap defines residual variance to characterize how well the low-dimensional Euclidean embedding captures the geodesic distances estimated from the neighborhood graph. Lower residuals indicate better-fitting solutions, with less metric distortion (Balasubramanian et al. 2002; Bae et al. 2008).

Bae et al. (2008) compared these four different ordination methods (i.e. PCA, DCA, NMDS, and Isomap) for patterning water quality of reservoirs, and concluded that PCA and NMDS appeared to be the most efficient methods based on the explanation power. Although conventional multivariate analyses are useful tools to explore ecological data, data mining by non-linear methods is preferred because a high degree of complexity resides in ecological phenomena (Blayo and Demartines 1991). SOMs are unsupervised artificial neural networks (ANNs) that allow non-linear data mining by means of biologically inspired learning algorithms, and are applicable to classification and association (Park et al. 2003; Chon et al. 2004). SOMs identify patterns and clusters in data effectively (Fig. 12.1), and visualize properties of a dataset. By contrast, supervised ANNs reveal input-output relationships within complex data that can be applied for predictive modelling (e.g. Recknagel et al. 1997). In supervised ANNs, a ‘teacher’ in the learning phase ‘tells’ the ANN how well it performs or what the correct behaviour should be. The most popular supervised ANN is a multi-layer perceptron with a back-propagation algorithm, which proves to be efficient for prediction and discrimination problems. In this chapter, we focus on the theory and application of SOMs in ecological modelling.



## 12.2 Properties of a Self-Organizing Map

SOMs were proposed by Kohonen (1982, 2001) in the early 1980s and are also known as Kohonen networks or Kohonen feature maps. An SOM approximates the probability density function of the input data and is used in clustering, visualization, and abstraction (Kohonen 2001). The algorithm performs a topology-preserving projection of the data space onto a regular low-dimensional space. Theoretically, there is no limitation in the dimensional space, but usually a two-dimensional space is preferred because of the ability of human perception. The SOM puts the dataset on the map preserving the neighbourhood, so similar patterns in the dataset are mapped close together on the grid.

In the learning process, the SOM calculates the distance between the samples and virtual computational units (details of this process are provided in following sections). This distance is influenced by the properties of the input datasets. Therefore, the final output stems from the input dataset. If the data have high variability due to undesired sources (e.g., noise), output data variation will not be represented properly in the reduced dimension. A preprocessing step is needed prior to SOM training to overcome the problems due to undesired sources such as missing values, outliers, and extremes. In addition, data transformation may be needed for handling extreme values, different data distribution patterns, periodic properties, etc. Therefore, we first consider the issues with data preparation.

## 12.3 Data Preparation

### 12.3.1 *Missing Values and Outliers*

Missing values can be treated in three ways: deletion, skipping, or replacement. (1) The simplest way to treat missing values is to delete any row or column of the data matrix containing missing values. However, this is the most costly method because valuable information present in the data can be lost when it is removed along with the missing values. (2) Missing values can be skipped during the numerical calculation by recording them in the data matrix. Conventionally, we use 'NaN' for 'not available' or '-9999' for values that are impossible to observe in the actual data. (3) Estimated values can be used to replace the missing values, including the mean, median, values obtained by regression or prediction models, and interpolated values by autocorrelation. This method is usually the most suitable when missing values are scattered over the data matrix.

Outliers are recorded values of measurements in variables that are outside the range of the bulk of the data (Ellison and Gotelli 2004; Osborne and Overbay 2004). They may be noise, but they may also reflect actual ecological processes. Therefore, they should be carefully considered.

### 12.3.2 Data Transformation

Transformations are conducted as a preprocessing procedure to meet the assumptions of applying statistical approaches if there are extreme differences or different distribution patterns among measured values for variables.

*Logarithmic transformation:* Logarithmic transformation may be used for variables with a high degree of variation. This type of transformation is one of the most commonly used in ecological studies. It compresses high values while spreading low values by expressing the values as orders of magnitude. To avoid the problem of  $\log(0)$  being undefined, a value of one is added to all data points before applying the transformation.

*Standardization:* Standardization, also called variance normalization, is a linear transformation that scales values with a mean = 0 and variance = 1, thus making the data dimensionless. It is useful for studying environmental variables in ecological studies. Then, the transformed variables can be compared to each other conveniently. Both positive and negative values are produced by the transformation, so it is not compatible with proportion-based distance measures such as a Sorensen distance (McCune and Grace 2002).

*Range normalization:* Range normalization is a linear transformation that scales the values between 0 and 1. This transformation is used to provide the same weights to different variables by rescaling. For example, in multivariate analysis of communities, abundant species contribute more than rare species do. To avoid this, species abundance can be transformed by rescaling the range between the minimum and maximum.

*Binary transformation:* Both quantitative and qualitative data can be converted to binary data (i.e. either 0 or 1). Quantitative data can be transformed to binary according to a threshold value, which can be the mean, median, or other values. If a value is less than or equal to the threshold value, it is designated as zero, otherwise it becomes a one. When several different threshold values are used, ordinal data is ranked. In addition, qualitative variables can be transformed into binary as dummy variables. This transformation allows the use of qualitative descriptors in multivariate analyses (Legendre and Legendre 1998).

Details on data transformation are available in statistical analysis and data mining books, including McCune and Grace (2002) and Legendre and Legendre (1998).

### 12.3.3 Distance Measure

The first step in the multivariate analysis is the calculation of distances or similarities among a set of samples. Euclidean distance (ED) and Bray-Curtis dissimilarity are most commonly used for distance measures in SOMs for ecological studies.

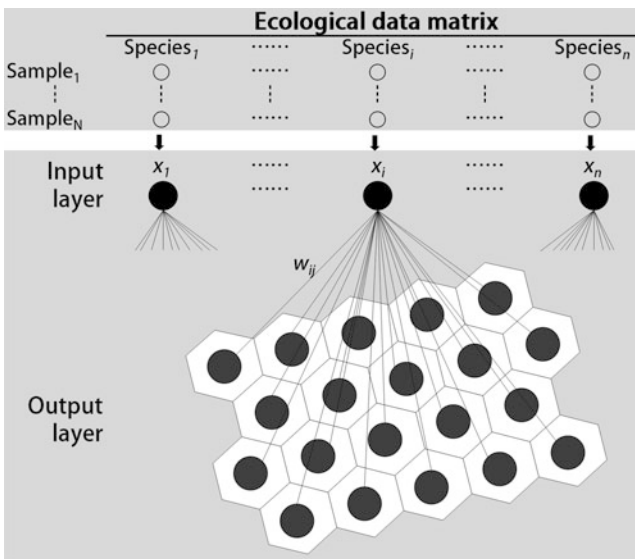
Euclidean distance ranges from zero to infinity. Euclidean distance is highly sensitive to outliers, because large differences are weighted more heavily than several small differences. Euclidean distance also suffers from double-zero problems (i.e. zero values observed at two samples concurrently), which are commonly observed in ecological community data. Due to these reasons, Bray-Curtis dissimilarity is used in ecological community studies, ranging between 0 and 1.

There are other distance measures including variants of ED such as squared ED and relative ED, relative Sorensen distance, Jaccard distance, Chi-square distance, Mahalanobis distance, and City-block (Manhattan) distance. They are well documented in the literature, including in Legendre and Legendre (1998) and in McCune and Grace (2002).

## 12.4 Self-Organizing Maps

### 12.4.1 Architecture

An SOM consists of an input layer and an output layer (Fig. 12.2). Each layer has a collection of computational units called neurons, or computation nodes. The input layer is connected to each vector of the dataset, and each variable in the data matrix corresponds to each input unit. The output layer consisting of  $D$  output units forms an array of units on which the distribution of the dataset is represented in an ordered



**Fig. 12.2** A two-dimensional SOM. The data matrix is given in the input layer, whereas patterned results are presented in the output layer

way in a low dimension. Input and output layers are connected by the connection intensities (weights) represented in the reference vectors.

### 12.4.2 Learning Algorithm

When an input vector  $\mathbf{x}$  of a sample is sent to the input layer as stated above, the distance between the weight vector  $\mathbf{w}$  and the  $\mathbf{x}$  is computed adaptively at all output units. At the beginning of the learning process, the  $\mathbf{w}$  is initialized with small random values. The distances between  $\mathbf{x}$  and all output units are calculated, and the output unit with the minimum distance is defined as the best matching unit (BMU), or so-called ‘winner’, of the given input vector. As with other clustering algorithms, many different kinds of distance measure algorithms can be applied as stated in Sect. 12.3.3.

The  $\mathbf{w}$  of the BMU and its neighbourhood units are updated by the SOM learning rule as follows:

$$\mathbf{w}_c(t+1) = \mathbf{w}_c(t) + \alpha(t)h_{cv}(t)(\mathbf{w}_c(t) - \mathbf{x}(t)) \quad (12.1)$$

where  $t$  is the iteration step,  $\mathbf{w}_c$  is the weight vector of BMU  $c$ ,  $\alpha(t)$  is the learning rate which is a decreasing function of the iteration time  $t$ , and  $h_{cv}(t)$  is the neighbourhood function that defines the distance between the neighbourhood  $v$  and the BMU  $c$  to be updated during the learning process.

The shape of the width of the neighbourhood can be chosen with various neighbourhood functions including block, Gaussian bell, triangular, and Mexican hat shaped (Kohonen 2001). Triangular and Gaussian bell shaped functions may lead to smoother formation of topology in the map and faster convergence of the weight vectors. The Mexican hat function is useful when constructing an SOM for classification purposes (Melssen et al. 1994).

This learning process is continued until a stopping criterion is met, usually, when weight vectors are stabilized or when a predefined number of iterations are completed. For good statistical accuracy, Kohonen (2001) recommends that the number of iterations must be at least 500 times the number of network units, whereas the number of input variables generally does not affect the number of iterations. The sequential training procedure is summarized in Box 12.1.

#### Box 12.1 Sequential Learning Algorithm of an SOM

Step 1. Initialize weights,  $\mathbf{w}$ , to small random values

Step 2. Present an input vector  $\mathbf{x}$

Step 3. Compute the distance between the weight vectors and the input vector

(continued)

**Box 12.1** (continued)

Step 4. Determine the best matching unit (BMU) with minimum distance for the input vector

Step 5. Determine neighbourhood whose distance to the BMU on the map of the network is less than or equal to neighbourhood radius  $r$

Step 6. Update weights  $w$ ; Learning rate and neighbourhood radius are decreased with time as convergence is reached.

Step 7. Go to Step 2 and repeat the process for all input vectors until a stopping criterion is met.

### 12.4.3 Evaluation of Trained Map Quality

After the learning process, the quality of trained results from the SOM can be evaluated. Among several map quality measures, *Quantization error* (QE) and *Topographic error* (TE) are commonly used. QE is calculated using the average distance between each data vector and its BMU to measure map resolution, whereas TE displays the proportion of all data vectors for which first and second BMUs are not adjacent to measure topology preservation (Kiviluoto 1996). Therefore, TE is used as an indicator of the accuracy of the mapping in preserving the topology, whereas QE is used to select the best map with the minimum value (Kohonen 2001; Park et al. 2003).

### 12.4.4 Optimum Map Size

Although a two-dimensional map in output layer is the most popular in practical applications, maps with one or higher numbers of dimensions may be applied. One dimension may be used specifically for emphasizing single dimensional data variation in output; however, data interpretation and visualization are generally not very appealing due to the dimensional limit, compared to two dimensions. On the other hand, a higher number of dimensions, such as three, can provide more detailed information on data variation while still maintaining comprehensibility. Maps with high numbers of dimensions, however, require a vast number of computations before convergence of the network. Moreover, it is rather cumbersome in a practical sense to visualize the output of the high dimensional map and to interpret the output dimension by dimension.

In a two-dimensional map, rectangular and hexagonal configurations are commonly used. However, a hexagonal lattice is preferred, because it does not favour horizontal or vertical directions as much as a rectangular array does (Kohonen 2001; Park et al. 2003).

The map size (number of output neurons) is important to detect any deviation in the data. If the map size is too small, it might not explain some important differences that should be detected. Conversely, if the map size is too big, the differences are too small (Wilppu 1997). The map size depends on the number of samples to be trained. Although no strict rules exist to define the optimal map size, there are several possible methods. First, setting the number of output neurons approximately equal to the number of the input samples seems to be a useful rule-of-thumb for many applications when the data sets are relatively small (Kaski 1997). However, attention should be paid to over-fitting problems when a large map size is used. This may happen when the number of output units is as large as or larger than the number of training samples.

The second method to determine the map size is by using the heuristic rule suggested by Vesanto (1999). According to their rule, the number of output neurons is determined as  $5 \times \sqrt{\text{number of samples}}$  which is defined in the SOM Toolbox (<http://www.cis.hut.fi/projects/somtoolbox/>). In this case, the two largest eigenvalues of the training data are calculated first, and then the ratio between the side lengths of the map grid is set to the ratio between the two maximum eigenvalues (C er ghino and Park 2009).

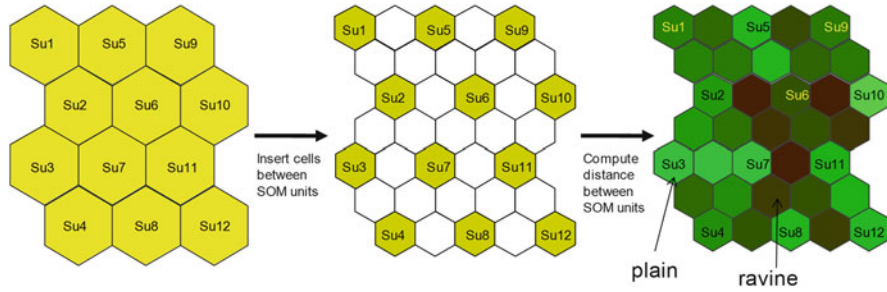
An alternative to the calculation of eigenvalues is to consider QE and TE, which indicate the quality of the trained map (Park et al. 2003). The optimum size is selected based on minimum values for QE and TE. However, it should be noted that QE and TE gradually decrease with increasing map size, so the optimum size is based on local minimum values for QE and TE.

### 12.4.5 Clustering SOM Units

The trained SOM ordines samples on the SOM output units based on the similarity of the input variables. However, it is difficult to distinguish subsets of SOM output units because there are still no boundaries among the SOM output units. Therefore, a further step is required to split the SOM map into several groups according to the similarity of input variables. Several different clustering algorithms can be allied.

First, the *unified distance matrix algorithm (U-matrix)*; Ultsch and Siemon (1990) is popular for presenting overall similarities among SOM units. Inserted between computation nodes, the U-matrix calculates distances between neighbouring map units, and these distances can be visualized to represent clusters using a grey scale display on the map (Fig. 12.3). Low distance is visualized as plain, whereas high distance is ravine, indicating possible clusters.

Alternatively, *hierarchical clustering analysis* is commonly used because it presents hierarchical similarities using a dendrogram among SOM output units based on linkage distances as criteria. However, both linkage methods and distance measures affect the clustering results.



**Fig. 12.3** Procedure for U-matrix visualization. New cells are inserted between SOM units, and then distances between the SOM units are calculated. The calculated distances in the U-matrix are visualized. Low distance is presented as plain, whereas high distance appears as ravine

A *k-means method* can also be applied. To determine the best number of clusters, the *Davies–Bouldin index* (DBI) (Davies and Bouldin 1979) is calculated. The DBI is a relative index of cluster validity, with smaller DBIs indicating better clustering. Small values of the DBI occur for a solution with low variance within clusters and high variance between clusters. Therefore, minimal DBI proposes the best number of clusters (Hruschka and Natter 1999). Moreover, DBI can be applied to the results of hierarchical clustering to determine the optimum number of clusters.

#### 12.4.6 Evaluation of Input Variables

During the learning process of an SOM, the SOM output units that are topographically close in the array activate each other to update their weights from the same input vector. This results in a smoothing effect on the weight vectors. These weight vectors tend to approximate the probability density function of the input vector. Therefore, the visualization of input variables is convenient to understand the contribution of each input variable with respect to the clusters on the trained SOM. This visualization map is called *component planes*.

*Indicator species analysis* (Dufrêne and Legendre 1997) is used to evaluate the contribution of input variables (indicator species) in each cluster defined in the SOM. Indicator species are used as ecological indicators of community or habitat types, environmental conditions, or environmental changes. An indicator value (IndVal) for each species  $i$  in cluster  $j$  is calculated as follows:

$$\text{IndVal}_{i,j} = (N_{\text{individuals}_{i,j}}/N_{\text{individuals}_i}) \times (N_{\text{sites}_{i,j}}/N_{\text{sites}_j}) \times 100 \quad (12.2)$$

where  $N_{\text{individuals}_{i,j}}$  is the mean number of individuals of species  $i$  across sites of cluster  $j$ ,  $N_{\text{individuals}_i}$  is the sum of the mean numbers of individuals of species  $i$  over all groups,  $N_{\text{sites}_{i,j}}$  is the number of sites in cluster  $j$  where species  $i$  is present, and  $N_{\text{sites}_j}$  is the total number of sites in cluster  $j$ .

De Cáceres et al. (2010) suggested improving indicator species analysis by considering all possible combinations of groups of sites and selecting the combination for which the species can be used best as an indicator. A package ‘*indicspecies*’ in R for indicator species is available from the CRAN (<https://cran.r-project.org/web/packages/indicspecies/>).

### ***12.4.7 Relations Between Biological and Environmental Variables***

It is necessary to understand the relationships between biological and environmental variables since natural distributions of organisms are determined primarily by their environment (Huntley 1999). To understand these relationships, environmental variables can be projected onto the SOM that has been trained with biological variables. At first, each environmental variable is submitted to the trained SOM, and then the mean value of each environmental variable is calculated in each output neuron of the trained SOM with samples belonging to the same neuron (Park et al. 2003). These mean values of environmental variables assigned on the SOM are visualized in grey scale as component planes, and then compared with maps of sampling sites as well as biological attributes.

## **12.5 Application in Ecological Modelling**

Since Chon et al. (1996) applied an SOM to the patterning of benthic communities in streams, SOMs have been used most widely in extracting information from ecological data: community grouping (Foody 1999; Giraudel and Lek 2001; Park et al. 2003), hydrosystems/landscapes (Tison et al. 2004), animal behaviours (Chon et al. 2004; Park et al. 2005), plankton community dynamics (Recknagel et al. 2006), cyanotoxins dynamics (Chan et al. 2007) and patterning of long-term fisheries data (Hyun et al. 2005). SOMs were also applied to natural resource and ecosystem management (Park et al. 2003, 2004, 2006, Gevrey et al. 2004, Park and Chung 2006), prediction of population and communities (Céréghino et al. 2001; Obach et al. 2001), dimensional reduction of large datasets (Park et al. 2006; Griebeler and Seitz 2006), computational policy simulations for natural hazard migrations (Samarasinghe and Strickert 2013), surface temperature anomaly and solar activity (Friedel 2012), spatial and temporal variations of benzene (Strebel et al. 2013), effects of landscape and morphometric factors on water quality of reservoirs (Park et al. 2014), and molecular ecology (Roux et al. 2007; Nikolic et al. 2009). Several papers were published in special issues of Ecological Informatics (Chon and Park 2006; Park and Chon 2015) and Ecological Modelling (Park and



Chon 2007). Kalteh et al. (2008) and Chon (2011) reviewed the applications of the SOM techniques in ecological and environmental sciences.

## 12.6 SOM Tools

The most efficient and popular SOM tool is the SOM Toolbox (Vesanto 1999) operated in MATLAB, which was developed by the Laboratory of Information and Computer Science at the Helsinki University of Technology. The SOM Toolbox is available freely from their website (<http://www.cis.hut.fi/projects/somtoolbox/>), and it provides default optimized initialization and training methods (Vesanto 1999).

Three R packages implementing standard SOMs are available from the CRAN (<https://cran.r-project.org/web/packages/>): *kohonen* (Wehrens 2015), *som*, and *wccsom*. Recently, Bottin et al. (2014) developed a package ‘*diatSOM*’ for R. Although it was specifically suited to diatom communities, it can be applied to other ecological data. The *diatSOM* package is available from the authors by request.

## 12.7 Example of SOM Application

### Problem

1. Summarize variability of aquatic insect species richness in streams
2. Characterize the relationship between species richness and environmental gradients

### Dataset

An SOM was applied to classify the sampling sites according to similarity of aquatic insect richness with a focus on four insect orders [i.e. Ephemeroptera, Plecoptera, Trichoptera, and Coleoptera (EPTC)] collected at 138 sampling sites (Park et al. 2003). EPTC richness is highly correlated to the overall macro-invertebrate richness in the study area (Céréghino et al. 2001). Therefore, it is a good estimator of the overall community richness.

EPTC richness was characterized using four environmental variables: altitude (m), stream order, distance from the source (km), and maximum water temperature (°C) in summer. A detailed description of these ecological data was also given in Cayrou et al. (2000) and Céréghino et al. (2001). The data were transformed to be proportionally normalized between 0 and 1 in the range of the minimum and maximum values (i.e. *range normalization*) in each taxon (each input variable).

### Training Network

The network consisted of an input layer with four input neurons (four taxa richness) and an output layer on a two-dimensional hexagonal lattice. The dataset, consisting of 138 samples with four taxa richness, was trained with a sequential learning algorithm. We used 54 ( $= 9 \times 6$ ) output units based on  $5 \times \sqrt{138}$  samples.

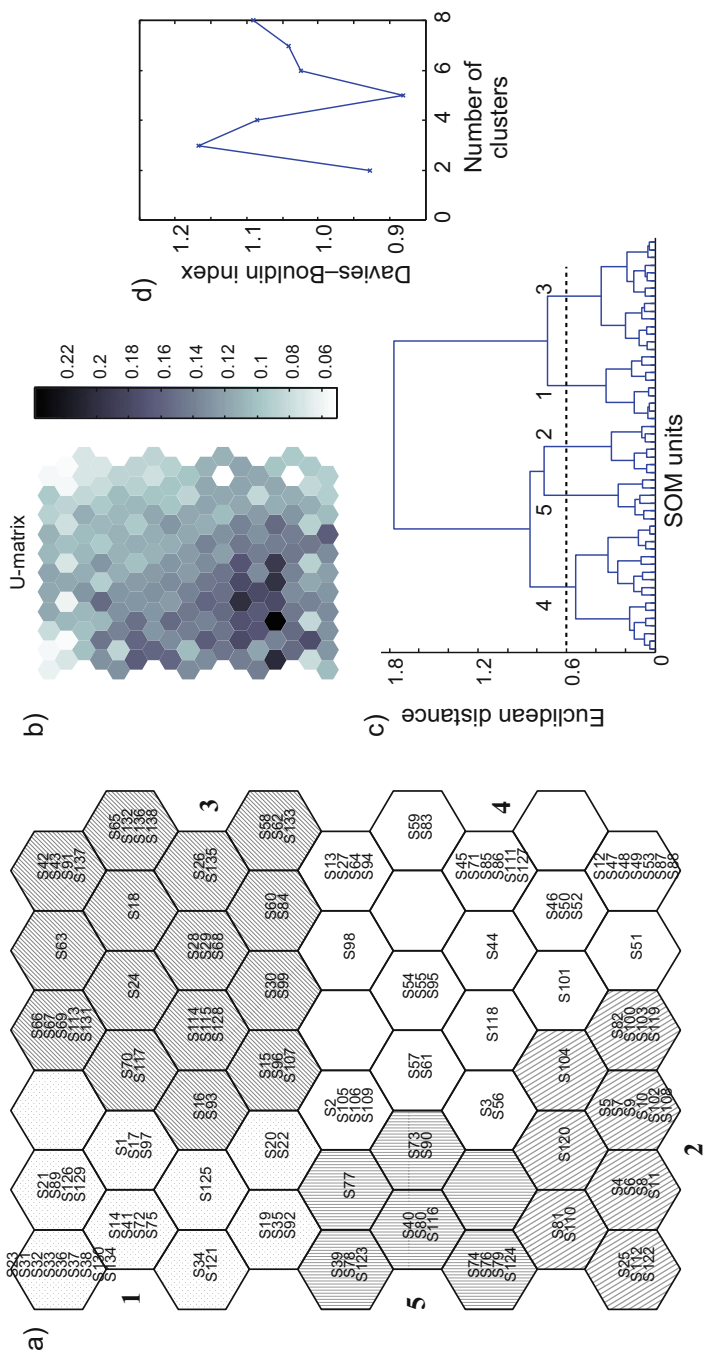
### Results

The SOM training converged after 8832 learning iterations. The final QE and final TE were 0.167 and 0.022, respectively. This result showed that only three pairs ( $= 0.022 \times 138$ ) of the first- and second-BMUs were not adjacent in the trained hexagonal map, therefore the SOM was trained smoothly in topology.

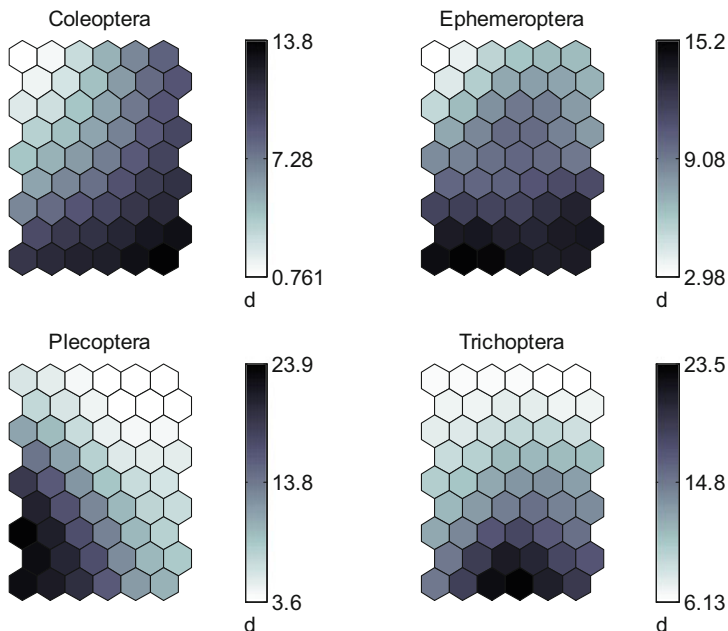
The trained SOM classified samples based on the similarities of EPTC richness are shown in Fig. 12.4. The SOM output units were classified further into five clusters (1–5) based on a hierarchical cluster analysis using the Ward linkage method with Euclidean distance measures. The DBI was the lowest with five clusters. The U-matrix showed similar patterns. Therefore, three different methods proposed five clusters as the optimum number in the network. Based on a dendrogram of hierarchical clustering and the DBI, the five clusters were parsed into two main groups: clusters 1 and 3 and clusters 2, 4, and 5.

Species richness was higher on the lower parts of the SOM units, but lower on the upper parts (Fig. 12.5). Samples in clusters 1 and 3 have mostly low species richness. In particular, the species richness of Coleoptera and Ephemeroptera were lowest in cluster 1, whereas that of Plecoptera was lower in cluster 3. Trichoptera richness was low in both clusters. Meanwhile, species richness of Coleoptera and Ephemeroptera was the highest in clusters 3 and 4, whereas that of Plecoptera was highest in cluster 2. Figure 12.6 shows the differences in EPTC richness of the five different clusters. Clusters 1 and 3 on the upper parts of the SOM map have low species richness, and clusters 2 and 5 on the lower part of the map have the highest values. These results indicate that occurrence patterns of EPTC still show variability to some degree, even though they inhabit mostly undisturbed areas.

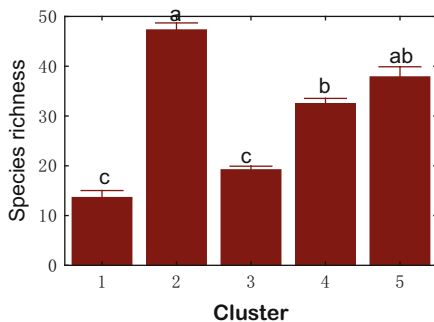
Samples in clusters 3 and 4 were from streams with high stream order (lower in the watershed), whereas samples in clusters 1 and 5 were mainly from small streams with low stream order (higher up in the watershed) (Fig. 12.7). The variation in maximum temperature in summer was low compared to that of the other variables. The distance from the source was clearly different between these clusters. In contrast, concerning altitude, clusters 1 and 5 represented high mountain streams, whereas clusters 3 and 4 were for lowland streams. Finally, these results display that EPTC richness was lower in the streams with higher stream order, but higher in the streams with average and lower stream orders. The relationship between species richness and environmental variables is summarized in Table 12.1.



**Fig. 12.4** Classification of samples on the trained SOM: (a) the U-matrix, (b) hierarchical cluster analysis, (c) Davies-Bouldin index, and (d) k-means method applied to set boundaries on the SOM map



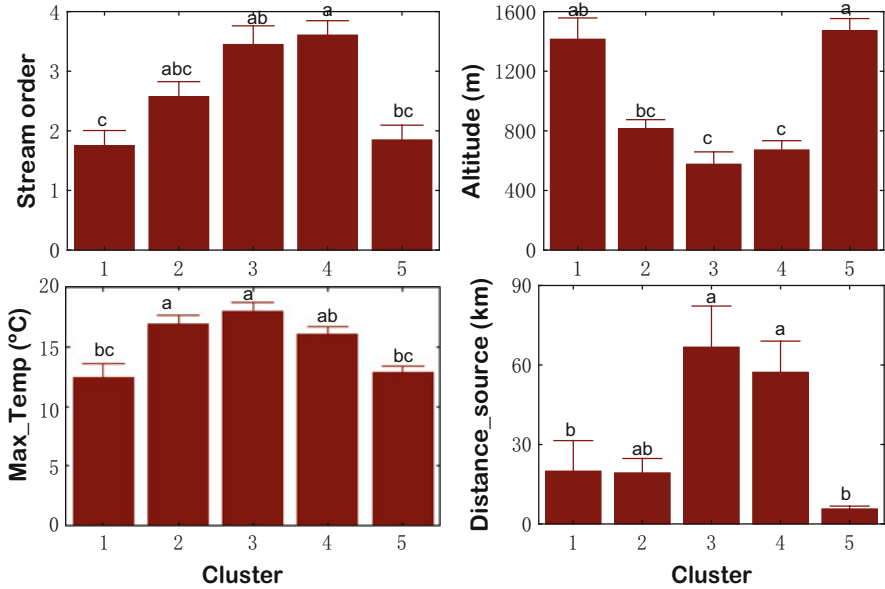
**Fig. 12.5** Visualization of EPTC richness calculated in the trained SOM. The values of EPTC richness were calculated during the learning process



**Fig. 12.6** Differences in EPTC richness for the five different clusters. The *different letters on the bars* indicate statistically significant differences based on Dunn’s multiple comparison test ( $P < 0.05$ )

## 12.8 Advantages and Disadvantages

An SOM is an efficient means of data mining to approximate the probability density function of complex input data with fewer dimensions (Kohonen 2001). Inspired by neural activity in biological organisms, the ‘winner’ has the chance to adjust its weight through local competition; and global convergence was reached adaptively



**Fig. 12.7** Differences in four environmental variables [stream order, altitude, maximum temperature in summer (Max\_temp), and distance from source (Distance\_source)] for the five different clusters defined in the SOM. The *different letters on the bars* indicate statistically significant differences based on Dunn’s multiple comparison test ( $P < 0.05$ )

**Table 12.1** Summary of the relationship between species richness and environmental variables

Variable	Cluster				
	1	2	3	4	5
EPTC richness	L	H	L	M	H
Stream order	L	M	H	H	L
Altitude	H	M	L	L	H
Max_temp <sup>a</sup>	L	H	H	H	L
Distance_source <sup>b</sup>	L	L	H	H	L
Overall	Low species richness at high mountain area	High species richness upper streams	Low species richness at down streams	Middle high species richness at down streams	High species richness at high mountain streams

<sup>a</sup>Maximum temperature in summer

<sup>b</sup>Distance from source

through repetition of numerous local competitions. Due to adaptive properties embedded within the network, an SOM has various advantages including utility for information extraction, flexibility in application, and network conformation.

### ***12.8.1 Utility for Training and Information Extraction***

Compared to other process-based or heuristic models, the model structure and algorithm in an SOM are relatively simple for training, requiring a short amount of time with few optimization techniques. Information extracted from local competition effectively accumulates, whereas the original topology of the initial data is preserved concurrently through the learning procedure. Although the algorithm is simple, its capacity for dimension compression is notable. An SOM is feasible for inferring the complex relationships among biological and environmental variables that are interwoven in natural ecosystems as stated above.

Consequently, an SOM is suitable for extracting information from large datasets consisting of numerous sample units and variables in different scales. In general, conventional multivariate analyses are not suitable to extract information from such large and complex datasets. In the dimensional reduction of principal component analysis, for instance, a large dataset with a large number of variables would produce a large number of significant principal components (i.e. relatively lower Eigenvalues for each component). Therefore, a few principal components may not be sufficient to address overall variation in the multidimensional datasets (Melssen et al. 1993).

### ***12.8.2 Visualization and Recognition***

An SOM has advantages in visualizing output presentations. Various styles are available to reveal associations between sample groups and input-output relationships. Profiles of multivariates, for instance, can be projected efficiently on component planes, providing a comprehensive view of the data structure as demonstrated in the example in Sect. 12.7 (Fig. 12.5).

An SOM is also useful for recognition of new datasets after training. An SOM is able to project the incoming input over the trained map through recognition: the association between new inputs to trained groups are conveniently identifiable. This process is a potential mechanism for automatic monitoring of ecosystem assessments. SOM sensitivity can be conducted to determine the importance of variables to address output correspondence responding to altered input data. Recently, prediction stability was checked in pest establishment risk by altering presence/absence data of species occurrence in Paini et al. (2010).

### ***12.8.3 Architecture Flexibility***

Due to the self-organizing property of the network, SOM networks can be flexibly modified to extract information according to problems *per se* in spatial and temporal domains. An SOM network can be flexibly modified to be applicable for spatial

variability. A geo-SOM was developed to fit geo-referenced data to take into account spatial dependency of variables in cluster formation (Bação et al. 2005).

An additional advantage of an SOM is noted by its feasibility in network architecture. Due to its self-organizing property, node configurations are flexible in evolving problem-oriented network architecture. One demonstration of SOM network architecture is modular networks. Stemming from operator maps allowing vector space transformation (Kohonen 1993; Kohonen et al. 1997), networks could be developed in a modular manner. SOM networks could be grown such that hypercubical output space is adaptively formed, enlarged by an extension of nodes and dimensions in the SOM architecture (Bauer and Villmann 1997; Villmann and Bauer 1998). The growing SOM allows the network to adaptively generate nodes to overcome the learning constraints when the current system does not sufficiently match the incoming data through the learning procedure (Marsland et al. 2002; Dittenbach et al. 2002).

#### ***12.8.4 Flexibility in Combining with Other Models***

The SOM is flexible in linking with other models. For instance, an MLP has been popularly combined with an SOM to exercise both supervised and unsupervised learning (Chon 2011). Data partitions were conducted first through an SOM, and then followed by an MLP to reveal the input-output relationships specifically.

#### ***12.8.5 Constraints on Measure Consistency and Output Variability***

As commonly shown in heuristic algorithms used in ANNs, distances between computation nodes in the output layer are not measured with the consistent scale on the map, whereas data variation could be expressed with parameters in multivariable statistics (i.e., Eigenvalue) (Peeters and Dassargues 2006). To overcome this disadvantage, alternatives are used to present the degree of association among grouped sample units on the map, including U-matrix (Ultsch and Siemon 1990) or hierarchical clustering (Ward 1963).

Another problem lies with variability in convergence. Because convergence is reached adaptively based on random processes in the learning procedure, the trained results would vary slightly based on each training condition, although the overall trend would be similar between different trials of training. Consequently, placement of grouped sample units may not be consistent on ordination maps because placement of the groups is relational to each other: actual placement of groups is determined adaptively through the convergence process. The main ordination (expressed as on a vertical, horizontal, or diagonal gradient) may be upside

down depending upon the convergence procedure. The subordinations could be repositioned according to the main ordinations on the map. U-matrix and hierarchical clustering could be helpful for addressing the degree of association across different levels of subgroups; similar groups could be allied closely depending upon their degree of association.

### ***12.8.6 Necessity of Sufficient Data***

Sufficient data is required for learning. The issue of data sufficiency can be applied to all learning methods. However, an SOM would be sensitive to the initial data size considering that topology preservation stems from the original data structure in dimensional compression. With a deficiency in sample number, dimensional compression could not be conducted accordingly from the first step. The number of samples should be sufficient to provide ample cases for training. For instance, caution is at least required when the sample size is lower than the number of variables.

## **12.9 Future Development**

The direction of future development for SOMs lies in solving the problems raised in constraints as stated above. Regarding problems in measure inconsistency and output variability, more computational approaches would be required regarding convergence, solution stability, and topology preservation (Fort 2006; Ritter and Schulten 1988). In addition, development of grouping techniques with statistical backgrounds is warranted in the future, extending the lines of methodology including U-matrix and hierarchical clustering.

Additionally, enhancing the advantages is needed in the future. An SOM could be developed further to be fluent in visualization, spatial and temporal data, network architecture development, and combining with other models to be fitted to complex ecosystem phenomena as stated above. Moreover, sensitivity tests and a supervised SOM garner attention regarding precision in addressing input-output relationships. For additional perspectives on the future development of SOMs, refer to Kalteh et al. (2008), Chon (2011), and Park et al. (2014).

## **12.10 Conclusions**

We explained the theory of SOMs and their application in ecological modelling, with a focus on learning processes, visualization, preprocessing of input data, and interpretation of results. SOMs are versatile in analysing non-linear and complex data, which are observed frequently in ecological systems.



## References

- Baao F, Lobo V, Painho M (2005) The self-organizing map, the Geo-SOM, and relevant variants for geosciences. *Comput Geosci* 31:155–163
- Bae M-J, Kwon Y-S, Hwang S-J, Park Y-S (2008) Comparison of four different ordination methods for patterning water quality of agricultural reservoirs. *Korean J Limnol* 41:1–10
- Balasubramanian M, Schwartz EL, Tenebaum JB et al (2002) The Isomap algorithm and topological stability. *Science* 295:7a
- Bauer H-U, Villmann T (1997) Growing a hypercubical output space in a self-organizing feature map. *IEEE Trans Neural Netw* 8:218–226
- Blayo F, Demartines P (1991) Data analysis: How to compare Kohonen neural networks to other techniques? *Proceedings of IWANN'91*. Springer, Berlin
- Borg I, Groenen P (1997) *Modern multidimensional scaling: theory and applications*. Springer, New York
- Bottin M, Giraudel J-L, Lek S, Tison-Rosebery J (2014) diatSOM: a R-package for diatom biotopology using self-organizing maps. *Diatom Res* 29:5–9
- Cayrou J, Compin A, Giani N, C er ghino R (2000) Species associations in lotic macroinvertebrates and their use for river typology: example of the Adour-Garonne drainage basin (France). *Annales de Limnologie-Int J Limnol* 36:189–202
- C er ghino R, Park Y-S (2009) Review of the self-organizing map (SOM) approach in water resources: commentary. *Environ Model Softw* 24:945–947
- C er ghino R, Giraudel J, Compin A (2001) Spatial analysis of stream invertebrates distribution in the Adour-Garonne drainage basin (France), using Kohonen self-organizing maps. *Ecol Model* 146:167–180
- Chan WS, Recknagel F, Cao H, Park HD (2007) Elucidation and short-term forecasting of microcystin concentrations in Lake Suwa (Japan) by means of artificial neural networks and evolutionary algorithms. *Water Res* 41:2247–2255
- Chon T-S (2011) Self-organizing maps applied to ecological sciences. *Ecol Inform* 6:50–61
- Chon T-S, Park Y-S (2006) Ecological informatics as an advanced interdisciplinary interpretation of ecosystems. *Ecol Inform* 1:213–217
- Chon T-S, Park YS, Moon KH, Cha EY (1996) Patterning communities by using an artificial neural network. *Ecol Model* 90:69–78
- Chon T-S, Park Y-S, Park KY et al (2004) Implementation of computational methods to pattern recognition of movement behavior of *Blattella germanica* (Blattaria: Blattellidae) treated with Ca<sup>2+</sup> signal inducing chemicals. *Appl Entom Zool* 39:79–96
- Daniel CL, Scott AR (2007) Abiotic and biotic factors explain independent gradients of plant community composition in ponderosa pine forests. *Ecol Model* 205:231–240
- Davies DL, Bouldin DW (1979) A cluster separation measure. *IEEE Trans Pattern Anal Mach Intell* 2:224–227
- De C aceres M, Legendre P, Moretti M (2010) Improving indicator species analysis by combining groups of sites. *Oikos* 119:1674–1684
- Dittenbach M, Rauber A, Merkl D (2002) Uncovering hierarchical structure in data using the growing hierarchical self-organizing map. *Neurocomput* 48:199–216
- Dufr ene M, Legendre P (1997) Species assemblages and indicator species: the need for a flexible asymmetrical approach. *Ecol Monogr* 67:345–366
- Ellison GN, Gotelli N (2004) *A primer of ecological statistics*. Sinauer, Sunderland, MA
- Foody GM (1999) The continuum of classification fuzziness in thematic mapping. *Photogramm Eng Remote Sens* 65:443–452
- Fort J-C (2006) SOM's mathematics. *Neural Netw* 19:812–816
- Friedel MJ (2012) Data-driven modeling of surface temperature anomaly and solar activity trends. *Environ Model Softw* 37:217–232
- Gauch HG (1982) *Multivariate analysis in community ecology*. Cambridge University Press, Cambridge

- Gevrey M, Rimet F, Park YS et al (2004) Water quality assessment using diatom assemblages and advanced modelling techniques. *Freshw Biol* 49:208–220
- Giraudel J, Lek S (2001) A comparison of self-organizing map algorithm and some conventional statistical methods for ecological community ordination. *Ecol Model* 146:329–339
- Griebeler EM, Seitz A (2006) The use of Markovian metapopulation models: reducing the dimensionality of transition matrices by self-organizing Kohonen networks. *Ecol Model* 192: 271–285
- Hill MO, Gauch HG (1980) Detrended correspondence analysis: an improved ordination technique. *Vegetation* 42:47–58
- Hruschka H, Natter M (1999) Comparing performance of feedforward neural nets and K-means for cluster-based market segmentation. *Eur J Oper Res* 114:346–353
- Huntley B (1999) Species distribution and environmental change: considerations from the site to the landscape scale. *Ecosystem management: questions for science and society*. Royal Holloway Institute for Environmental Research, Virginia Water
- Hyun K, Song M-Y, Kim S, Chon T-S (2005) Using an artificial neural network to patternize long-term fisheries data from South Korea. *Aquat Sci* 67:382–389
- Jongman RH, Ter Braak CJ, Van Tongeren OF (1995) *Data analysis in community and landscape ecology*. Cambridge University Press, Cambridge
- Kalteh AM, Hjorth P, Berndtsson R (2008) Review of the self-organizing map (SOM) approach in water resources: analysis, modelling and application. *Environ Model Softw* 23:835–845
- Kaski S (1997) Data exploration using self-organizing maps. *Acta polytechnica scandinavica, mathematics, computing and management in engineering series no. 82*. Finnish Academy of Technology, Espoo, Finland
- Kiviluoto K (1996) Topology preservation in self-organizing maps. In: *Proceedings of ICNN'96*. IEEF. Service Center, Piscataway
- Kohonen T (1982) Self-organized formation of topologically correct feature maps. *Biol Cybern* 43:59–69
- Kohonen T (1993) Physiological interpretation of the self-organizing map algorithm. *Neural Netw* 6:895–905
- Kohonen T (2001) *Self-organizing maps*. Springer, Berlin
- Kohonen T, Kaski S, Lappalainen H (1997) Self-organized formation of various invariant-feature filters in the adaptive-subspace SOM. *Neural Comput* 9:1321–1344
- Legendre P, Legendre L (1998) *Numerical ecology*. Elsevier, Amsterdam
- Mahechaa MD, Martínez A, Lischeida G, Beckc E (2007) Nonlinear dimensionality reduction: alternative ordination approaches for extracting and visualizing biodiversity patterns in tropical montane forest vegetation data. *Ecol Inform* 2:138–149
- Marsland S, Shapiro J, Nehmzow U (2002) A self-organising network that grows when required. *Neural Netw* 15:1041–1058
- McCune B, Grace JB (2002) *Analysis of ecological communities*. MjM, Grededen Beach, Oregon
- Melssen W, Smits J, Rolf G, Kateman G (1993) Two-dimensional mapping of IR spectra using a parallel implemented self-organising feature map. *Chemom Intell Lab Syst* 18:195–204
- Melssen W, Smits J, Buydens L, Kateman G (1994) Using artificial neural networks for solving chemical problems: Part II. Kohonen self-organising feature maps and Hopfield networks. *Chemom Intell Lab Syst* 23:267–291
- Nikolic N, Park Y-S, Sancristobal M et al (2009) What do artificial neural networks tell us about the genetic structure of populations? The example of European pig populations. *Genet Res* 91: 121–132
- Obach M, Wagner R, Werner H, Schmidt H-H (2001) Modelling population dynamics of aquatic insects with artificial neural networks. *Ecol Model* 146:207–217
- Osborne JW, Overbay A (2004) The power of outliers (and why researchers should always check for them). *Pract Assess Res Eval* 9:1–12
- Paini DR, Worner SP, Cook DC et al (2010) Using a self-organizing map to predict invasive species: sensitivity to data errors and a comparison with expert opinion. *J Appl Ecol* 47:290–298
- Park Y-S, Chon T-S (2007) Biologically-inspired machine learning implemented to ecological informatics. *Ecol Model* 203:1–7

- Park Y-S, Chon T-S (2015) Editorial: ecosystem assessment and management. *Ecol Inform* 29: 93–95
- Park Y-S, Chung Y-J (2006) Hazard rating of pine trees from a forest insect pest using artificial neural networks. *For Ecol Manage* 222:222–233
- Park Y-S, C  r  ghino R, Compin A, Lek S (2003) Applications of artificial neural networks for patterning and predicting aquatic insect species richness in running waters. *Ecol Model* 160: 265–280
- Park Y-S, Chon T-S, Kwak I-S, Lek S (2004) Hierarchical community classification and assessment of aquatic ecosystems using artificial neural networks. *Sci Total Environ* 327:105–122
- Park Y-S, Chung N-I, Choi K-H et al (2005) Computational characterization of behavioral response of medaka (*Oryzias latipes*) treated with diazinon. *Aquat Toxicol* 71:215–228
- Park Y-S, Tison J, Lek S et al (2006) Application of a self-organizing map to select representative species in multivariate analysis: a case study determining diatom distribution patterns across France. *Ecol Inform* 1:247–257
- Park Y-S, Kwon Y-S, Hwang S-J, Park S (2014) Characterizing effects of landscape and morphometric factors on water quality of reservoirs using a self-organizing map. *Environ Model Softw* 55:214–221
- Peeters L, Dassargues A (2006) Comparison of Kohonen’s self-organizing map algorithm and principal component analysis in the exploratory data analysis of a groundwater quality dataset. Proceedings of 6th international conference on geostatistics for environmental applications. Rhodes, Greece, 25–27 October 2006, pp 1–12
- Recknagel F, French M, Harkonen P, Yabunaka K (1997) Artificial neural network approach for modelling and prediction of algal blooms. *Ecol Model* 96(1–3):11–28
- Recknagel F, Talib A, van der Molen D (2006) Phytoplankton community dynamics of two adjacent Dutch lakes in response to seasons and eutrophication control unraveled by non-supervised artificial neural networks. *Ecol Inform* 1:277–286
- Ritter H, Schulten K (1988) Convergence properties of Kohonen’s topology conserving maps: fluctuations, stability, and dimension selection. *Biol Cybern* 60:59–71
- Roux O, Gevrey M, Arvanitakis L et al (2007) ISSR-PCR: tool for discrimination and genetic structure analysis of *Plutella xylostella* populations native to different geographical areas. *Mol Phylogenet Evol* 43:240–250
- Samarasinghe S, Strickert G (2013) Mixed-method integration and advances in fuzzy cognitive maps for computational policy simulations for natural hazard mitigation. *Environ Model Softw* 39:188–200
- Shepard RN (1962) The analysis of proximities: multidimensional scaling with an unknown distance function. II. *Psychometrika* 27:219–246
- Strebel K, Espinosa G, Giralt F et al (2013) Modeling airborne benzene in space and time with self-organizing maps and Bayesian techniques. *Environ Model Softw* 41:151–162
- Tenenbaum YB, de Silva V, Langford JC (2000) A global geometric framework for nonlinear dimensionality reduction. *Science* 290:2319–2323
- Tison J, Giraudel J, Coste M et al (2004) Use of unsupervised neural networks for ecoregional zoning of hydrosystems through diatom communities: case study of Adour-Garonne watershed (France). *Arch Hydrobiol* 159:409–422
- Ulsch A, Siemon HP (1990) Kohonen’s self organizing feature maps for exploratory data analysis. In: Proceedings of INN’90. Kluwer Academic, Dordrecht
- Vesanto J (1999) SOM-based data visualization methods. *Intelligent Data Anal* 3:111–126
- Villmann T, Bauer H-U (1998) Applications of the growing self-organizing map. *Neurocomputing* 21:91–100
- Ward JH Jr (1963) Hierarchical grouping to optimize an objective function. *J Am Stat Assoc* 58: 236–244
- Wehrens R (2015) Package ‘kohonen’. version 2.0.19
- Wilppu E (1997) The visualisation capability of self-organizing maps to detect deviations in distribution control. TUCS technical report no. 153. Turku Centre for Computer Science, Finland

# Chapter 13

## GIS-Based Data Synthesis and Visualization

**Duccio Rocchini, Carol X. Garzon-Lopez, A. Marcia Barbosa, Luca Delucchi, Jonathan E. Olandi, Matteo Marcantonio, Lucy Bastin, and Martin Wegmann**

**Abstract** Synthesizing and properly visualizing data in 2D systems is a key issue when aiming at explaining spatial patterns by spatial processes.

In this chapter we address the topics synthesis and visualization in relation to following ecological issues: (1) synthesizing species distribution models relying on virtual species, (2) visualizing spatial uncertainty in species distribution based on cartograms, (3) fuzzy methods to synthesize species distribution uncertainty,

---

D. Rocchini (✉)

Center Agriculture Food Environment, University of Trento, Via E. Mach 1, 38010 S. Michele all'Adige (TN), Italy

Centre for Integrative Biology, University of Trento, Via Sommarive, 14, 38123 Povo (TN), Italy

Department of Biodiversity and Molecular Ecology, Fondazione Edmund Mach, Research and Innovation Centre, San Michele all'Adige, Trento, Italy  
e-mail: [duccio.rocchini@unitn.it](mailto:duccio.rocchini@unitn.it)

C.X. Garzon-Lopez

Ecology and Vegetation Physiology Group (EcoFiv), Universidad de los Andes, Bogotá, Colombia

A.M. Barbosa

Centro de Investigacao em Biodiversidade e Recursos Geneticos (CIBIO), InBIO Research Network in Biodiversity and Evolutionary Biology, University of Evora, Evora, Portugal

L. Delucchi • J.E. Olandi

Department of Biodiversity and Molecular Ecology, Fondazione Edmund Mach, Research and Innovation Centre, San Michele all'Adige, Trento, Italy

M. Marcantonio

Department of Pathology, Microbiology, and Immunology, School of Veterinary Medicine, University of California Davis, Davis, CA, USA

L. Bastin

School of Computer Science, Aston University, Aston, Birmingham, UK

European Commission, Joint Research Centre (JRC), Directorate D - Sustainable Resources, School of Computer Science, Aston University, Aston, Birmingham, UK

M. Wegmann

Department of Remote Sensing, Remote Sensing and Biodiversity Research Group, University of Wuerzburg, Wuerzburg, Germany

(4) remote sensing data synthesis by exploratory analysis and replotting data in new systems, (5) measuring and visualizing ecological diversity from space based on generalized entropy, and (6) neutral landscape for testing ecological theories. We will make use of examples from the free and open source software GRASS GIS and R.

## 13.1 Introduction

In spatial ecology synthesizing and properly visualizing data in 2D systems is a key issue when aiming at explaining spatial patterns by spatial processes. This has been demonstrated in a number of ecological and geographical studies, dealing with different scientific aims (e.g. Rocchini et al. 2016). Increasing availability of open ecological data through networks like the Global Biodiversity Information Facility or the DataONE (Data Observation Network for Earth), on the one hand, and remote sensing data on the other, makes it necessary to promote methods for data synthesis and visualization.

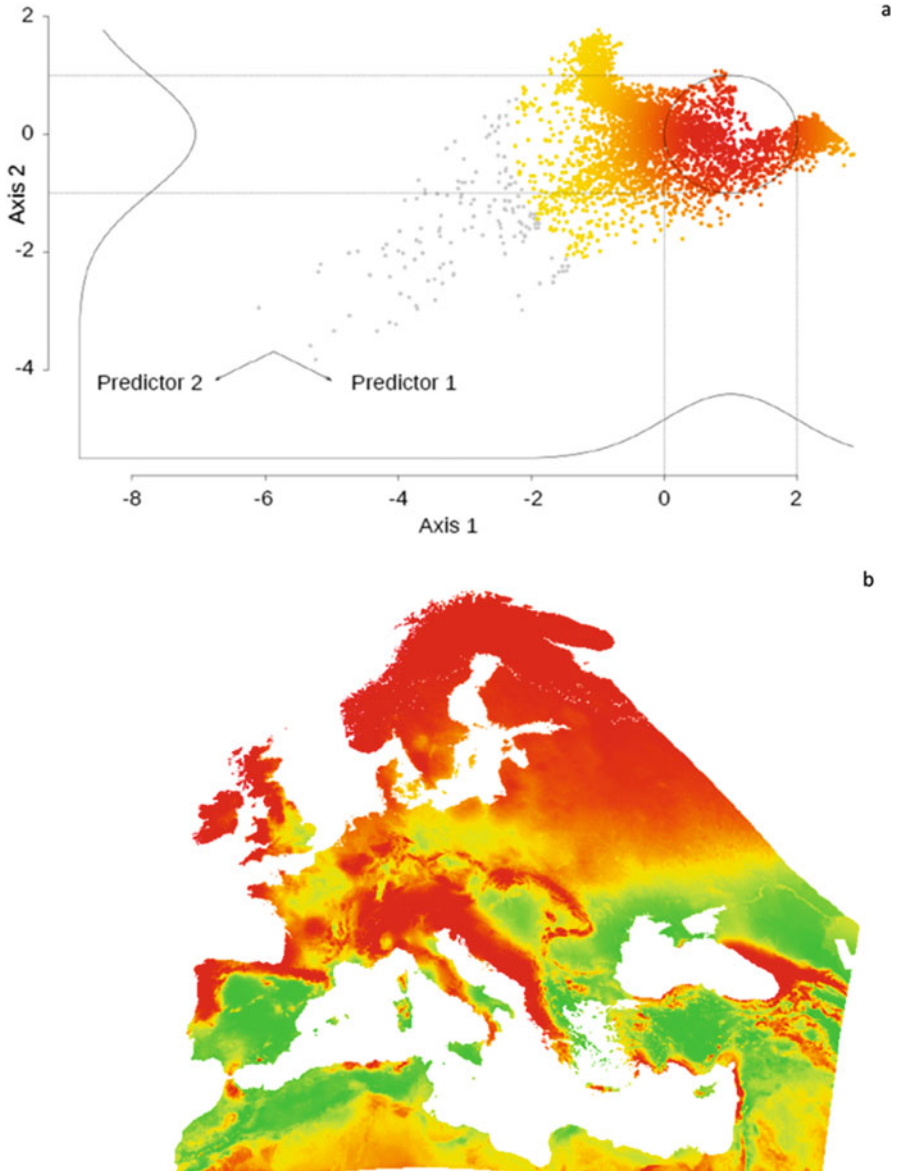
In this book chapter we address currently used methods for synthesis and visualization, and provide examples from the free and open source software GRASS GIS (Neteler et al. 2012) and R (R Development Core Team 2016).

## 13.2 Synthesizing Species Distributions by Virtual Species

Virtual species represents a powerful approach to build species distributions based on known ecological parameters for illustrating species spread. The package “virtualspecies” in the R software allows to create habitat suitability maps as shown in Fig. 13.1. The bioclimatic variables annual temperature and annual precipitation were used as proxies of habitat suitability, and obtained from the bioclim dataset at 1 km spatial resolution (Hijmans et al. 2005). The resulting map shows a species with a wide niche distributed throughout Europe (Fig. 13.1b).

## 13.3 Cartograms to Synthesize and Visualize Sampling Effort Bias

In ecology, a number of studies have dealt with the prediction of species distribution and diversity over space and its changes over time based on a set of environmental predictors related to environmental variability, productivity, spatial constraints and climate drivers. Species distribution models have been acknowledged as the most powerful methods to map the spread of plant and animal species. The basic approach used to create maps based on predictors is to rely on linear models to create gridded landscapes of potential distribution of species based on



**Fig. 13.1** A virtual species distribution might be useful to synthesize species spread conditional to known ecological drivers. Panel (a) environmental suitability of the virtual species in the predictor space, represented with two climate predictor variables. *Red*, high suitability, *orange*, medium suitability, *yellow*, low suitability. Panel (b) the habitat suitability for the virtual species created. Suitability is represented from low in *red* to high in *green*

point or polygon local data. In most cases, the output is a density function in two dimensions representing the distribution  $S_x$  of the  $x$  species. In general, boundaries are defined sharply based on thresholds of predictors or factors such as land cover e.g. Comber et al. (2013) or based on continuous variability of predictors such as air temperature. Uncertainty in such generalized linear models, generalized additive models, maximum entropy models is mainly derived from pseudo-absent input data (Foody 2011) as well as from models' bias, i.e. the error deriving from the selected model. Hence, the visualization of uncertainty in two dimensions is strongly suggested (Comber et al. 2012; Rocchini et al. 2013).

Concerning bias related to sampling effort, we rely on one of the most commonly used datasets in biodiversity studies at large spatial extents, namely the GBIF dataset. GBIF data comprises a huge range of species occurrence observations collected with a wide variety of sampling approaches. It spans from well-established plot censuses to direct observations collected during field trips. Consequently, some of the data points are at the centre of censured grids i.e. each point comprises the species located at a species  $c$ -size quadrant, or correspond to single observations of individuals of the same species. These differences also depend on the methodologies used to observe occurrences per taxon. Plots within transects are commonly employed in vegetation censuses, while transects, point counts and live traps are preferred in the case of animals. Moreover, the variation in factors such as country-specific biodiversity monitoring schemes, funding schemes, focal ecosystems, accessibility to remote areas add more sources of variation, especially at multinational scales (Barbosa et al. 2013).

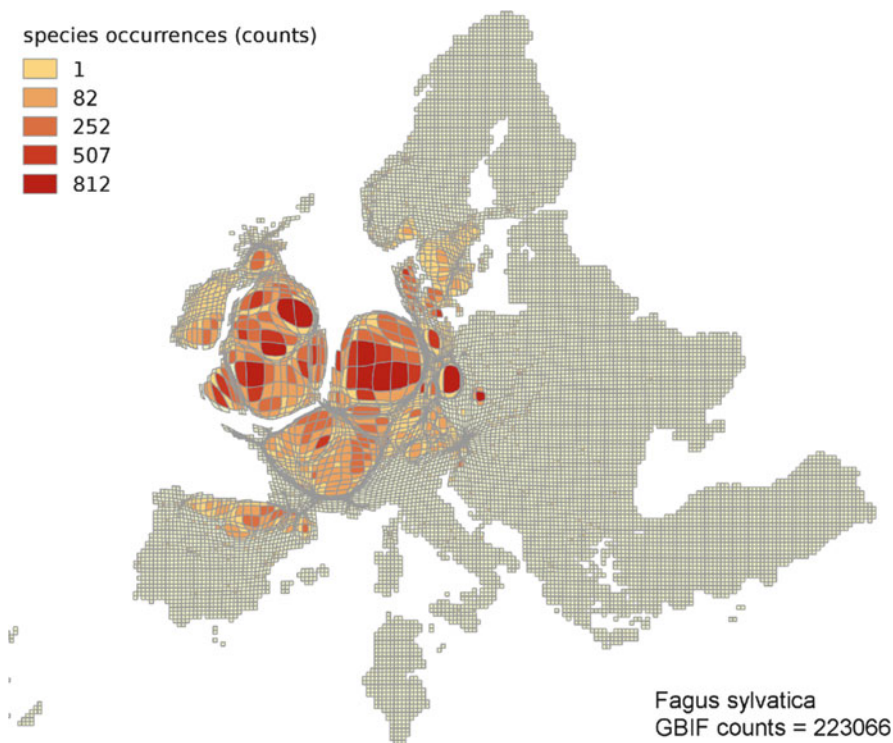
Undoubtedly, all those sources of variation result in non-homogeneous samples that influence not only the development of accurate species distribution maps but more importantly conservation and management decisions focusing on such a distribution of biodiversity (Rocchini et al. 2011).

In this study we synthesize spatial uncertainty in the sampling effort of the GBIF data, by explicitly taking into account potential area effects of European countries. We aim at quantifying and mapping the uncertainty derived from the variation in observations due to differences in sampling efforts. Cartograms serve well this purpose where the shape of objects is directly related to a certain property, such as uncertainty. Cartograms build on the standard treatment of diffusion, in which the current density is given by:

$$J = v(r,t) * p(r,t) \quad (13.1)$$

where  $v(r,t)$  and  $p(r,t)$  are the velocity and density at position  $r$  and time  $t$ . For more details see Gastner and Newman (2004).

Cartograms allow to visualize spatial uncertainty in the results by changing the size of the polygons based on its information density e.g. number of observations, variation. Based on this approach the spatial distribution of a species (e.g. *Fagus sylvatica*) can be represented in the coloured grid in Fig. 13.2 where the colour represents the abundance of the species and the distortion of the shape of each grid cell might represent the sampling bias, i.e. more distorted cells may have been oversampled compared to others.



**Fig. 13.2** Cartograms can be used to show the sampling effort bias in species distribution modelling. In this case, oversampled cells are more distorted than the others; hence in such cells the higher abundance of *Fagus sylvatica* might be an artifact of oversampling

### 13.4 Fuzzy Methods to Synthesize Species Distribution Uncertainty

Beside sampling bias, taxonomic bias may occur when different operators or scientists deal with the association of each individual to a taxonomic category. Fuzzy set theory can assist in processing information uncertainty related to each species (hereafter also generally related to class as in fuzzy set theory). The concept of fuzzy sets has been introduced by Zadeh (1965) widely been used in ecology since.

The principle behind fuzzy set theory is that the situation of one class being exactly right, other classes being equally or exactly wrong often does not exist. Conversely, Gopal and Woodcock (1994) suggests that there is a gradual change from membership to non-membership.

A fuzzy set is defined as follows:

if  $U$  denotes a universe of entities  $u$ , the fuzzy set  $F$  is represented as:



$$F = (u, \mu(u)) \mid u \in U \quad (13.2)$$

where the membership function  $\mu(u)$  associates for each entity  $u$  the degree of membership into the set  $F$  and ranges within the interval  $[0,1]$ .

Hence, fuzzy sets might represent a good starting point for continuously mapping species by relying on each species as:

$$F_i = (u, \mu_i(u)) \mid u \in U \quad (13.3)$$

$$F_j = (u, \mu_j(u)) \mid u \in U \quad (13.4)$$

In this case, for each species  $i$  and  $j$  a map is derived based on fuzzy training data taken in the field that represents species probability of occurrence. In this case, according to Boggs (1949), uncertainty is explicit in the sense that a probability of occurrence of each sampled individual to each species is mapped instead of determining a crisp set of species with 100% accuracy.

A fuzzy determination of a species can be derived e.g. as the probability of correct determination by different operators or scientists. Figure 13.3 represents an example for the foraminifera species *Keratella quadrata*. It shows a map of the species' presence per country or region together with the probability (as inverse distance) of occurrence of each individual within the species or group. The analysis was performed by means of the fuzzySym package (Barbosa 2015) for the R software.

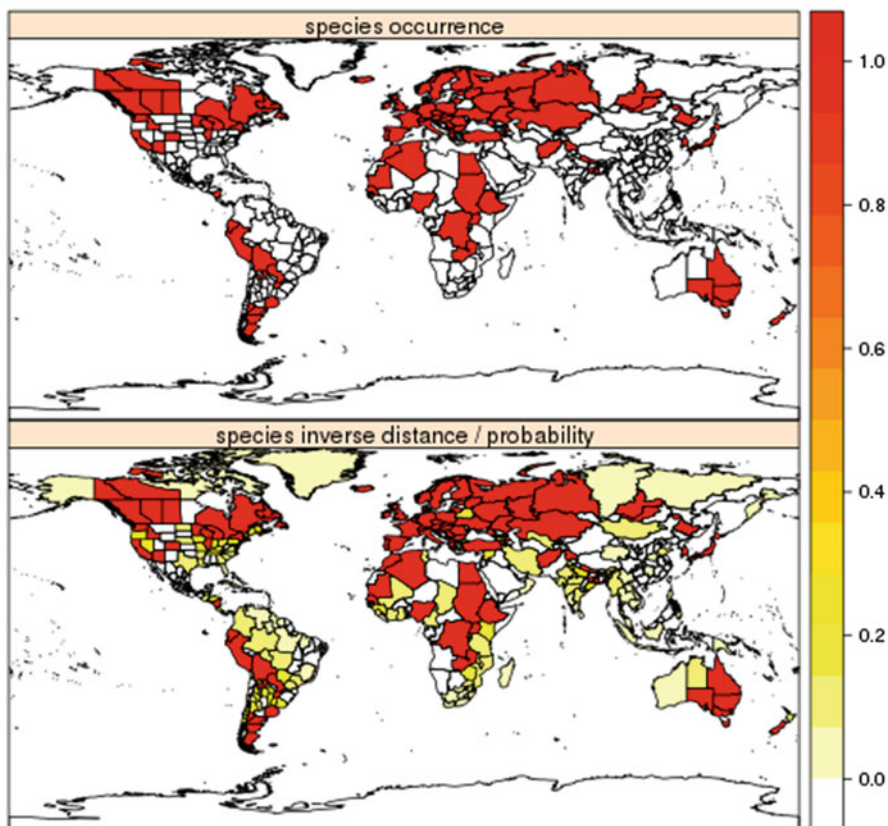
## 13.5 Synthesis of Remote Sensing Data

### 13.5.1 Exploratory Data Analysis

In some cases remote sensing data are correlated to each other such as high reflectance in a certain region of the electromagnetic spectrum might be related to that in another one. In other cases, indices derived from remote sensing data are implicitly correlated. This is the case when calculating texture measures.

#### 13.5.1.1 Correlation of Remotely Sensed Bands by Hexagon Binning

Hexagon binning is a powerful technique for synthesizing geographical data, especially those based on huge 2D matrices. Figure 13.4 displays an example from two Landsat images freely available in the North Carolina dataset of GRASS GIS (<https://grass.osgeo.org/download/sample-data/>). Hexagon binning by means of the R package "hexbin" clumps point clouds into hexagons once matrices are imported to R by using the package *rgrass7* (Bivand et al. 2016). It visualises the contrast between Landsat NIR infrared versus Landsat red (Fig. 13.4).

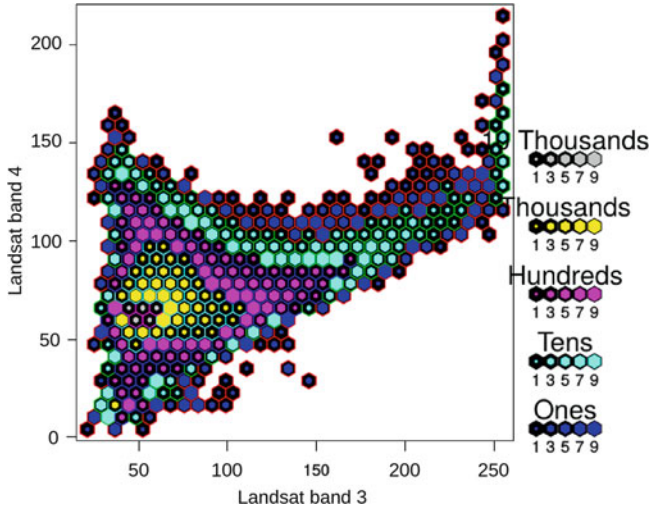


**Fig. 13.3** Representation of the presence (*top*) of the foraminifera species *Keratella quadrata* and the probability (as inverse distance) of occurrence (*down*) of each determined individual to that species. While the presence/absence map has obviously only *red* (1—presence) and *white* (0—absence) colour, the probability map based on inverse distance covers the whole range of decimal values from 0 to 1

In contrast to normal plots, hexagon binning allows to display the amount of points per each value in the point cloud.

### 13.5.1.2 Correlation Among Several Layers by Texture Measures

Texture measures provide information about the amount of variability in a neighbourhood. This has a number of repercussions related to biodiversity studies in which local spatial heterogeneity is used as a proxy for species diversity (Rocchini et al. 2016). In most cases texture measures are implicitly correlated. Showing such correlation is important to synthesize the texture system and avoid redundant information.



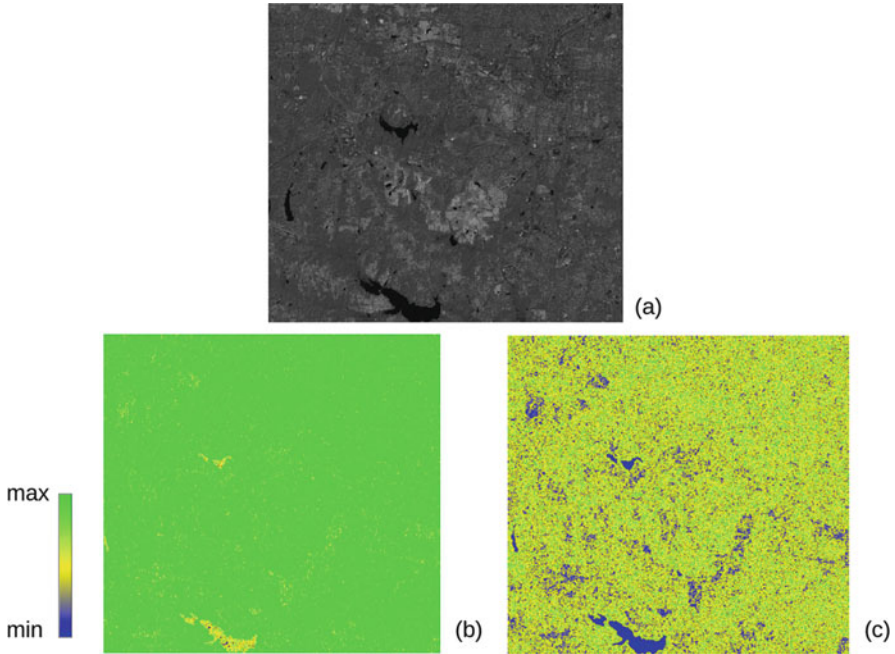
**Fig. 13.4** Starting from two Landsat ETM+ bands, hexagon binning allows to explore their relationship by also showing the amount of data per each value

The following example suggested by Haralick et al. (1973) applies GRASS GIS to calculate the texture measures in a neighbourhood of pixels by following steps: (1) the angular second moment, as a measure of local homogeneity; (2) the contrast, a grey-level variation with respect to neighbour pixels; (3) the correlation, a linear dependency value; (4) the variance in the neighbouring moving window (see also r.neighbors); (5) the entropy, an index of randomness; (6) the sum average; (7) the sum entropy; (8) the sum variance; (9) the difference in variance; (10) the difference in entropy; (11) the inverse distance moment, i.e. the inverse of the previously described contrast measure; and (12) the maximal correlation coefficient. We refer to Haralick et al. (1973) for a detailed description of all the measures. Figure 13.5 presents two of the aforementioned maps for entropy and variance generated from a Landsat ETM+ image. In order to represent the level of correlation of such measures, R package ‘corrplot’ allows to create a graphical matrix of correlation coefficients provided that data have been imported from GRASS GIS by the R package ‘rgrass7’.

Figure 13.6 illustrates the amount of correlation among texture measures. During modelling of ecosystem complexity most of the texture measures should be first synthesized by a graphical output since they are strongly correlated.

### 13.5.2 *Fourier Transformations*

Remote sensing data are a powerful input for studying landscape transformations in space and time. In some cases, such transformations cannot be inspected in the



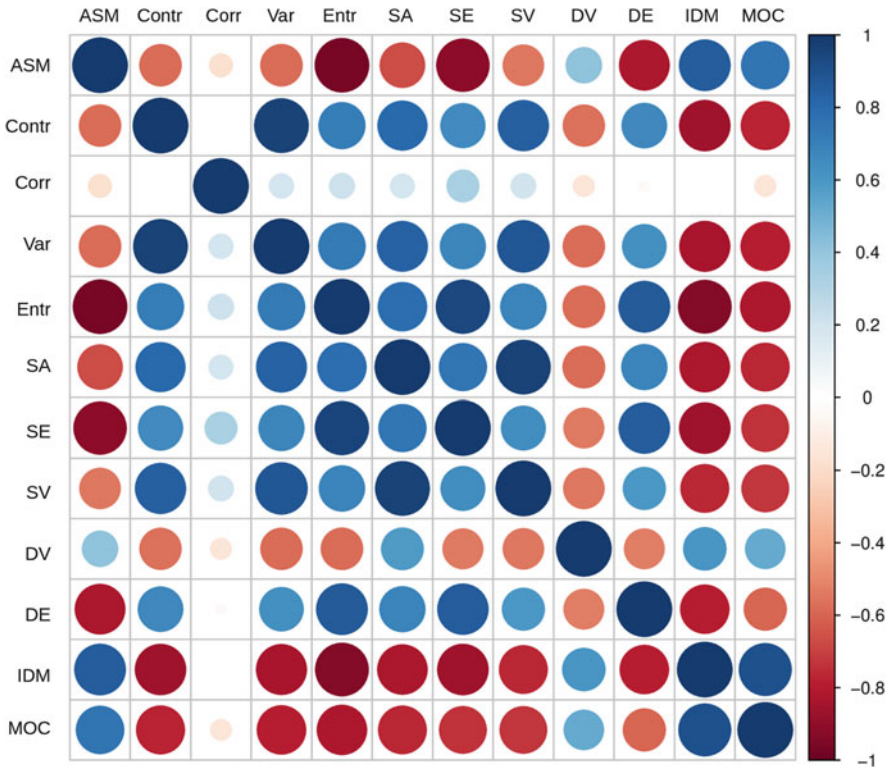
**Fig. 13.5** Examples of texture measures derived from a Landsat ETM+ band 4. (a) NIR, (b) entropy and (c) variance

normal space but need to be further transformed to highlight such difference. The use of transformations within frequency spaces to measure variation in a signal has long been acknowledged. While methods are known based on orthonormal series such as rectangular decomposition of waves (Walsh 1923), the most commonly used Fourier transformations (Fourier 1822) rely on continuous waves. The methods for detecting landscape change based on continuous instead of classified information rely on continuous functions which neither require a-priori field information nor specific models based on the data being used. In view of this fact the Fourier transformation appears most suitable.

The continuous function  $f(x)$  may be described in a spatial domain. According to Fourier (1822), every  $f(x)$  can be transformed into a continuum of sinusoidal functions of varying frequency as follows:

$$F(\omega) = \int_{-\infty}^{\infty} f(x)e^{-2\pi i\omega x} dx \quad (13.5)$$

where  $\omega$  = frequency, also known as radian frequency since it is expressed in radians per spatial units. Extending Eq. (13.5) to two dimensions implies



**Fig. 13.6** A corplot by R allows to directly show the amount of correlation among remotely sensed layers. In this example, the system composed by texture measures [sensu Haralick et al. (1973)] is generally highly positively or negatively correlated. Refer to the main text for additional information on single measures’ acronyms. Reproduced from Rocchini et al. (2013)

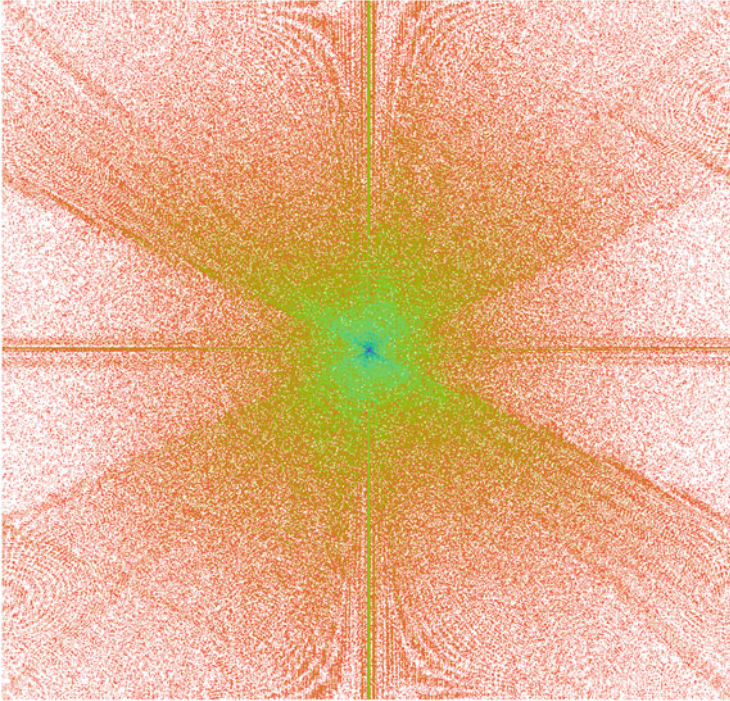
considering a two-dimensional function  $f(x,y)$ , e.g. a raster matrix. Its Fourier transformations reads as follows:

$$F(\omega, \nu) = \int \int_{-\infty}^{\infty} f(x, y) e^{-2\pi i(\omega x + \nu y)} dx, dy \tag{13.6}$$

where  $\omega, \nu$  = frequency coordinates.

Figure 13.7 illustrates the Fourier space where high frequency values (high heterogeneity) are at the border of the image while low frequency values (high homogeneity) are at the centre. Hence the higher the value of pixels at the border, the higher the heterogeneity or complexity of the whole image.





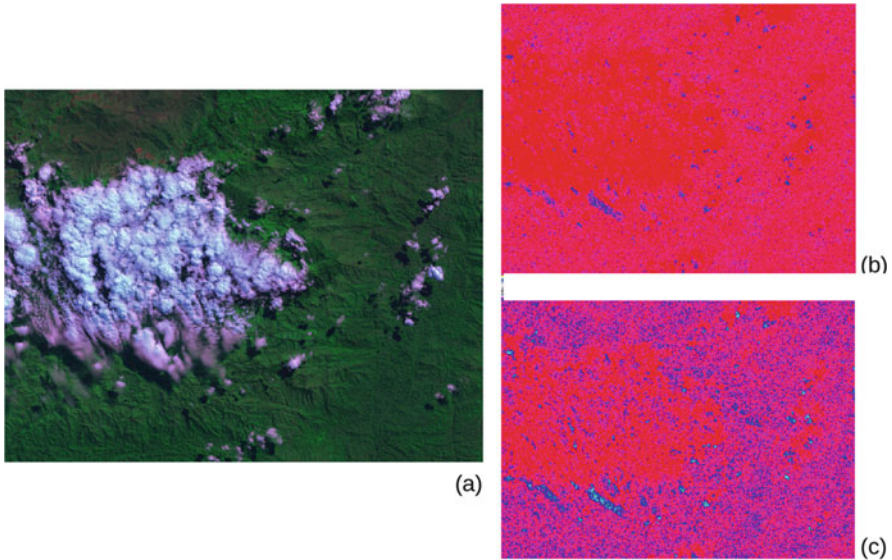
**Fig. 13.7** A Fourier image gathered applying Eq. (13.6) to a remotely sensed image. The external part of a Fourier frequency space contains high frequency values while the part near the centre contains low frequency values. Hence the higher the amount of *red values* (higher values) occupying the *white* (low values) external part, the higher will be the heterogeneity in the landscape

### 13.6 Synthesizing Diversity Measurements from Space: The Case of Generalized Entropy

From a practical point of view, distinct diversity measures summarize a large multivariate data set into one single value based on distinct objectives and approaches. Distinct diversity measures always result in a loss of information, and summary statistics seems to be capable of unequivocally synthesizing all aspects of diversity (Ricotta 2005). However, Renyi (1970) proposed a generalized entropy as follows:

$$H_{\alpha} = \frac{1}{1 - \alpha} \times \ln \sum p^{\alpha}$$

which is extremely flexible and powerful since many popular diversity indices are simply special cases of  $H_{\alpha}$ . As an example, for  $\alpha = 0$ ,  $H_0 = \ln(N)$  namely the logarithm of richness ( $N =$  number of Digital Numbers), i.e. the maximum Shannon



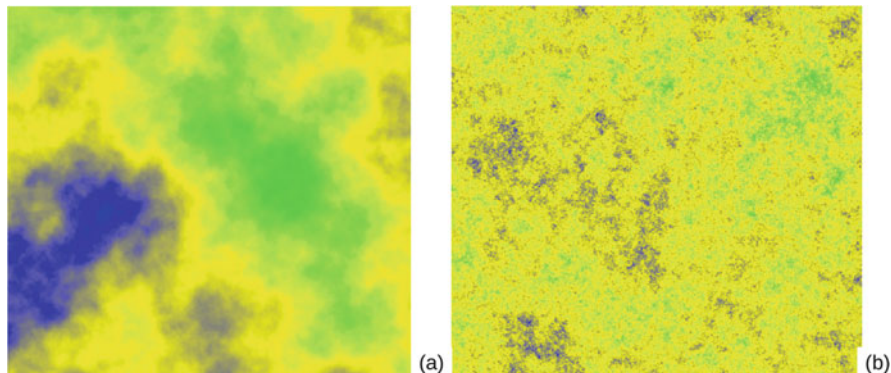
**Fig. 13.8** Starting from the same remotely sensed image (*left*) R\_enyi generalized entropy based on di\_ erent alpha values can lead to di\_ erent maps to better synthesize the continuous variation of ecological diversity in space. This panels are related to calculations in GRASS GIS

entropy index ( $H_{\max}$ ) which is used as the denominator of the Pielou index, while for  $\alpha = 2$ ,  $H_2 = \ln(1/D)$  where  $D$  is the Simpson Dominance index (impossible to make sense of this sentence). For  $\alpha = 1$  the Renyi entropy is defined in the limiting sense using l'Hospital's rule of calculus, and  $H_1 = \text{Shannon's entropy } H$ .

Renyi's framework offers a continuum of possible diversity measures, which differ in their sensitivity to rare and abundant DNs, becoming increasingly regulated by the most common DNs when increasing the values of  $\alpha$ . In this view, changing  $\alpha$  can be considered as a scaling operation that takes place not in the real but in the data space. That is why Renyi's generalized entropy has been referred to as a continuum of diversity measures (Ricotta and Avena 2003). Changing the parameter  $\alpha$  will change the behaviour of the formula generating different maps of diversity as displayed in Fig. 13.8. As a result Fig. 13.8 represents a continuum of diversity values over space instead of single measures.

### 13.7 Neutral Landscapes

Patterns in the field can be correlated to random patterns by calculating the deviation from random expectations in two dimensions as suggested by Hanspach et al. (2011). To accomplish this goal, different kinds of lattice surfaces can be generated, including completely random surfaces, Gaussian distribution, and fractal



**Fig. 13.9** Artificial landscapes with different fractal dimensions: (a) 2.1 and (b) 2.96

surfaces with a predefined fractal dimension. Lattice surfaces help to compare real patterns found in landscape ecology with neutral landscape to determine if the real patterns show a significant deviation from random (neutral) expectations.

Landsat images can be tested against random surfaces in order to find clumped parts of a Landsat image which significantly deviate from random expectations over space. A more sophisticated but still straightforward neutral model can be represented by a Gaussian surface, which should graphically not be different from a random surface but would have normally distributed values in two dimensions, and means and standard deviations can be defined a-priori.

Figure 13.9 provides an example for fractal surfaces according to Mandelbrot and Blumen (1989) based on the assumption that surfaces with a fractal dimension from 2 to 3 might represent severe differences in their roughness or complexity (Imre et al. 2011). Fractal surfaces can be used to test the complexity of real patterns against lattice images.

## 13.8 Conclusions

This chapter has demonstrated the use of free and open-source software to synthesize and visualize spatio-ecological data. Different approaches for processing spatio-ecological information have been discussed for a variety of research fields. We referred largely to methods that have already been implemented and tested in GRASS GIS and R packages. However the GRASS GIS platform and R software allows users to contribute new features to the already existing extensive software libraries.



## References

- Barbosa AM (2015) fuzzySim: applying fuzzy logic to binary similarity indices in ecology. *Methods Ecol Evol* 6:853–858
- Barbosa AM, Pautasso M, Figueiredo D (2013) Species- people correlations and the need to account for survey effort in biodiversity analyses. *Divers Distrib* 19:1188–1197
- Bivand R, Krug R, Neteler M, Jeworutzki S (2016) rgrass7 – interface Between GRASS 7 Geographical Information System and R. R software package
- Boggs S (1949) An atlas of ignorance: a needed stimulus to honest thinking and hard work. *Proc Am Philos Soc* 93:9–258
- Comber A, Fisher P, Brunsdon C, Khmag A (2012) Spatial analysis of remote sensing image classification accuracy. *Remote Sens Environ* 127:237–246
- Comber A, See L, Fritz S, Van der Velde M, Perger C, Foody GM (2013) Using control data to determine the reliability of volunteered geographic information about land cover. *Int J Appl Earth Obs Geoinf* 23:37–48
- Foody GM (2011) Impacts of imperfect reference data on the apparent accuracy of species presence/absence models and their predictions. *Glob Ecol Biogeogr* 20:498–508
- Fourier J (1822) *Thorie Analytique de la Chaleur*. Didot, Paris
- Gastner M, Newman M (2004) Diffusion-based method for producing density- equalizing maps. *Proc Natl Acad Sci U S A* 15:7499–7504
- Gopal S, Woodcock C (1994) Theory and methods for accuracy assessment of thematic maps using fuzzy sets. *Photogramm Eng Remote Sens* 60:181–188
- Hanspach J, Kuehn I, Schweiger O, Pompe S, Klotz S (2011) Geographical patterns in prediction errors of species distribution models. *Glob Ecol Biogeogr* 20:779–788
- Haralick R, Shanmugam K, Dinstein I (1973) Textural features for image classification. *IEEE Trans SMC* 6:610–621
- Hijmans RJ, Cameron SE, Parra JL, Jones PG, Jarvis A (2005) Very high resolution interpolated climate surfaces for global land areas. *Int J Climatol* 25:1965–1978
- Imre AR, Cseh D, Neteler M, Rocchini D (2011) Korcak dimension as a novel indicator of landscape fragmentation and re-forestation. *Ecol Indic* 11:1134–1138
- Mandelbrot BB, Blumen A (1989) Fractal geometry: What is it, and what does it do? *Proc R Soc Lond A* 423:3–16
- Neteler M, Bowman MH, Landa M, Metz M (2012) GRASS GIS: a multi-purpose open source GIS. *Environ Model Softw* 31:124–130
- R Development Core Team (2016). R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. <http://www.R-project.org/>
- Rényi A (1970) *Probability theory*. North Holland Publishing Company, Amsterdam
- Ricotta C (2005) On possible measures for evaluating the degree of uncertainty of fuzzy thematic maps. *Int J Remote Sens* 26:5573–5583
- Ricotta C, Avena G (2003) On the relationship between Pielou's evenness and landscape dominance within the context of Hill's diversity profiles. *Ecol Indic* 2:361–365
- Rocchini D, Hortal J, Lengyel S, Lobo JM, Jiménez-Valverde A, Ricotta C, Bacaro G, Chiarucci A (2011) Accounting for uncertainty when mapping species distributions: the need for maps of ignorance. *Prog Phys Geogr* 35:211–226. doi:10.1177/0309133311399491
- Rocchini D, Foody GM, Nagendra H, Ricotta C, Anand M, He KS, Amici V, Kleinschmit B, Foerster M, Schmidlein S, Feilhauer H, Ghisla A, Metz M, Neteler M (2013) Uncertainty in ecosystem mapping by remote sensing. *Comput Geosci* 50:128–135. doi:10.1016/j.cageo.2012.05.022
- Rocchini D, Boyd DS, Féret JB, Foody GM, He KS, Lausch A, Nagendra H, Wegmann M, Pettorelli N (2016) Satellite remote sensing to monitor species diversity: potential and pitfalls. *Remote Sens Ecol Conserv* 2:25–36. doi:10.1002/rse2.9
- Walsh JL (1923) A closed set of orthogonal functions. *Am J Math* 45:5–24
- Zadeh L (1965) Fuzzy sets. *Inf Control* 8:338–353

**Part IV**  
**Communicating and Informing Decisions**

# Chapter 14

## Communicating and Disseminating Research Findings

Amber E. Budden and William K. Michener

**Abstract** This chapter provides guidance on approaches and best practices for communicating and disseminating research findings to technical audiences via scholarly publications such as peer-reviewed journal articles, abstracts, technical reports, books and book chapters. We also discuss approaches for communicating findings to more general audiences via newspaper and magazine articles and highlight best practices for designing effective figures that explain and support the research findings that are presented in scientific and general audience publications. Research findings may also be presented verbally to educate, change perceptions and attitudes, or influence policy and resource management. Key topics include simple steps for giving effective presentations and best practices for designing slide text and graphics, posters and handouts. Websites and social media are increasingly important mechanisms for communicating science. We discuss forms of commonly used social media, identify simple steps for effectively using social media, and highlight ways to track and understand your social media and overall research impact using various metrics and altmetrics.

### 14.1 Introduction

The ingredients of good science are obvious—novelty of research topic, comprehensive coverage of the relevant literature, good data, good analysis including strong statistical support, and a thought-provoking discussion. The ingredients of good science reporting are obvious—good organization, the appropriate use of tables and figures, the right length, writing to the intended audience—do not ignore the obvious. Bourne (2005)

Researchers communicate their findings for several reasons. First and foremost, basic and applied researchers strive to enhance scientific knowledge; communicating and disseminating new information and knowledge represent a cornerstone of the scientific process. Second, many researchers focus on communicating research findings that can positively contribute to improved natural resource management, conservation and decision-making. Third, many researchers are motivated to

---

A.E. Budden (✉) • W.K. Michener  
University of New Mexico, Albuquerque, NM, USA  
e-mail: [aebudden@dataone.unm.edu](mailto:aebudden@dataone.unm.edu); [william.michener@gmail.com](mailto:william.michener@gmail.com)

educate the next generation of scientists and to increase public awareness through broad communication. Finally, all scientists, to varying degrees, communicate and disseminate their findings so that they may be recognized for their contributions to science; such contributions may be characterized by number of citations and impact factor, altmetrics, impact on resource management and decision-making, as well as their influence on tenure and promotion decisions.

Research findings may be communicated and disseminated using various types of media to reach different audiences and to achieve different objectives. Written communications such as scholarly articles, technical reports, abstracts, books and textbooks, newspaper and magazine articles, blogs, infographics, posters, and website content are used to convey research findings to both technical and lay audiences. Some media such as scholarly publications and technical reports are frequently aimed at more expert audiences, whereas textbooks are focused on particular age ranges and educational levels (e.g., high school, college) and newspaper articles are aimed at the broad public and usually written at a middle school level (e.g., 6th through 8th grade). Research findings may also be communicated verbally at professional society meetings via talks and poster presentations; public meetings (e.g., lectures, Town Halls); television and radio interviews, podcasts and videos; and webinars. Scientific content that is presented verbally at conferences and meetings may also be disseminated via recorded videos (e.g., YouTube, vimeo) and by sharing slide and poster presentations (e.g., slideshare). Research findings may also be embedded in data, tables and illustrations that are preserved and discoverable through archives, data directories and data aggregators (Cook et al. 2017; Michener 2017). Last, many scientists keep current with the newest research findings through social media such as twitter and facebook.

In this chapter, we describe approaches and best practices for communicating and disseminating research findings via publications and online resources aimed at scientific and general audiences (Sect. 14.2), presentations (Sect. 14.3), and social media (Sect. 14.4). We conclude with a description of metrics and altmetrics and how these tools are used to measure the impact of research findings (Sect. 14.5).

## 14.2 Publishing Research Findings

Research findings are most frequently and directly communicated to one's peers via scholarly publications. These publications typically undergo peer review to assure their value and have been the traditional method of research communication since the first scientific journal, *Philosophical Transactions of the Royal Society*, was founded in 1665 (Kronick 1976).

Synthesized research findings are communicated via textbooks and books to students and others that are interested in detailed treatments of particular topics. Highlights of research findings are often presented to more general-interest audiences via newspaper and magazine articles. Below, we present some of the best practices associated with communicating findings via scholarly publications,

general audience outlets and illustrations. Additional details pertaining to writing non-fiction and technical documents can be found in books by Strunk and White (1999), Zinsser (2006), Alred et al. (2011), and the University of Chicago Press Staff (2010).

### ***14.2.1 Scholarly Publications***

Scholarly publications include peer-reviewed journal articles, abstracts, technical reports and books and book chapters. Original research findings frequently first appear in journal articles, abstracts, and technical reports. Books and book chapters often take longer to appear in publication form and generally include findings that are synthesized from numerous studies.

#### **14.2.1.1 Journal Articles**

Individual journals normally have specific publication requirements that address page limitations, text content and format (e.g., required sections), table and figure guidelines, and citation style. Table 14.1 describes ten simple rules for writing research papers as defined by Zhang (2014). The rules include many principles and practices that, if followed, can increase the likelihood that a research paper will be published, interesting and impactful.

#### **14.2.1.2 Abstracts**

Research findings may also be presented in abstracts. Such abstracts are often published prior to or following scientific meetings and are associated with posters and presentations that are given at the meetings. Publishers and scientific societies often have very specific abstract guidelines that authors are required to follow such as word limits (e.g., 200–500 word limits) and specific paragraph content (e.g., methods, key findings). Recommendations for writing good abstracts differ from those associated with writing research papers. Weinberger et al. (2015) analyzed one million abstracts from numerous scientific domains and found that citation rates were positively associated with longer and more detailed abstracts containing prolix prose and statements that signified the novelty and importance of the work. In addition, highly cited abstracts frequently contained superlatives, pleasant words, active words and words that easily evoked images. In contrast to scientific articles where conciseness is a virtue, authors of abstracts benefit from using all space at their disposal to highlight novel findings and to emphasize the importance of their work.

**Table 14.1** Ten simple rules for writing research papers [adapted from Zhang (2014)]

Number	Rule	Description
1	“Make it a driving force”	“Design a project with an ultimate paper firmly in mind”
2	“Less is more”	“Fewer but more significant papers serve both the research community and one’s career better than more papers of less significance”
3	“Pick the right audience”	“This is critical for determining the organization of the paper and the level of detail of the story, so as to write the paper with the audience in mind.”
4	“Be logical”	“The foundation of “lively” writing for smooth reading is a sound and clear logic underlying the story of the paper.” “An effective tactic to help develop a sound logical flow is to imaginatively create a set of figures and tables, which will ultimately be developed from experimental results, and order them in a logical way based on the information flow through the experiments.”
5	“Be thorough and make it complete”	Present the central underlying hypotheses; interpret the insights gleaned from figures and tables and discuss their implications; provide sufficient context so the paper is self-contained; provide explicit results so readers do not need to perform their own calculations; and include self-contained figures and tables that are described in clear legends
6	“Be concise”	“The delivery of a message is more rigorous if the writing is precise and concise”
7	“Be artistic”	“Concentrate on spelling, grammar, usage, and a “lively” writing style that avoids successions of simple, boring, declarative sentences”
8	“Be your own judge”	Review, revise and reiterate. “. . . put yourself completely in the shoes of a referee and scrutinize all the pieces—the significance of the work, the logic of the story, the correctness of the results and conclusions, the organization of the paper, and the presentation of the materials.”
9	“Test the water in your own backyard”	“. . . collect feedback and critiques from others, e.g., colleagues and collaborators.”
10	“Build a virtual team of collaborators”	Treat reviewers as collaborators and respond objectively to their criticisms and recommendations. This may entail redoing research and thoroughly re-writing a paper.

### 14.2.1.3 Technical Reports

Technical reports are generally used to convey technical information about a topic or project to a particular audience in a clear, well-organized format. Examples include project reports that summarize a project’s progress and findings, literature reviews, and authoritative syntheses of the state-of-knowledge about a specific topic. Such reports may be targeted to research sponsors, knowledgeable colleagues and peers, and, in some cases, decision-makers and the public. Technical reports

**Table 14.2** Common components of a technical report

Component	Description
Title page	Brief descriptive title of the report; may include authors, date, citation, etc.
Executive summary, summary or abstract	Summarizes major findings of the report; an executive summary provides a high-level summary of the report and is typically short in length (e.g., 1–5 pages)
Table of contents, list of figures, list of tables	Presents the document’s structure and indicates where specific sections, figures and tables may be found
Introduction	Helps the reader understand the structure and content of the report and, often, the reason the report was written
Section(s)	The main body of the report normally includes one or more sections that are often titled and numbered (e.g., background, methods, results, conclusions, recommendations, etc.)
Acknowledgments	Brief paragraph that acknowledges sponsors and other contributors
References	Literature cited in the report
Appendices	Detailed information such as raw or summarized data, survey forms, scientific code, and design specifications are frequently included in one or more appendices

written for research sponsors, agencies and professional societies must often follow specific requirements with respect to format, length and content. Generally, though, technical reports include most or all of the components described in Table 14.2. The same rules pertaining to writing research papers presented in Table 14.1 (especially rules 3–8) also apply to writing technical reports. Also, see Alred et al. (2011) for recommendations about writing good technical reports.

#### 14.2.1.4 Books and Book Chapters

Books and book chapters may be written for general or expert audiences and may either be self-published or published by a university press or commercial publisher. Publishers frequently require that a book proposal be submitted and be peer-reviewed. Such proposals typically follow a specific format and include information about the intended audience, a detailed outline of the book, anticipated length and number of tables and figures, and learning objectives if the book is intended for the textbook market. The outline is particularly important as it highlights how content is structured and, ideally, demonstrates that the author(s) have envisioned a logical flow for the presentation of the material. If a book proposal is approved, the authors are normally provided with specific guidelines for formatting text, tables, figures, references and other elements.

### ***14.2.2 Newspaper and Magazine Articles for General Audiences***

Researchers rarely write newspaper and magazine articles. Instead, they are more likely to be interviewed by a reporter who is writing an article about a particular area of research or a new research finding. The goal of the reporter or author is typically to convey important and interesting facts to the reading public in a brief, word-limited story, although some articles may be opinion pieces that are designed to sway public opinion to a particular point of view. Before granting an interview, it is good practice to determine whether the publication is reputable and whether or not the piece is designed to inform readers or editorialize a position that you may or may not support.

Regardless of the objective, newspaper and magazine articles are normally written in a style that differs markedly from how scientific articles are written. Scientific articles follow a logical progression where the author: (1) introduces key hypotheses or a problem statement along with background information in the introduction; (2) describes how the research was conducted in the materials and methods section; (3) presents basic findings and supporting data in the results section; and (4) explains the major findings and their significance in the discussion section. In contrast, newspaper articles include the most notable finding or conclusion in the lead paragraph. The remainder of the article contains supporting facts, quotes and anecdotes that pertain to what was discovered or is known about the topic, who did the work and when and where they did it and, possibly, why and how the work was done.

As a researcher, your primary objective may be to ensure that your research findings are presented in an accurate and compelling way to readers, and that they may positively impact conservation, resource management, policy, education or decision-making. There are several ways to do this. First, it is always a good idea to plan out what you want to say to the reporter. You may even ask them to send you a list of questions in advance of the interview. Jot down answers to the what, who, when, where, why and how of the research. Also, compose one or more brief, memorable quotes that explain what is noteworthy about the research and possibly include an anecdote that may help the reporter and readers relate to the finding or discovery (e.g., how this research impacts the lives of the readers). Second, provide the reporter with factual responses to their questions and offer to clarify any confusing or complex points during the interview. Third, offer to review the article before it goes to press and answer any subsequent questions that arise. A little preparation work can lead to good press coverage that will further promulgate your research findings and discoveries.

### ***14.2.3 Designing Effective Figures***

Effective graphics and illustrations help explain and support the concepts and research findings that are presented in scientific and general audience publications. Many good references provide guidance on creating effective graphics and



illustrations (e.g., Few 2012; Robbins 2013; Tufte 1983, 1990, 1997; Wong 2013). Here, we present general guidance for creating good figures, highlight some of the tools that can be used and provide examples of effective figures.

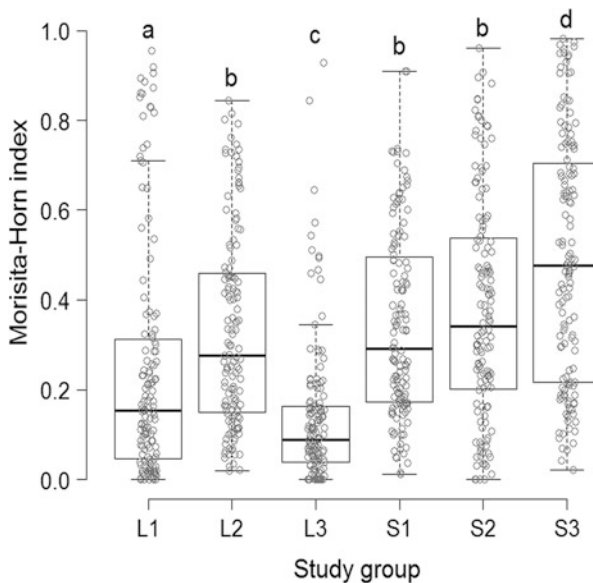
Table 14.3 lists ten simple rules for creating better figures. Tailoring the message to the audience is as important to creating a good illustration (Table 14.3, Rule 1) as it is to writing a good research paper (Table 14.1, Rule 3). Readers of a scientific journal will wish to see a figure that conveys all information relevant to a key point or finding such as error bars that allow one to judge the significance of the results. Illustrations for students and the general public can be more effective if they are simpler, contain fewer details, and include explanations of the most salient points. Rules 2 and 9 focus on the importance of having a clear and understandable

**Table 14.3** Ten simple rules for creating better figures [adapted from Rougier et al. (2014)]

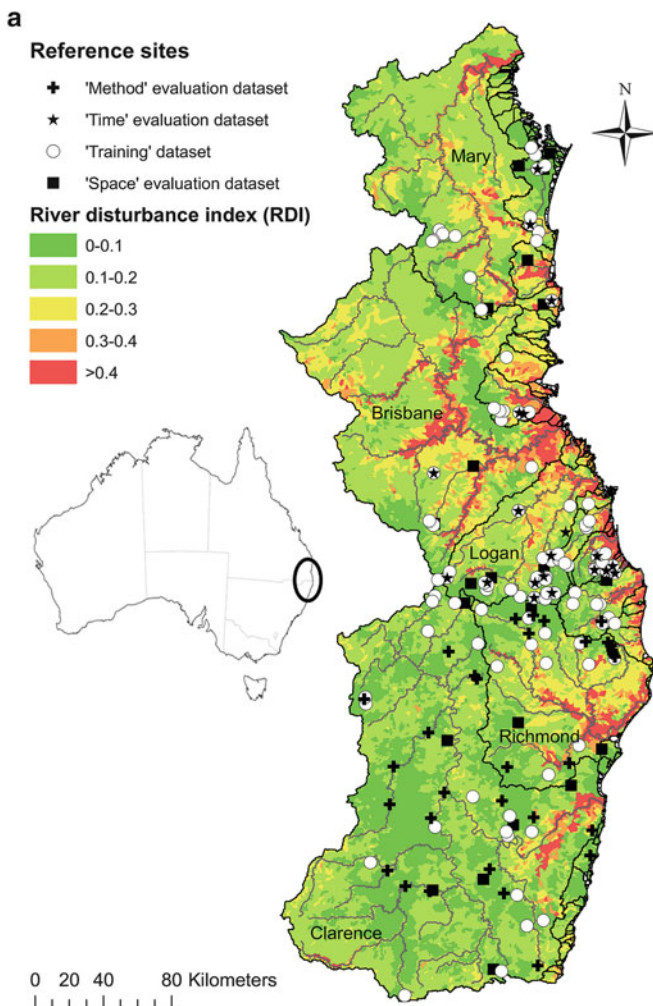
Number	Rule	Description
1	“Know your audience”	“. . . identify, as early as possible in the design process, the audience and the message the visual is to convey.”
2	“Identify your message”	Clarify the underlying message that you wish to convey in a figure and how a figure can best serve that purpose.
3	“Adapt the figure to the support medium”	Adapt the figure content (e.g., amount of detail) and presentation style (e.g., degree of complexity, color contrast, line thickness) to the medium (e.g., journal vs newspaper article, poster, presentation, web page).
4	“Captions are not optional”	Figures should be accompanied by captions that explain how to interpret the figure and that provide other details that may not be included directly in the figure (e.g., probability values).
5	“Do not trust the defaults”	Default settings (e.g., choice of font, line thickness, colors, tick marks) should be manually adjusted for a specific type of plot.
6	“Use color effectively”	“If you decide to use color, you should consider which colors to use and where to use them.” Color can be used to enhance a message, but use of too many similar colors causes color blindness.
7	“Do not mislead the reader”	“As a rule of thumb, make sure to always use the simplest type of plots that can convey your message and make sure to use labels, ticks, title, and the full range of values when relevant.”
8	“Avoid “Chartjunk””	Avoid “unnecessary or confusing visual elements . . . that do not improve the message (in the best case) or add confusion (in the worst case). For example, chartjunk may include the use of too many colors, too many labels, gratuitously colored backgrounds, useless grid lines, etc.”
9	“Message trumps beauty”	“Remember, in science, message and readability of the figure is the most important aspect while beauty is only an option.”
10	“Get the right tool”	Many tools exist that can facilitate the creation of good figures and save time.

message whereas Rules 3–8 provide guidelines for designing and creating effective figures. Rule 10 highlights the benefits of using an appropriate tool that can do the desired job and save time in the process. Many open-source (e.g., GRASS, Matplotlib, QGIS, R, VisTrails; Hampton et al. 2015; Rougier et al. 2014) and commercial (e.g., JMP, MATLAB, OmniGraffle, SAS, Tableau) software packages can be used to create good figures.

Figures 14.1, 14.2, 14.3 and 14.4 exemplify many of the best characteristics of good illustrations identified in Table 14.3. Figure 14.1 is from a study by Chaves and Bicca-Marques (2016) that tested the hypothesis that brown howler monkeys (*Alouatta guariba clamitans*) can adjust their diet in response to changes in resource availability by comparing the diet of six free-ranging groups inhabiting three small and three large forest fragments in southern Brazil. The box-whisker plot figure is an information-rich black and white illustration that enables one to see the spread and skewness of the data in each of six groups. The figure shows all data values and indicates significant differences between the study groups (as indicated by the letters). Overall, the figure provides a large amount of information about the distribution of data within each of the study groups and allows one to better visualize how significant the differences are between groups. In addition, the legend is clear and concise and aids in interpretation by the viewer.

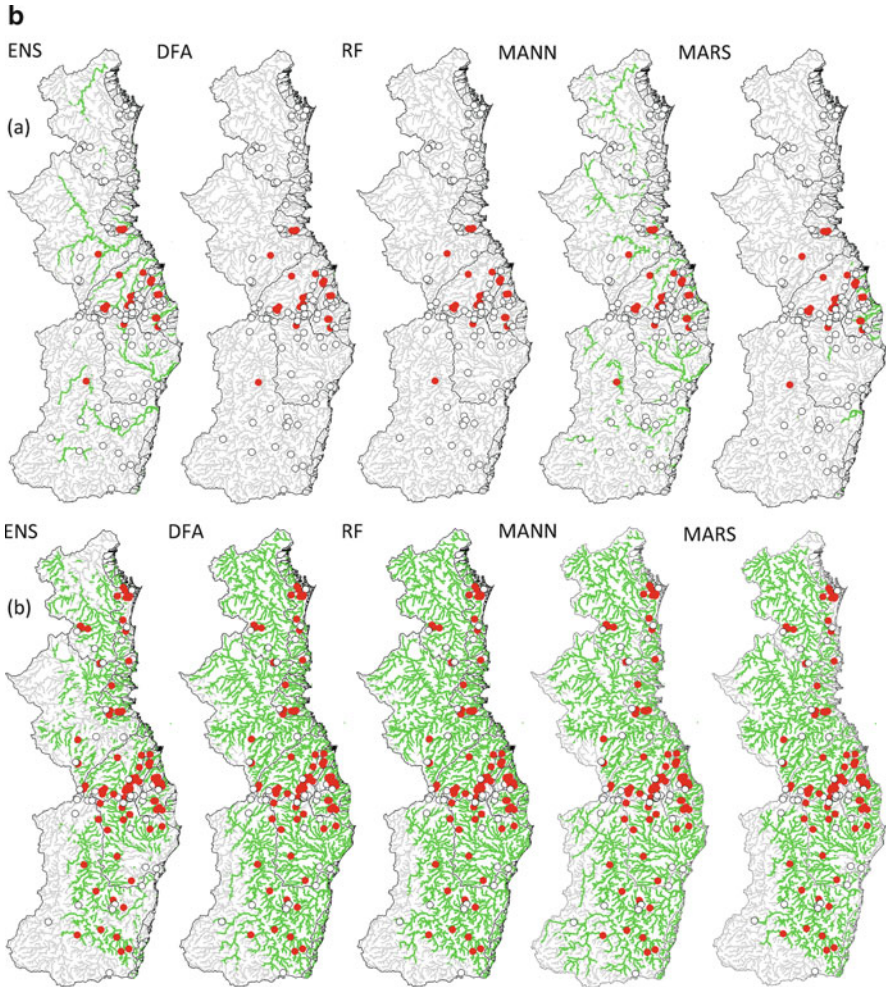


**Fig. 14.1** Figure legend quoted directly from Fig. 3 (Chaves and Bicca-Marques 2016): Fig. 3. Intermonth diet similarity between study groups inhabiting small and large fragments. The *line within a box* represents the median of the Morisita-Horn index, the *box* represents the 25% and 75% interquartiles (IQR), and the *whiskers* represent the IQR multiplied by 1.5. *Dots* represent the actual data points for each group. *Different letters* indicate significant differences ( $P < 0.05$ ). doi:10.1371/journal.pone.0145819.g003



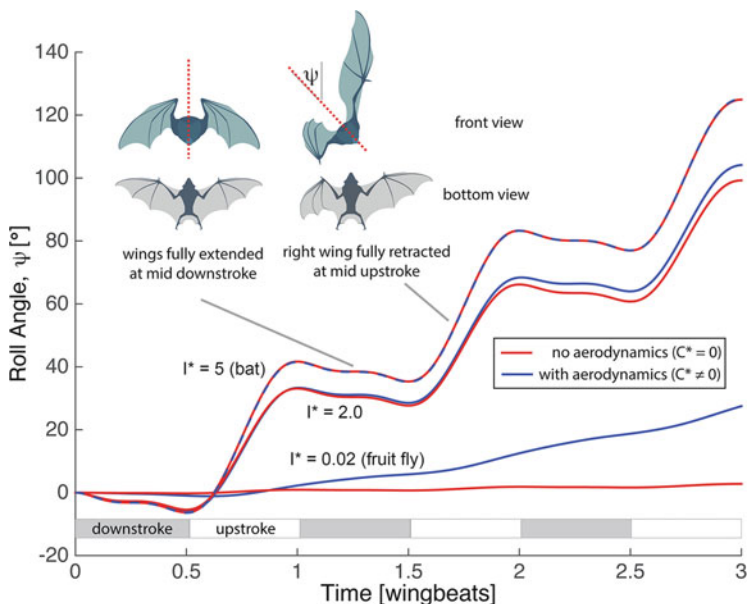
**Fig. 14.2 (a)** Figure legend quoted directly from Fig. 1 (Rose et al. 2016): Fig. 1. Reference site locations for each dataset used, and major river systems in the study area. The river disturbance index (RDI—see [47] for details of its derivation) provides context for the ‘least disturbed’ reference sites; low RDI values indicate low levels of human pressures in the upstream catchment. Note that the season dataset is represented by all ‘training’ sites in the SEQ section of the study area. doi:10.1371/journal.pone.0146728.g001.

**(b)** Figure legend quoted directly from Fig. 2 (Rose et al. 2016): Fig. 2. Projected species distributions (at a cut-off threshold of 0.5) for **(a)** *Hypseleotris klunzingeri* and **(b)** *Melanotaenia duboulayi*. Green stream segments are predicted presences; grey segments are predicted absences. The circles are sites that were sampled in autumn/winter 2013 (i.e. the training and space datasets; n = 128). Red circles are observed presences, open circles are observed absences. ENS—Single species ensemble model; DFA—RIVPACS community model using a discriminant function classifier; RF—RIVPACS model using a random forest classifier; MANN—Multi-species response artificial neural network model; MARS—Multi-species response multivariate adaptive regression splines model. doi:10.1371/journal.pone.0146728.g002



**Fig. 14.2** (continued)

Figure 14.2 combines two figures from a study by Rose et al. (2016) that tested the accuracy of five species distribution models for predicting fish assemblages at reference stream segments in coastal subtropical Australia. Figure 14.2a is a clear, uncluttered map that shows the location of the study area in Australia, depicts four different types of reference sites using easily distinguishable symbols, and highlights five different categories of river disturbance that range from least disturbed (forest green and light green) to most disturbed (yellow, orange and red). The figure includes key cities, a 4-point compass rose, and a bar scale for reference. Figure 14.2b illustrates the results of the study and shows stream segments where each of the two species modeled (*Hypseleotris klunzingeri* and *Melanotaenia duboulayi*) were predicted to be present (green) or absent (grey). In addition, filled



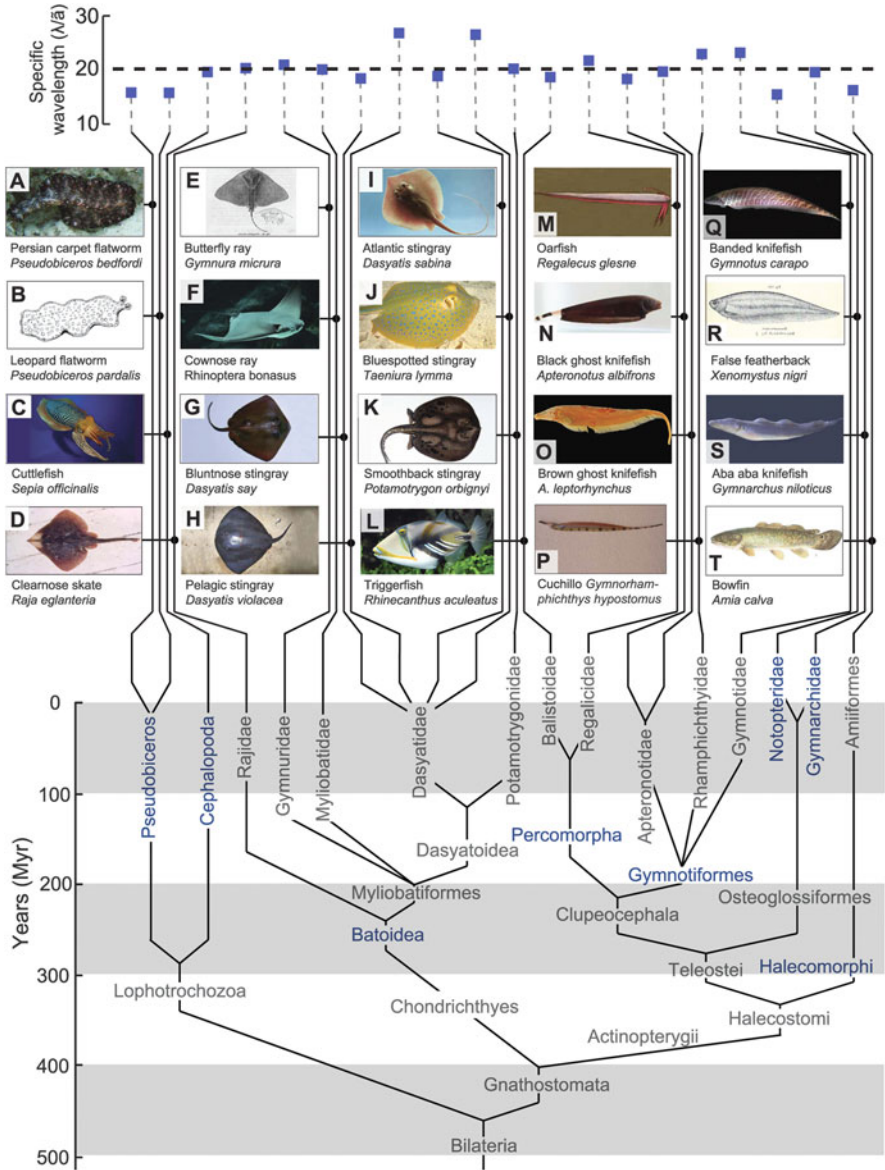
**Fig. 14.3** Figure legend quoted directly from Fig. 4 (Bergou et al. 2015): Fig. 4. Minimal model of bat dynamics applied to body roll maneuver. Both wings are fully extended ( $e_r = e_l = 1$ ) at mid-downstroke, while one wing is fully retracted ( $e_r = 0$ ) at mid-upstroke. For morphological parameters matched to those of *C. perspicillata* ( $I^* = 5$ ), simulations show that this asymmetric wing extension induces body roll and that aerodynamic forces do not influence the motion significantly. The response is insensitive to modest changes in the relative wing inertia ( $I^* = 2$ ), although when the morphological parameters are matched to those of fruit flies ( $I^* = 0.02$ ,  $C^* = 0.05$ ), aerodynamic forces dominate, while inertial forces have minimal effect on the body orientation. MATLAB code available in file `minimal_simulation.zip` from the Dryad Digital Repository, doi:10.5061/dryad.21qs5 [31]. doi:10.1371/journal.pbio.1002297.g004

red circles indicated observed presences at reference sites and open circles indicated predicted absences. Combined, the two figures provide the reader with substantial information about the study area and sampling strategy as well as a highly effective colored illustration that highlights the differences and similarities among the different models and species.

Figure 14.3 illustrates the results of a bat dynamics model developed by Bergou et al. (2015) that shows that bats perform body rolls by selectively retracting one wing during the flapping cycle and that the complex maneuver does not rely on aerodynamic forces. The figure combines a graph that shows roll angle over time under three different morphological conditions with a bar that highlights downstroke and upstroke periods. In addition, front and bottom illustrations of bats with their wings extended at downstroke and retracted at upstroke aid the viewer in interpreting model results. The figure is notable for its judicious use of color, inclusion of bat sketches and comprehensive legend.

Figure 14.4 is from a study of convergent evolution by Bale et al. (2015) that demonstrates that an optimal method of swimming has evolved independently at





**Fig. 14.4** Figure legend quoted directly from Fig. 1 (Bale et al. 2015): Fig. 1. Undulatory median/paired fin swimmer phylogenetic relationships and  $SW = \lambda/\bar{a}$ , where  $\lambda$  and  $\bar{a}$  are wavelength and mean amplitude of undulations present along the fin, respectively. The eight instances of independent emergence of elongated median/paired fin swimming are highlighted in blue. The SW of these organisms and sources of the data are also tabulated in the S2 Table. Not shown here because of space constraints is the SW for the ray *Dasyatis americana*, which has an SW of 25.1, and the weakly electric knifefish *Eigenmannia virescens*, which use two counter-propagating waves on their fin during slow speed swimming and have an average SW of 17.7. See S4 Table. The following images are licensed under CC-BY: (c) *Sepia officinalis* image courtesy of Hans Dappen.

least eight times in vertebrate and invertebrate swimmers across three different phyla. The figure is very effective at showing when the swimming trait emerged and in which organisms; pictures of example organisms are included as well as the Optimal Specific Wavelength (i.e., measure of fin undulations). The figure is notable because of its clear presentation of phylogenetic relationships, inclusion of functional data and the addition of pictures of the organisms to aid the reader in understanding the paper's findings.

### 14.3 Communicating Research Findings Outside of Publications

Communication is more than simply presenting material; communication is defined by an interaction between the 'sender', or presenter/author, and the 'receiver', or audience. Communication is effective when the message elicits a desired response by the receiver. In communicating research findings, the desired response may be a change in policy or practice, continued or increased funding, or increased understanding of the research area. If this is not achieved then communication has largely failed. It is critical to both understand the goals of your presentation and to know your audience. This is true for both written and verbal communication. For verbal communication, there is an added dimension of environment. Authors have little control over the environment where their printed material is read. However, the presenter must consider and respond to the environment for verbal presentations to be effective.

In this section we discuss steps for creating an effective presentation, highlight best practices for designing slides and supporting materials such as handouts, and provide tips for creating appealing, information-rich poster presentations.

#### 14.3.1 Simple Steps for Giving an Effective Presentation

The basis of a good presentation lies in storytelling. Presentations are a ubiquitous, default form of communication in scientific research and other domains, and are largely treated as such in their preparation and execution. They can be dull. They



**Fig. 14.4** (continued) (d) *Raja eglanteria* image courtesy of George Burgess. (f) *Rhinoptera bonasus* image courtesy of Juan Aguerre. (j) *Taeniura lymma* image courtesy of Nicolai Johannesen. (m) *Regalecus glesne* image courtesy of Sandstein. (n) *Apteronotus albifrons* image courtesy of Clinton and Charles Robertson. (o) *Apteronotus leptorhynchus* image courtesy of the Harvard Museum of Comparative Zoology. (p) *Gymnorhamphichthys hypostomus* image courtesy of Mark Sabaj with support from IXingu Project (NSF DEB-1257813). (s) *Gymnarchus niloticus* image courtesy of Masashi Kawasaki. All remaining images are public domain. doi:[10.1371/journal.pbio.1002123.g00](https://doi.org/10.1371/journal.pbio.1002123.g00)

include facts, statistics and other information but may do so without a compelling story. Storytelling is the ability to weave details into a compelling narrative and create emotional connections. It is common across all cultures. Stories are the most powerful art form or tool for delivering information (Duarte 2010). Becoming a storyteller to more effectively communicate information may seem daunting, but you don't need to be an outstanding orator. You simply need to be authentic and to make a human connection.

Knowing your audience is central to delivering an impactful presentation (Bourne 2007, Table 14.4): What grade level will you be talking with? Is it a specialized audience familiar with your field? Are these decision makers? What are they motivated by? Where does their interest lie? Make your presentation about them and not about you: Why would your research interest them? How does this

**Table 14.4** Ten simple rules for good oral presentations [adapted from Bourne (2007)]

Number	Rule	Description
1	“Talk to the audience”	This is akin to knowing your audience. “Prepare presentations that address the target audience”.
2	“Less is more”	“Your knowledge of the subject is best expressed through a clear and concise presentation”. You will have the opportunity to go into details in response to questions.
3	“Only talk when you have something to say”	Be realistic about what you will have available to present and do not take advantage of the audience's time by presenting “uninteresting preliminary material”
4	“Make the take-home message persistent”	Aim for the audience to be able to remember three key points a week after the presentation.
5	“Be logical”	Set your presentation up like a story. “There is a logical flow—a clear beginning, middle and an end.”
6	“Treat the floor as a stage”	“Presentations should be entertaining” but remain true to yourself and know your limits. “A good entertainer will captivate the audience”.
7	“Practice and time your presentation”	This will become easier but it is always good to practice with a friendly audience. “An important talk should not be given for the first time to an audience of peers. You should have delivered it to your research collaborators who will be kinder and gentler”
8	“Use visuals sparingly but effectively”	Presentation styles and the need for visuals vary. Practice will help refine how many visuals you need. “A useful rule of thumb [is] one visual for each minute you are talking.”
9	“Review audio and/or video of your presentations”	This will help you easily see where you can improve your presentation.
10	“Provide appropriate acknowledgements”	You may “acknowledge people at the beginning or at the point of their contribution so that their contributions are very clear”.



relate? Why should they be invested in the outcome? Ultimately, the goal is for your audience to embrace your message and this requires them to engage with it. Authenticity is required. Build upon the emotional connections developed in storytelling and do not distance your audience by using jargon, heavy-handed visuals and other crutches. Knowing your audience is the first of ten simple rules for good oral presentations.

Storyboarding is critical for preparing a presentation. Storyboarding is a concept that is derived from film production and means to use drawings (or text notes) that represent critical concepts and elements from your story to build the flow of your presentation. Storyboarding enables you to streamline your material, focus only on the critical information, ensure that your message is persistent and stands out, and maintain a logical flow (i.e., rules 2–5 in Table 14.4). Storyboarding can be done in many different formats but you should be able to easily edit, cut and revise the story elements; thus, working outside of presentation software is often an advantage. Sticky notes or index cards, for example, can readily be physically rearranged, enabling you to quickly change your narrative without being encumbered by software (Duarte 2008).

When constructing a storyboard, adhering to a form that contains a ‘beginning’, ‘middle’ and ‘end’ is more likely to result in a successful presentation and one that provides the logical flow identified by Bourne (2007). Duarte (2010) draws from literary and cinematic structures and refers to the transitions between the beginning and middle as a “call to adventure”. This call to adventure creates a perceived imbalance by stating ‘*what could be*’ versus ‘*what is*’. ‘*What could be*’ is the desired outcome that a body of research is moving us towards and ‘*what is*’ reflects the current status or knowledge in the field. Likewise, Duarte (2010) refers to the transition between middle and end as “a call to action” where the presenter articulates the finish line that the audience is to cross, whether it be a figurative or literal call to action.

### ***14.3.2 Best Practices for Slides***

The previous section introduced how to give an effective presentation and largely focused on story development. Few speakers will give an oral presentation without visual aids and PowerPoint or Keynote are widely used to generate supporting graphics. In developing your storyboard and identifying the take-home messages and transitions, it will quickly become apparent which sections can be best supported with visuals as well as the type of visual information needed to elucidate your points. Visuals should be used sparingly but effectively (Rule 8, Table 14.4; Bourne 2007). In this section we provide some basic guidance on how to design effective slides.

### 14.3.2.1 Slide Design

The arrangement and organization of slides can have a significant impact on whether the message of your presentation is clearly communicated. This is true not just in terms of the story arc across a set of slides but also with respect to elements within individual slides. Text, figures or images may occur alongside other slide elements and careful design can maximize the clarity of the slide (Duarte 2008). In particular, attention should be paid to:

1. **Contrast:** Contrast helps to establish relationships between elements. By highlighting something as different through color, shape, size, etc., the audience can quickly understand that the item warrants attention. For example, using bold words within a sentence is a common way to employ contrast and bring attention to the significant text. Use contrast sparingly to create notable differences.
2. **Flow:** A western reading pattern flows left to right, top to bottom and this is the way most viewers will process slides. Slides should follow this convention for ease of interpretation. If it is necessary to deviate from this pattern, you should use cues to direct the audience.
3. **Hierarchy:** Hierarchies enable the audience to readily interpret relationships, or order, in elements. An example of this is the title font being larger than the main text, which may be larger than supporting text. Another is the use of bullets. Ensure your font size choices are intentional since variation in text size infers meaning.
4. **Unity:** Adopting an underlying grid structure to your slides allows for consistency in information presentation. Verify that items are aligned, images, text or graphics are consistently placed and the transition between slides is not distracting. A unified structure also creates a more organized or branded presentation.
5. **Proximity:** As with hierarchy and text size, the spatial relationship of elements to one another carries meaning and elements should therefore be placed intentionally.

### 14.3.2.2 Text Slides

Distill information down to the most salient points when representing a textual concept or series within a slide. Include only key concepts that anchor the audience to your narrative since the audience will be listening to you at the same time they are reading the slide. Impact statements (single phrases occupying a whole slide) or bulleted points might be appropriate depending on the nature of the information. Duarte (2008) argues that there are no set rules for the amount of text or number of bullets that should appear on a slide. Rather, bullets should be treated as headlines and used sparingly whereas sub-bullets should be avoided. Information should be reduced down to key points so that bullets serve as mnemonics for the narrative (Duarte 2008). Bullets can sometimes be replaced with images; however, practice

of the accompanying narrative and repetition are critical (Rule 7, Bourne 2007; Duarte 2008).

Sans serif fonts are more legible when reading at a distance and are preferred for presentation purposes. There are many fonts to choose from and different fonts have different personalities (Cho 2013). It is best to include no more than two font types (Duarte 2008). Additional emphasis can be achieved with color, weight and italics to add variation without the introduction of a third font type. With respect to font size you should try to stay above 28 pt. A good test is to put your file into slide sorter view and look at the slides at 66 percent size. If you can still read them, so can your audience (Duarte 2008).

### 14.3.2.3 Graphics

Figures in slide presentations must be read and interpreted significantly faster than those in a paper publication. So while the guidance presented previously for designing effective figures holds true for published material, additional consideration is required for slide presentation. Rather than provide all details of your research within a figure, first identify the intended conclusion that you want the audience to reach and then make sure that this message stands out during the graphic design. This can be achieved by choosing the appropriate tool or chart type, keeping the chart simple or free from ‘chartjunk’ (Tufte 2003, 2006) and using graphical elements to emphasize key data or points.

Sullivan (2011) and Few (2009) provide information on the limitations of particular chart types and demonstrate how to simplify charts so that the important information stands out. Duarte (2008) extends these concepts to PowerPoint presentations and provides ‘make-overs’ for an example pie chart, vertical bar chart and horizontal bar chart [Fig. 14.5; see also Robbins (2016) and Few (2016)]. Some consistent guidelines include: avoid 3D graphics; de-emphasize or remove non data elements (such as gridlines); be consistent in the use of color; order bars by size versus alphabetically for easier comparison; choose an aspect ratio that shows variation in the data; and use visually prominent elements to emphasize key points. Note the use of contrast in Fig. 14.5b, c to highlight the important message.

Chart design should also take into consideration the way in which we decode graphical information. For example, shape, color, position, etc. can be used to compare or contrast data points. However, the ability to perceive these differences varies. Cleveland and McGill (1985) identified the order in which people are able to most accurately compare information (Table 14.5). Position along a common scale is easiest for viewers to judge. For example, imagine two points in a scatter plot. We are quickly able to see which point is farther from the axis than the other. Cleveland and McGill (1985) found that judging differences in these elements was more accurate than judging the length of bars within a stacked chart, for example. They observed that our ability to perceive differences in angle (*pie charts*) and slope (*regression lines*) was tied and ranked midway in their list of graphical attributes.

Before

a)

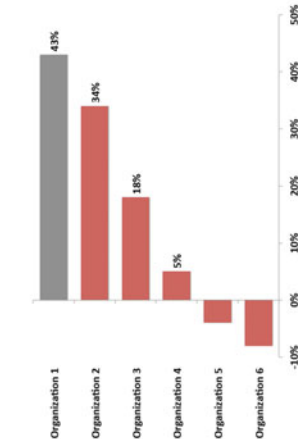
Primary language (2016)



After



Primary Language (2016)



Before

a)

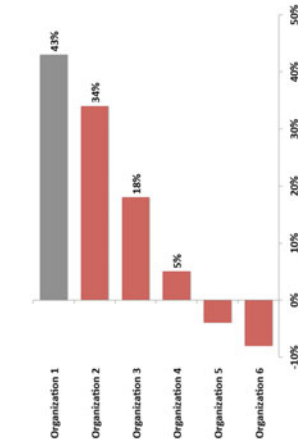
Primary language (2016)



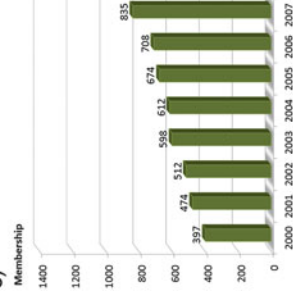
After



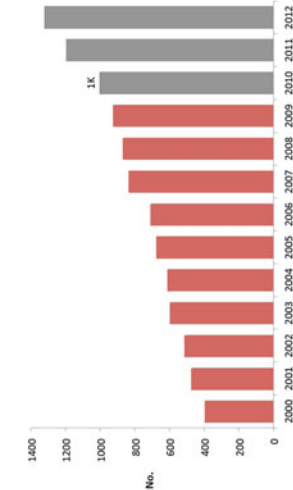
Primary Language (2016)



b)



c)



**Fig. 14.5** Example cleaning of typical pie, horizontal bar and vertical bar charts adapted from Duarte (2008). In all cases the figure title has been omitted to focus on the graphical elements. (a) 3-dimensional pie charts tend to give more prominence to the data in the foreground. (b) Too many colors can obscure the focal point. The horizontal scale and single color change help emphasize growth of the target organization. (c) Depth can visually skew the data and has been removed. Color was added to create a distinct take-away message; the year that membership reached 1000

**Table 14.5** Ordering of elementary graphical perception tasks by element [adapted from Cleveland and McGill (1985)]

Rank	Aspect judged	Example
1	Position along a common scale	Scatter plot
2	Position on identical but non-aligned scales	Multiple scatter plots
3	Length	Bar chart
4	Angle; slope (tied)	Pie chart, regression line
5	Area	Bubble chart
6	Volume; density; color saturation (tied)	Heatmap
7	Color hue	Multiple

Differences in hue (*color*) were the most difficult to discern. Graphs should exploit the highest ranking attributes as much as possible.

### 14.3.3 Handouts

Both the presenter and the audience can benefit from the judicious use of handouts. Handouts provide an opportunity to go into greater detail, enabling presenters to keep the presentation slides as simple as possible. Handouts also allow the presenter to leave the audience with a reminder of the presentation and provide contact information for future discussion opportunities. Audience members can listen to the presentation without the need to take notes and are presented an opportunity for follow-up. Handouts should be designed explicitly for the audience to augment the material presented and should not simply repeat the presentation. Some generic best practices for creating effective handouts (adapted from Witt 2016) include:

1. **Simplicity:** While a handout enables you to provide additional detail, it should focus on the key concepts of your presentation. You should also aim to keep the handout short but this may vary with the length of your presentation.
2. **Relationship to presentation:** Create the handout using the same story format as the presentation. You do not need to cover all the slides presented but the audience should be able to follow the handout sequentially with the talk.
3. **Provide additional detail:** Use the handout to expand on points, including additional figures as needed. Consider that your handout may need to stand alone if someone is reading it months after your presentation. A handout can also be used to provide full citations for references.
4. **Visually appealing:** As with presentations, follow good style guidelines for handouts. Use appropriate fonts, allow for white space in the design, ensure that break points are logical and that proximity, hierarchy and contrast are used appropriately.

5. Know when to distribute: If your handout is intended for the audience to follow along and take notes, or allow for deeper understanding of the material, distribute it before the presentation. However, if it is intended primarily as a means to remind them of the talk and provide your contact information, you can distribute it after your presentation.
6. Allow ample time: A good handout will be designed alongside the presentation so that as you reduce content in your slides, you will better understand what needs to be provided in a handout. The handout should not be an afterthought.

For researchers that work with RStudio or in LaTeX, Edward Tufte's (2016) distinctive clean graphical style has been converted into programs that can help structure Tufte-esque handouts (CTAN 2016; R Markdown 2016).

### 14.3.4 Posters

An effective poster can require considerably more time to prepare and print than is necessary to prepare an oral presentation. However, posters do have some distinct advantages over talks. Posters can provide more opportunities to interact with your audience than might be available through a moderated question and answer session that follows a series of presentations. You can respond to questions, provide additional details or context, use the poster as a spring-board for discussion, and receive direct feedback on your work. Due to the interpersonal nature of poster presentations, the communication style is different. Preparation is important. However you will necessarily go 'off script' and the practice and repetition recommended for a talk (Rule 7, Table 14.4) does not hold true for posters in the same way. Indeed, Erren and Bourne (2007) suggest that authors should take advantage of the unique nature of posters (Rule 6, Table 14.6). The less formal presentation format allows the author to be more speculative and poster presentations are great opportunities to distribute related materials.

A good poster will employ many of the good design elements identified for presentations in Sect. 14.3.2 including font choices, color schemes, contrasting elements, flow, image use, etc. However, since posters can also act as stand-alone communication mechanisms, the poster must be able to speak for you. For this reason, it is important to quickly attract and capture the attention of the viewer. There are many examples of poster best practices or 'tips' for creation and presentation (e.g., Plunkett 2016; Malson 2015; Purrington 2016) and multiple printing companies provide free tutorials and templates. When developing your poster it is a good idea to review other examples and identify what worked and what didn't, what stood out as the take home message, and what was lost in the details. The blog "Better Posters" (Faulkes 2016) provides examples of real posters that have been critiqued or improved upon, as well as discussion posts, and is a great resource for improving poster design.

**Table 14.6** Ten simple rules for good poster presentations [adapted from Erren and Bourne (2007)]

Number	Rule	Description
1	“Define the purpose”	This will vary by the intent and nature of the work. “Do you want the person .. to engage in discussion? .. try something for themselves? .. collaborate?” Ask yourself these questions before beginning.
2	“Sell your work in ten seconds”	You will likely be competing with many other posters and first impressions count. You need to ‘sell’ your work.
3	“The title is important”	Following from Rule 2, your “title is a good way to sell your work”. It should be “short and comprehensible to a broad audience”.
4	“Poster acceptance means nothing”	Acceptance is not an endorsement of your work. That comes from peers through good science and a well-presented poster.
5	“Many of the rules for writing a good paper apply to posters too”	“Identify your audience and provide the appropriate scope and depth of content”
6	“Good posters have unique features not pertinent to papers”	“Posters allow you to be more speculative. There is the opportunity to say more than you would in traditional literature.”
7	“Layout and format are critical”	As with slides, use natural flow and directions to guide the reader through the content.
8	“Content is important but keep it concise”	“Economy of words .. is particularly important for posters because of their inherent space limitations”. Clarity and precision of expression are also key.
9	“Posters should have your personality”	“Think of your poster as an extension of your personality” and use it to connect with passersby.
10	“The impact of a poster happens both during and after a poster session”	Make sure to engage your audience on the day and ‘present’ your poster. Also, “make it easy for a conference attendee to contact you afterward.”

### 14.4 Communication in a Virtual Environment

Social media refers to the collective of online communication channels that enable people to interact with each other by both sharing and consuming information. There are many different forms of social media including websites and applications dedicated to forums (question and answer discussion environments), blogging (online journaling), social networking (communication across networks of friends and colleagues), social curation (collaborative management of online information) and wikis (community contributed and curated websites). Here we discuss some commonly used social and online media, and provide guidance for using them effectively for research communication.

### ***14.4.1 Websites***

Websites, like newspaper and magazine articles, can be used to promote and enable discovery of your research findings and scholarly outputs; they can provide access to data, research findings, articles, videos and other project resources. Many universities and other organizations encourage their researchers to provide a brief curriculum vitae (including a list of recent publications) as a webpage on their institution's website. There are advantages to this approach. In particular, the institution maintains and supports the website and the researcher need only routinely update the content. In addition to institutional webpages, many individual researchers create a separate website for their individual or laboratory's scholarly output. For instance, a laboratory website normally highlights publications emanating from the laboratory as well as the students, post-doctoral associates, and technical staff that are associated with the laboratory. Such websites can be important for recruiting new students as well as for promoting the laboratory's research foci, findings and data. An individual website can be tailored to your individual needs but also requires time, money and, possibly, personnel for design, maintenance and update, and system administration. Furthermore, individual websites frequently must adhere to institutional requirements.

Most research networks and scientific organizations maintain institutional websites that are used to: (1) highlight discoveries and new publications; (2) announce meetings, news, and job openings to their members; (3) enable access to research data and tools; (4) raise funds; and (5) build and promote a sense of community. Examples of research network websites include:

- Global Lake Ecological Observatory Network (GLEON 2016);
- The Long Term Ecological Research Network (LTER 2016); and
- Nutrient Network: A Global Research Cooperative (NutNet 2016).

At least one individual working part-time is necessary to support a website for a research network and costs depend on the degree of functionality, amount and diversity of content, and frequency of updates. Large organizations may, of course, support websites that are created and maintained by a team of individuals. A prime example is the Cornell Lab of Ornithology website (Cornell University 2016) which includes links to its numerous citizen science programs (e.g., eBird, Project FeederWatch, NestWatch, etc.) as well as its research, education, technology and conservation programs. Many excellent reference books provide guidance on how to design and build good websites (e.g., Duckett 2011; Krug 2014; Robbins 2012).

### ***14.4.2 Types and Uses of Different Social Media***

Traditional online media, such as websites, will often have a social component to them: the ability to provide user comments or integration of social media applications within the site. However, there is enormous variety in social media



applications that might specialize in particular media (images, video) or facilitate a specific type of communication.

Statista (2016) lists Facebook as the most widely used social media network with 1590 million users. WhatsApp, a cross platform mobile messaging service, was listed second with 1000 million users, Instagram was 9th with 400 million users and Twitter was 10th with 320 million users. The leading social networks are usually available in multiple languages and enable users to connect with people across geographical, political or economic borders (Statista 2016).

Social media applications have been used widely in business for market research, communication, promotions, community development and e-commerce, and they are increasingly being used for research communication. Here we briefly summarize some of the more commonly used social media platforms within the scientific community and how they might support research dissemination.

- **Facebook** is an online social networking service that allows users to share their profiles, post updates and share media with their immediate network or the public. Facebook can be used to notify your networks of your research products or activities and create a moderated page dedicated to a particular topic. This latter option is often used by teams of researchers or research organizations (e.g., DataONE 2016; NEON 2016; NSF 2016).
- **Twitter** is also an online social networking service that enables users to communicate with their network. However, the form of communication (tweets) is presently limited to 140-character messages. These messages may contain links to online sources of images and Twitter is often used to promote activities and events as well as provide brief updates. The use of #hashtags (relevant phrases or keywords) facilitates discussion around a thematic topic and enables users to follow a ‘stream’ of conversation. They are also used to facilitate communication around a specific event, such as a conference.
- **Google+** (Google Plus) is an interest based social network platform, owned and operated by Google Inc. that provides multiple functions. Users have public profiles that provide standard demographic information. Users can provide updates about their activities and follow the updates from others that are organized into groups or ‘circles’. One of the most commonly used features of Google+ is the ‘hangouts’ communication service. Hangouts enables text chat, group video conferencing (with screen share functionality) and ‘Hangouts on Air’; the ability to live webcast from Google+ and stream to Youtube.
- **Slideshare** is a web based slide hosting service where users can upload files in public or private status for viewing and sharing. Researchers commonly use it as an online library of presentations they have given and for promoting a particular presentation. The site accepts multiple file formats and allows users to rate, comment on and share the material.
- **Figshare** is an online digital repository that enables researchers to preserve and share their research outputs. File formats include figures, datasets, images, and videos. Users can share, embed or download content exposed via Figshare.

- **Youtube/Vimeo** are video sharing platforms that enable individuals to upload and share video files. Users can like and comment on individual videos, as well as share them through social media. Video files can also be embedded in websites by the owner, rendering Youtube and Vimeo as video hosting services. “Video Abstracts” accompanying publications are becoming increasingly common, as are video data as supplemental material.

### ***14.4.3 Simple Steps for Effective Use of Social Media***

Twitter and Facebook have harnessed two of the largest communities of academics using social media and these are good places to start if social media is new to you. Whether Twitter or Facebook, building your social media presence comprises two primary activities (Leek 2016). First, you need to build your network; follow other people, and have other people follow you. It is not hard to acquire a long list of individuals, journals or organizations to follow; having individuals follow you is more challenging. To build a following requires that you post content that they find of interest. Posting solely about your work will not garner a broad following and Leek (2016) suggests a strategy of acting as a ‘content curator’ where you promote the work of others and share anything exciting, creative or important. Second, once you have established a network, you can use social media to build an audience for your scientific work.

Many best practices for businesses seeking brand engagement (Forant 2013) apply equally to individuals promoting their research. For example, make sure you follow back and interact with others, stay unique to your style and tone, and be as transparent as possible but don’t over-share. Leek (2016) also suggests you avoid ‘hot button’ issues unless they are directly relevant to your message. Controversy is rife on the internet and as a scientist using social media to promote your work, it may not be advantageous to engage in all discussions.

Increasingly, scientists are using Twitter as a means to live (micro) blog activities at research conferences (Lister et al. 2010; Ekins and Perlstein 2014). This use of social media enables non-attendees to stay connected with current activities, enables attendees to follow concurrent sessions and provides a platform for presenters to promote their work. Not surprisingly, a set of Twitter guidelines has emerged to help attendees, conference organizers and interested parties extend the value of the scientific content beyond the auditorium (Ekins and Perlstein 2014; Croxall 2014). Table 14.7 provides 10 simple rules of live tweeting at scientific conferences (Ekins and Perlstein 2014). Lister et al. (2010) extend recommendations beyond Twitter users and provide guidelines for conference organizers, bloggers and presenters who will be giving talks in an environment that is open to live blogging.

**Table 14.7** Ten simple rules of live tweeting at scientific conferences [adapted from Ekins and Perlstein (2014)]

Number	Rule	Description
1	“Short conference hashtag”	“Organizers should claim a short descriptive # that includes the year”
2	“Promote the hashtag”	“Highlight the # in all conference materials”
3	“Encourage tweeting”	“Session chairs can facilitate this and relay questions”
4	“Conference twitter etiquette”	“Keep questions short and on the science .. encourage responsible tweeting”
5	“Conference tweet layout”	“List speaker name, affiliation and conference # in the first tweet; surname or initials and meeting # are sufficient thereafter”
6	“Keep conference discussion flowing”	“Summarize presentations concisely, use # for keywords, and use “@ reply” to engage individuals”
7	“Differentiate your opinions from the speaker’s”	“Separate your own comments/viewpoints on the speaker or science .. from the speaker’s own words.”
8	“Bring questions up from outside”	“Check for and raise questions from those outside the conference, returning the speaker responses.”
9	“Meet other live tweeters face to face”	Build relationships and collaboration opportunities by organizing or participating in tweetups
10	“Emphasize impact of live tweeting”	“Ensure that positive effects of tweeting at conferences, such as discoveries, publications, or collaborations, are highlighted”

#### 14.4.4 Understanding Your Social Media Impact

Just as citations and downloads can indicate the use or impact of a journal article, social media analytics can provide equivalent information (Eysenbach 2011). By gathering data from blogs, websites and social media applications you can get insights into the reach and potential impact of your material. However, comparing analytics can be challenging since different applications collect different types of usage statistics. Widrich (2013) suggests that some of the most important metrics to track include click rates on social shares, Facebook “talking about this”, twitter followers, and your Klout score (Klout Inc. 2015). The Klout score provides a single number between 1 and 100 that represents your social influence (i.e., the more influential you are, the higher your Klout Score) and is based on more than 400 signals from 8 different networks. However, this does not enable you to access more detailed information on the demographics and behavior of your audience such as where they are located, what time of day they were most active, how long they engaged with your material, etc. Hines (2015) provides a comprehensive overview of these metrics for Facebook, Twitter, Google+, LinkedIn, Pinterest and Google Analytics that you may wish to explore.

## 14.5 Metrics and Altmetrics

Zhang (2014) postulates that “fewer but more significant papers serve both the research community and one’s career better than more papers of less significance” (Table 14.1, Rule 2). Various metrics and altmetrics have been proposed as measures of the significance of an individual scholarly publication as well as the cumulative contributions of an individual researcher. Some metrics that are commonly provided for individual publications include number of citations, page views and downloads. Similarly, the h-index was proposed as a mechanism for assessing the quality of a researcher’s output, and reflects the number of publications and the number of citations per publication (Hirsch 2005).

Many metrics like numbers of downloads, page views and Wikipedia citations increase gradually over time (Brody et al. 2006) making it difficult to assess the significance of recent publications. Consequently, altmetrics have been developed as one way to identify recent, potentially impactful scholarly publications by also tracking citations in various social web services (e.g., Piwowar 2013; Priem et al. 2012). Lin and Fenner (2013) grouped article-level metrics into five categories that can be related to increasing amount of engagement with the research article: (1) viewed (lowest level of engagement)—e.g., HTML views, PDF downloads; (2) saved—e.g., Mendeley, CiteULike; (3) discussed—e.g., science blogs, journal community, Twitter, Facebook; (4) recommended—e.g., F1000 Prime; and (5) cited (highest level of engagement)—e.g., Web of Science, CrossRef.

Several services calculate and track altmetrics including Altmetric (2016) and Impactstory (2016). Altmetrics is an active field of research; nevertheless, there is strong evidence that altmetrics like tweets, Facebook wall posts, research highlights, blog mentions, mainstream media mentions and forum posts are positively associated with citation counts (Thelwall et al. 2013). Social media usage is discussed above in Sect. 14.4.3 and guidance for reporting altmetrics in a curriculum vita is provided by Piwowar and Priem (2013).

## 14.6 Conclusion

Science is dramatically changing. The total global scientific output is doubling roughly every 9 years (Bornmann and Mutz 2015), tens of millions of scientific papers have been published in tens of thousands of scientific journals, and we are drowning in a sea of data and information. Discovering a particular research finding can be equated to finding the proverbial needle in the haystack. Given that science is predicated on advancing the state of knowledge, scientists increasingly must play a central role in clearly communicating and documenting their research findings in ways that are tailored to and reach appropriate audiences.

Research findings are more likely to be incorporated in the corpus of knowledge if they are communicated and disseminated to targeted stakeholders (e.g., readers of

high impact journals, attendees of high profile scientific conferences). Successful communication depends on having a clear, concise and impactful message that is accompanied by attractive and understandable supporting visuals and is underpinned by high quality and well-documented data.

The best practices highlighted in this chapter can help in advancing knowledge by promoting more effective storytelling, providing more impactful graphics and figures, and reaching both broader audiences and targeted stakeholders through general-audience outlets and social media. Scientific advancement, now more than ever, demands that we not only do good science, but that we also clearly communicate and broadly disseminate our discoveries and findings to other scientists, citizens, resource managers and decision-makers.

## References

- Alred GJ, Brusaw CT, Oliu WE (2011) Handbook of technical writing, 10th edn. St. Martin's Press, Bedford
- Altmetric (2016) Altmetric. <https://www.altmetric.com>. Accessed 18 Aug 2016
- Bale R, Neveln ID, Bhalla APS et al (2015) Convergent evolution of mechanically optimal locomotion in aquatic invertebrates and vertebrates. *PLoS Biol* 13(4):e1002123. doi:10.1371/journal.pbio.1002123
- Bergou AJ, Swartz SM, Vejdani H et al (2015) Falling with style: bats perform complex aerial rotations by adjusting wing inertia. *PLoS Biol* 13(11):e1002297. doi:10.1371/journal.pbio.1002297
- Bornmann L, Mutz R (2015) Growth rates of modern science: a bibliometric analysis based on the number of publications and cited references. *J Assn Inf Sci Technol* 66:2215–2222. doi:10.1002/asi.23329
- Bourne PE (2005) Ten simple rules for getting published. *PLoS Comput Biol* 1(5):e57. doi:10.1371/journal.pcbi.0010057
- Bourne PE (2007) Ten simple rules for making good oral presentations. *PLoS Comput Biol* 3(4):e77. doi:10.1371/journal.pcbi.0030077
- Brody T, Harnad S, Carr L (2006) Earlier web usage statistics as predictors of later citation impact. *J Am Soc Inf Sci Technol* 57:1060–1072. doi:10.1002/asi.20373
- Chaves ÓM, Bicca-Marques JC (2016) Feeding strategies of brown howler monkeys in response to variations in food availability. *PLoS One* 11(2):e0145819. doi:10.1371/journal.pone.0145819
- Cho M (2013) The science behind fonts (and how they make you feel). <http://thenextweb.com/dd/2013/12/23/science-behind-fonts-make-feel/>. Accessed 1 Aug 2016
- Cleveland WS, McGill R (1985) Graphical perception and graphical methods for analyzing scientific data. *Science* 229(4716):828–833
- Cook RB, Wie Y, Hook LA et al (2017) Preserve: protecting data for long-term use, Chapter 6. In: Recknagel F, Michener W (eds) *Ecological informatics. Data management and knowledge discovery*. Springer, Heidelberg
- Cornell University (2016) The Cornell Lab of Ornithology. <http://www.birds.cornell.edu/>. Accessed 18 Aug 2016
- Croxall B (2014) Ten tips for tweeting at conferences. <http://chronicle.com/blogs/profhacker/ten-tips-for-tweeting-at-conferences/54281>. Accessed 15 June 2016
- CTAN (2016) CTAN: Package tufte-latex. <http://www.ctan.org/pkg/tufte-latex>. Accessed 26 Jul 2016
- DataONE (2016) DataONE. <https://www.facebook.com/DataONEorg/>. Accessed 25 Nov 2016

- Duarte N (2008) *slide:ology: the art and science of creating great presentations*. O'Reilly Media, Sebastopol, CA
- Duarte N (2010) *Resonate: present visual stories that transform audiences*. Wiley, Hoboken, NJ
- Duckett J (2011) *HTML and CSS: design and build websites*, 1st edn. Wiley, Indianapolis
- Ekins S, Perlstein EO (2014) Ten simple rules of live tweeting at scientific conferences. *PLoS Comput Biol* 10(8):e1003789. doi:[10.1371/journal.pcbi.1003789](https://doi.org/10.1371/journal.pcbi.1003789)
- Erren TC, Bourne PE (2007) Ten simple rules for a good poster presentation. *PLoS Comput Biol* 3(5):e102. doi:[10.1371/journal.pcbi.0030102](https://doi.org/10.1371/journal.pcbi.0030102)
- Eysenbach G (2011) Can tweets predict citations? Metrics of social impact based on Twitter and correlation with traditional metrics of scientific impact. *J Med Internet Res* 13(4):e123
- Faulkes Z (2016) Better posters. <http://betterposters.blogspot.com/>. Accessed 26 Jul 2016
- Few S (2009) *Now you see it: simple visualization techniques for quantitative analysis*. Analytics Press, Oakland, CA
- Few S (2012) *Show me the numbers: designing tables and graphs to enlighten*, 2nd edn. Analytics Press, Burlingame, CA
- Few S (2016) Perceptual edge – examples. <http://www.perceptualedge.com/examples.php>. Accessed 26 Jul 2016
- Forant T (2013) 10 social media best practices for brand engagement. <https://www.marketingcloud.com/blog/social-media-best-practices-for-brand-engagement/>. Accessed 16 May 2016
- GLEON (2016) GLEON: global lake ecological observatory network. <http://gleon.org>. Accessed 18 Aug 2016
- Hampton SE, Anderson SS, Bagby SC et al (2015) The Tao of open science for ecology. *Ecosphere* 6(7):1–13. doi:[10.1890/ES14-00402.1](https://doi.org/10.1890/ES14-00402.1)
- Hines K (2015) All of the social media metrics that matter. <http://sproutsocial.com/insights/social-media-metrics-that-matter/>. Accessed 26 Jul 2016
- Hirsch JE (2005) An index to quantify an individual's scientific research output. *Proc Natl Acad Sci U S A* 102:16569–16572
- Impactstory (2016) Impactstory. <https://impactstory.org>. Accessed 18 Aug 2016
- Klout Inc. (2015) Klout. <https://klout.com>. Accessed 10 Aug 2016
- Kronick DA (1976) *A history of scientific & technical periodicals: the origins and development of the scientific and technical press, 1665-1790*. Scarecrow Press, Metuchen, NJ
- Krug S (2014) *Don't make me think, revisited: a common sense approach to web usability*, 3rd edn. New Riders, San Francisco
- Leek J (2016) *How to be a modern scientist*. Lean Publishing, Victoria, BC
- Lin J, Fenner M (2013) Altmetrics in evolution: defining and redefining the ontology of article-level metrics. *Inf Stand Q* 25:20–26
- Lister AL, Datta RS, Hofmann O et al (2010) Live coverage of scientific conferences using web technologies. *PLoS Comput Biol* 6(1). doi:[10.1371/journal.pcbi.1000563](https://doi.org/10.1371/journal.pcbi.1000563)
- LTER (2016) The long term ecological research network. <http://www.lternet.edu>. Accessed 18 Aug 2016
- Malson G (2015) Preparing a research poster for a conference. *Clin Pharm* 7(3). doi:[10.1211/CP.2015.20068193](https://doi.org/10.1211/CP.2015.20068193)
- Michener WK (2017) Data discovery, Chapter 7. In: Recknagel F, Michener W (eds) *Ecological informatics. Data management and knowledge discovery*. Springer, Heidelberg
- NEON (2016) National Ecological Observatory Network. <https://www.facebook.com/NEONScienceData/>. Accessed 25 Nov 2016
- NSF (2016) National Science Foundation. <https://www.facebook.com/US.NSF/>. Accessed 25 Nov 2016
- NutNet (2016) Nutrient network: a global research cooperative. <http://www.nutnet.umn.edu>. Accessed 18 Aug 2016
- Piwowar H (2013) Altmetrics: value all research products. *Nature* 493:159. doi:[10.1038/493159a](https://doi.org/10.1038/493159a)

- Piwowar H, Priem J (2013) The power of altmetrics on a CV. *Bull Am Soc Inf Sci Technol* 39:10–13
- Plunkett S (2016) Tips on poster presentations at professional conference. [http://www.csun.edu/plunk/documents/poster\\_presentation.pdf](http://www.csun.edu/plunk/documents/poster_presentation.pdf). Accessed 1 Jun 2016
- Priem J, Piwowar HA, Hemminger BM (2012) Altmetrics in the wild: using social media to explore scholarly impact. *ArXiv.org*. <http://arxiv.org/abs/1203.4745>. Accessed 5 Feb 2016
- Purrington CB (2016) Designing conference posters. <http://colinpurrington.com/tips/poster-design>. Accessed 24 May 2016
- R Markdown (2016) Tufte handouts. [http://rmarkdown.rstudio.com/tufte\\_handout\\_format.html](http://rmarkdown.rstudio.com/tufte_handout_format.html). Accessed 26 Jul 2016
- Robbins JN (2012) *Learning web design: a beginner's guide to HTML, CSS, JavaScript, and web graphics*, 4th edn. O'Reilly Media, Sebastopol, CA
- Robbins NB (2013) *Creating more effective graphs*. Wiley, Hoboken, NJ
- Robbins NB (2016) <http://www.nbr-graphs.com/examples/>. Accessed 26 Jul 2016
- Rose PM, Kennard MJ, Moffatt DB et al (2016) Testing three species distribution modelling strategies to define fish assemblage reference conditions for stream bioassessment and related applications. *PLoS One* 11(1):e0146728. doi:10.1371/journal.pone.0146728
- Rougier NP, Droettboom M, Bourne PE (2014) Ten simple rules for better figures. *PLoS Comput Biol* 10(9):e1003833. doi:10.1371/journal.pcbi.1003833
- Statista (2016) Leading social networks worldwide as of April 2016, ranked by number of active users (in millions). <http://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/>. Accessed 05 May 2016
- Strunk W Jr, White EB (1999) *The elements of style*, 4th edn. Longman, New York
- Sullivan C (2011) Strata 2011 [day 1]: communicating data clearly. [http://infosthetics.com/archives/2011/02/strata\\_2011\\_communicating\\_data\\_clearly.html](http://infosthetics.com/archives/2011/02/strata_2011_communicating_data_clearly.html) Accessed 10 Aug 2016
- Thelwall M, Haustein S, Larivière V et al (2013) Do altmetrics work? Twitter and ten other social web services. *PLoS ONE* 8(5):e64841. doi:10.1371/journal.pone.0064841
- Tufte EG (1983) *The visual display of quantitative information*. Graphics Press, Cheshire, CT
- Tufte ER (1990) *Envisioning information*. Graphics Press, Cheshire, CT
- Tufte ER (1997) *Visual explanations: images and quantities, evidence and narrative*. Graphics Press, Cheshire, CT
- Tufte ER (2003) *The cognitive style of Powerpoint*. Graphics Press, Cheshire, CT
- Tufte ER (2006) *Beautiful evidence*. Graphics Press, Cheshire, CT
- Tufte E (2016) *The work of Edward Tufte and Graphics Press*. <https://www.edwardtufte.com/tufte/>. Accessed 26 Jul 2016
- University of Chicago Press Staff (2010) *The Chicago manual of style*, 16th edn. University of Chicago Press, Chicago
- Weinberger CJ, Evans JA, Allesina S (2015) Ten simple (empirical) rules for writing science. *PLoS Comput Biol* 11(4):e1004205. doi:10.1371/journal.pcbi.1004205
- Widrich L (2013) 5 essential social media metrics to track and how to improve them. <https://blog.bufferapp.com/social-media-metrics-improve>. Accessed 26 Jul 2016
- Witt C (2016) *Effective handouts*. <http://wittcom.com/effective-handouts/>. Accessed 26 Jul 2016
- Wong DM (2013) *The Wall Street Journal guide to information graphics: the dos and don'ts of presenting data, facts, and figures*. Norton, New York
- Zhang W (2014) Ten simple rules for writing research papers. *PLoS Comput Biol* 10(1):e1003453. doi:10.1371/journal.pcbi.1003453
- Zinsser W (2006) *On writing well: the classic guide to writing nonfiction*, 30th edn. Harper Perennial, New York

# Chapter 15

## Operational Forecasting in Ecology by Inferential Models and Remote Sensing

Friedrich Recknagel, Philip Orr, Annelie Swanepoel, Klaus Joehnk, and Janet Anstee

**Abstract** This chapter addresses the demand of environmental agencies and water industries for tools enabling them to prevent and mitigate events of rapid deterioration of environmental assets such as contamination of air, soils and water, declining biodiversity, desertification of landscapes. Getting access to reliable early warning signals may avoid excessive ecological and economic costs.

Here we present examples of recently emerging technologies for predictive modelling and remote sensing suitable for early warning of outbreaks of toxic cyanobacteria blooms in freshwaters that pose a serious threat to public health and biodiversity. As demonstrated by two case studies, inferential models developed from *in situ* water quality data by evolutionary computation prove to be suitable for up to 30 days forecasting of population dynamics of cyanobacteria and concentrations of cyanotoxins in drinking water reservoirs with different climates. The models not only forecast daily concentrations of cyanobacteria and cyanotoxins but also daily proliferation rates. Proliferation rates exceeding  $0.2 \text{ day}^{-1}$  serve as criteria for early warning. Alarm is triggered if forecasted concentrations of cyanobacteria or cyanotoxins exceed predefined threshold values and proliferation

---

F. Recknagel (✉)  
University of Adelaide, Adelaide, SA, Australia  
e-mail: [friedrich.recknagel@adelaide.edu.au](mailto:friedrich.recknagel@adelaide.edu.au)

P. Orr  
Griffith University, Nathan, QLD, Australia  
e-mail: [p.orr@griffith.edu.au](mailto:p.orr@griffith.edu.au)

A. Swanepoel  
Rand Water, Vereeniging, South Africa  
e-mail: [aswanepo@randwater.co.za](mailto:aswanepo@randwater.co.za)

K. Joehnk  
CSIRO Land and Water, Canberra, ACT, Australia  
e-mail: [klaus.joehnk@csiro.au](mailto:klaus.joehnk@csiro.au)

J. Anstee  
CSIRO Ocean and Atmosphere, Canberra, ACT, Australia  
e-mail: [Janet.Anstee@csiro.au](mailto:Janet.Anstee@csiro.au)



rates exceed  $0.2 \text{ day}^{-1}$ , constituting a bloom event. Findings from these case studies suggest that cyanobacteria blooms can be forecasted up to 30 days ahead in real-time mode solely based on online water quality data monitored by multi-sensor data loggers.

Advanced remote sensing technology allows to quantify absorption/reflectance characteristics of algal pigments of a water column for deriving chlorophyll-*a* concentrations as indicator for algal biomass, or discriminating cyanobacteria by specific pigments such as cyano-phycoyanin and cyano-phycoerithrin. It has the potential to monitor spatio-temporal distribution of water quality parameters and cyanobacteria blooms based on sufficient spatial, temporal and spectral resolution of the sensors, and the availability of suitable algorithms to match satellite information with high-resolution in-situ measurements. The chapter discusses the prospect of using remote sensing technology for forecasting seasonal trajectories of cyanobacteria blooms that requires the combination of in-situ monitoring and remote sensing data with hydrodynamic models. By deriving vertical light attenuation in the water column from remote sensing data, hydrodynamic models will be enabled to predict seasonally occurring cyanobacteria blooms.

## 15.1 Introduction

Some events occur suddenly and spread rapidly temporarily disturbing ecosystem states or, possibly, causing irreversible ecosystem change. Temporary disturbances may be caused by local wild fires, sporadic pathogenic or toxic pollution, whereas irreversible changes may be caused by accumulations of pollutants, bioinvasions or climate change. Ecological and economic costs of both cases can be high, and tools for operational forecasting are needed to avoid or minimise these costs.

Operational forecasting of sudden, detrimental events in ecosystems is a very challenging task that can be approached by quantifying tipping points (e.g. Scheffer et al. 2009; Huber et al. 2012; Recknagel et al. 2014a), through real-time forecasting (e.g. Recknagel et al. 2014b; Ye et al. 2014) or remote sensing (e.g. Lunetta et al. 2015; Matthews and Odermatt 2015).

This chapter provides examples of operational forecasting of outbreaks of harmful algal blooms (HABs) by inferential models developed using evolutionary algorithms, and by remotely sensed data.

## 15.2 Early Warning of HABs Based on Inferential Modelling

Two case studies of real-time forecasting and early warning of cyanobacteria blooms are presented based on the early warning scheme shown in Fig. 15.1. The scheme suggests the use of routine and *in situ* monitoring as sources of water

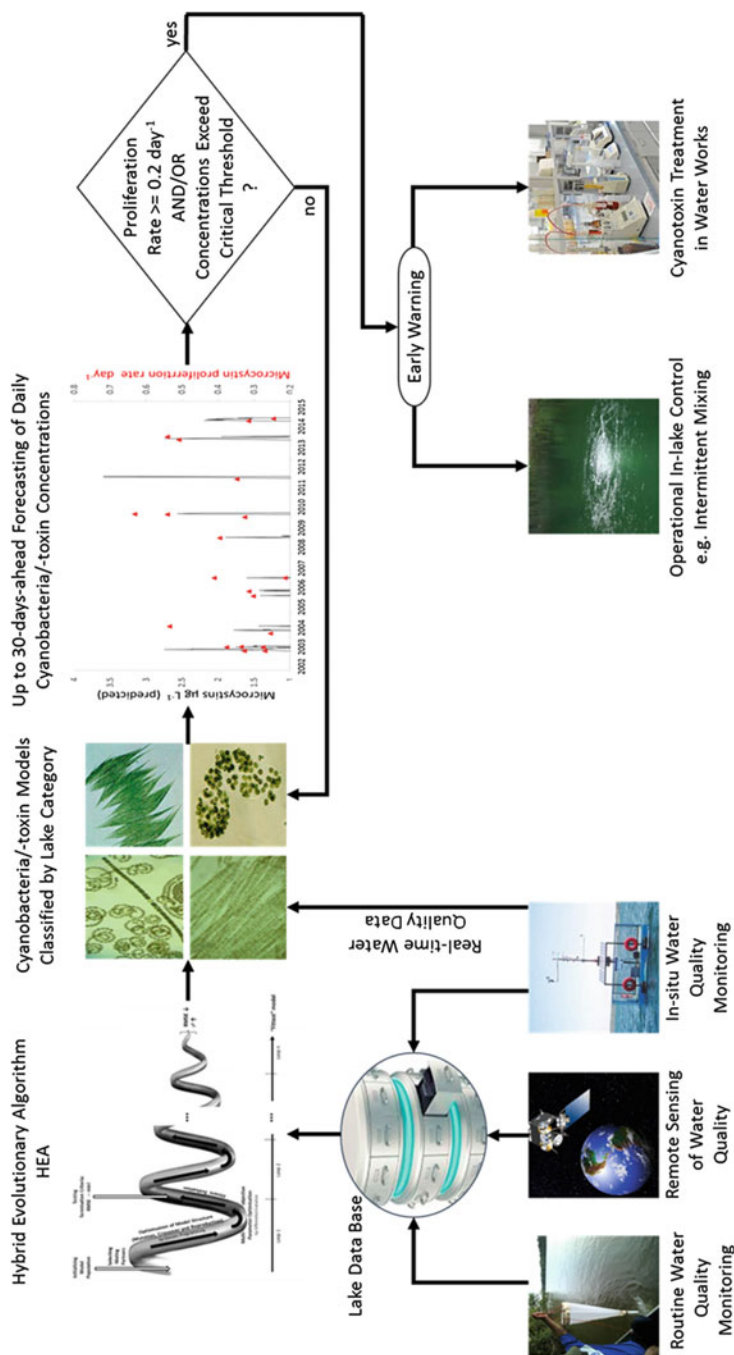


Fig. 15.1 Conceptual diagram for real-time forecasting, early warning and alarm of HABs by inferential models based on HEA (Recknagel et al. 2017)

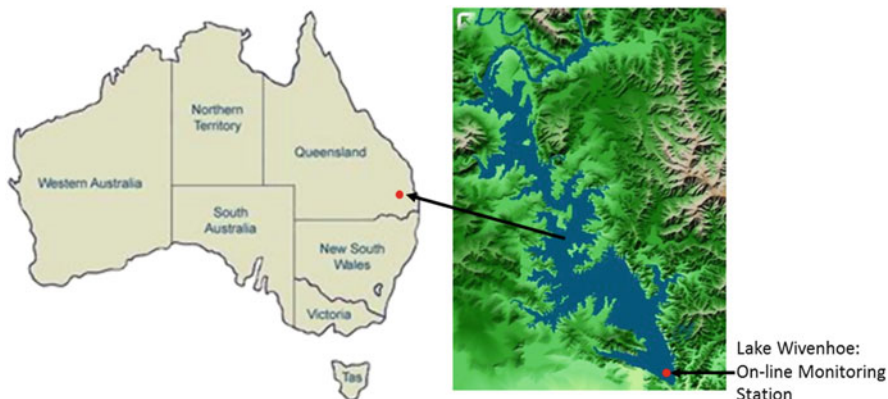
quality data that are pre-processed and archived in a lake data base prior to inferential modelling using the hybrid evolutionary algorithm (HEA; see Chap. 9). Species-specific cyanobacteria models are validated before being used for up to 30-day forecasts based on in-situ water quality data in real-time mode. Daily forecasted concentrations of cyanobacteria populations or cyanotoxins are analysed in relation to thresholds of related proliferation rates. If proliferation rates and concentrations exceed critical levels defined by local water authorities, early warnings of HABs are issued to lake management and water works. As displayed in Fig. 15.1, early warnings may activate intermittent mixing of the lake and *ad hoc* cyanotoxin treatment in water works.

The two case studies demonstrate that the HEA enables development of predictive inferential models that:

1. can be applied in real-time mode for early warning of cyanobacteria blooms. The models are solely based on electronically measurable predictor variables such as water temperature (WT), dissolved oxygen (DO), turbidity (TURB), pH, electrical conductivity (EC) and chlorophyll-a (Chla) being monitored *in situ* by multi-probe data loggers (e.g. YSI 6920 V2-2).
2. directly target the threat from cyanobacteria blooms posed by cyanotoxins rather than inadvertently forecasting cyanobacteria blooms of non-toxic strains.

The first example is from Lake Wivenhoe (Australia) and takes advantage of *in situ* water quality monitoring by multi-probe data loggers that have operated near the lake outlet since 2007. Daily *in situ* data and weekly to biweekly cyanobacteria cell counts from 2007 through 2015 are used for modelling population dynamics of *Cylindrospermopsis raciborskii*. The second example is from Vaal Reservoir (South Africa) and models microcystin concentrations based on *in situ* water quality data recorded at weekly/biweekly intervals from 2002 through 2015. Since the six *in situ* water quality variables were not measured in real-time, interpolated daily data were used for modelling.

Details of the design and functioning of HEA (Cao et al. 2014; Recknagel et al. 2014a, b) are presented in Chap. 10. Since HEA induces models from long-term data patterns, it appears that the more event-related patterns the historical data contains, the more generic the models tend to become, and the more likely the model's predictive validity reaches beyond the data limits. Ongoing ecosystem evolution requires that models be regularly updated using the most recent monitoring data.



**Fig. 15.2** Location of Lake Wivenhoe in Queensland, Australia

### 15.2.1 *Cyanobacterium Cyndrospermopsis* in Lake Wivenhoe (Australia)

Lake Wivenhoe is a warm-monomictic and mesotrophic reservoir located near Brisbane in the subtropical southeast of Queensland, Australia (Fig. 15.2). The lake has a [catchment area](#) of 7020 km<sup>2</sup>, an average depth of 11 m with a surface area of 108 km<sup>2</sup> and a volume of 1.165 million ML.

Blooms of the filamentous cyanobacterium *Cyndrospermopsis raciborskii* occur annually in Lake Wivenhoe (Orr et al. 2010). *C. raciborskii* produces the hepatotoxic cylindrospermopsin, which presents a risk to human health (e.g. Hawkins et al. 1985) and must be removed during water treatment. Controlling the development of *C. raciborskii* within the reservoir is a key goal of Seqwater ([www.seqwater.com.au](http://www.seqwater.com.au)), the water authority responsible for reservoir management. However, this cyanobacterium is ecologically adaptable and can form blooms under a range of light, temperature and nutrient regimes, and may tolerate nitrogen-depleted waters through the ability to fix atmospheric nitrogen (N<sub>2</sub>) (Bouvy et al. 2000; Moisander et al. 2008).

Daily *in situ* data and weekly cyanobacterium cell counts from 2007 through 2015 (Table 15.1) were used to develop HEA models for 10, 20, and 30-day forecasts of *C. raciborskii*. Resulting models are tested for their capacity to provide early warnings to employ operational in-lake measures for HAB control such as intermittent mixing, and to alert within days the drinking water treatment plant of high concentrations of *C. raciborskii* possibly present in source water.

Both cross- and split-sample-validation displayed good correspondence between observed and 10-day-ahead forecasts of *C. raciborskii* population dynamics in Lake Wivenhoe for 2007 through 2015 (Fig. 15.3b, e and f). The model in Fig. 15.3a achieved a coefficient of determination  $r^2 = 0.48$  with an IF-condition that separates high and low cell numbers depending on distinct ranges of pH and electrical conductivity (see Fig. 15.3d). Based on the assumption that 8000 cells mL<sup>-1</sup> of

**Table 15.1** Limnological data of Lake Wivenhoe

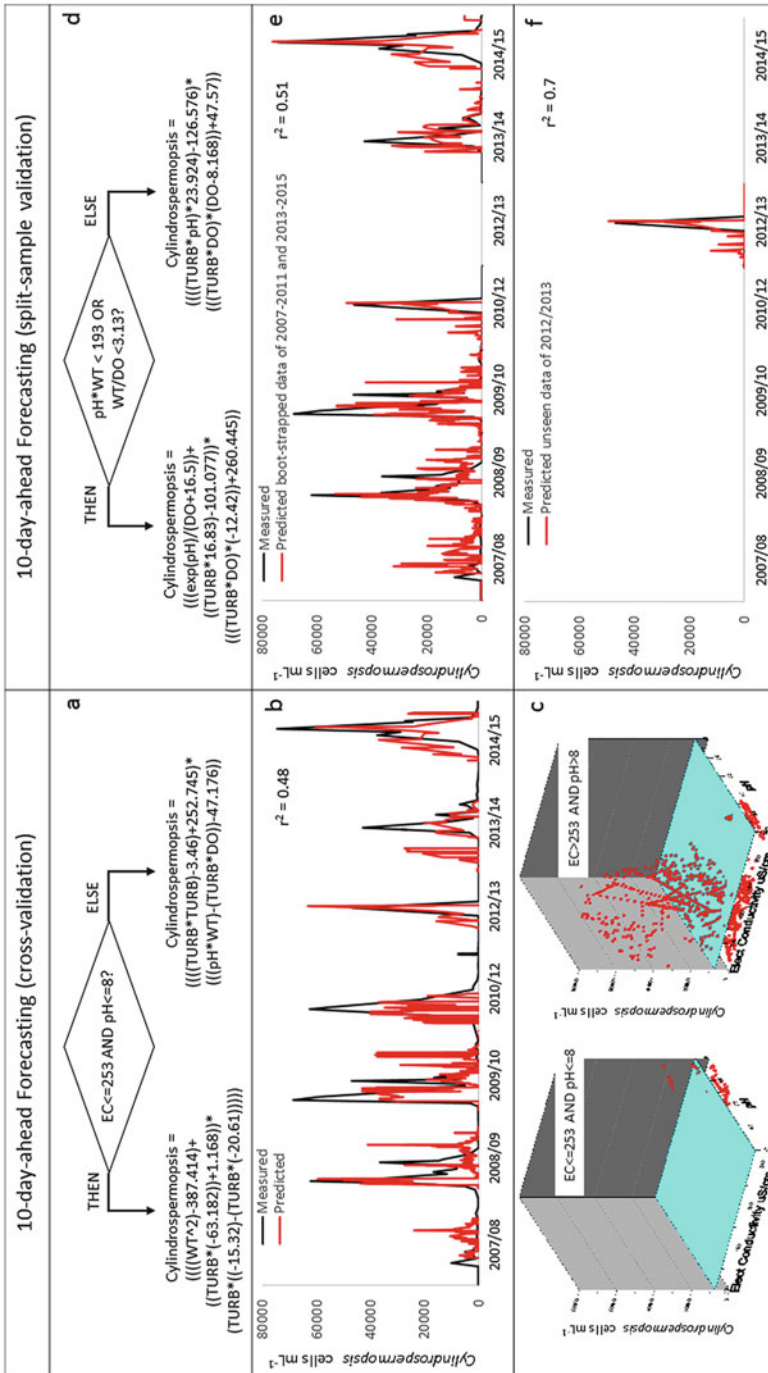
Water quality variables	Units	2007–2015
		Mean/Min/Max
Physical-chemical parameters		
WT (Temperature)	°C	14.83/22.59/29.5
TURB (Turbidity)	NTU	10.08/1.5/288.1
pH		8.1/6.5/9.7
DO (dissolved oxygen)	mg L <sup>-1</sup>	7.96/4.85/11.16
EC (electrical conductivity)	μS cm <sup>-1</sup>	335/140/496.5
Biological parameters		
<i>Cylindrospermopsis raciborskii</i>	cells mL <sup>-1</sup>	7126/1/74700

*C. raciborskii* corresponds to the 1 μg L<sup>-1</sup> threshold concentration for cylindrospermopsins in drinking water currently being considered by the WHO (2003), the timing of outbreaks of major blooms in the period between 2008 and 2015 has been accurately forecasted except for a minor bloom event in summer 2007/2008 (Fig. 15.3b). Although magnitudes of blooms in 2009, 2010 and 2013 were lower than the observed data, models predicted the observed bloom events of more than 8000 cells mL<sup>-1</sup>.

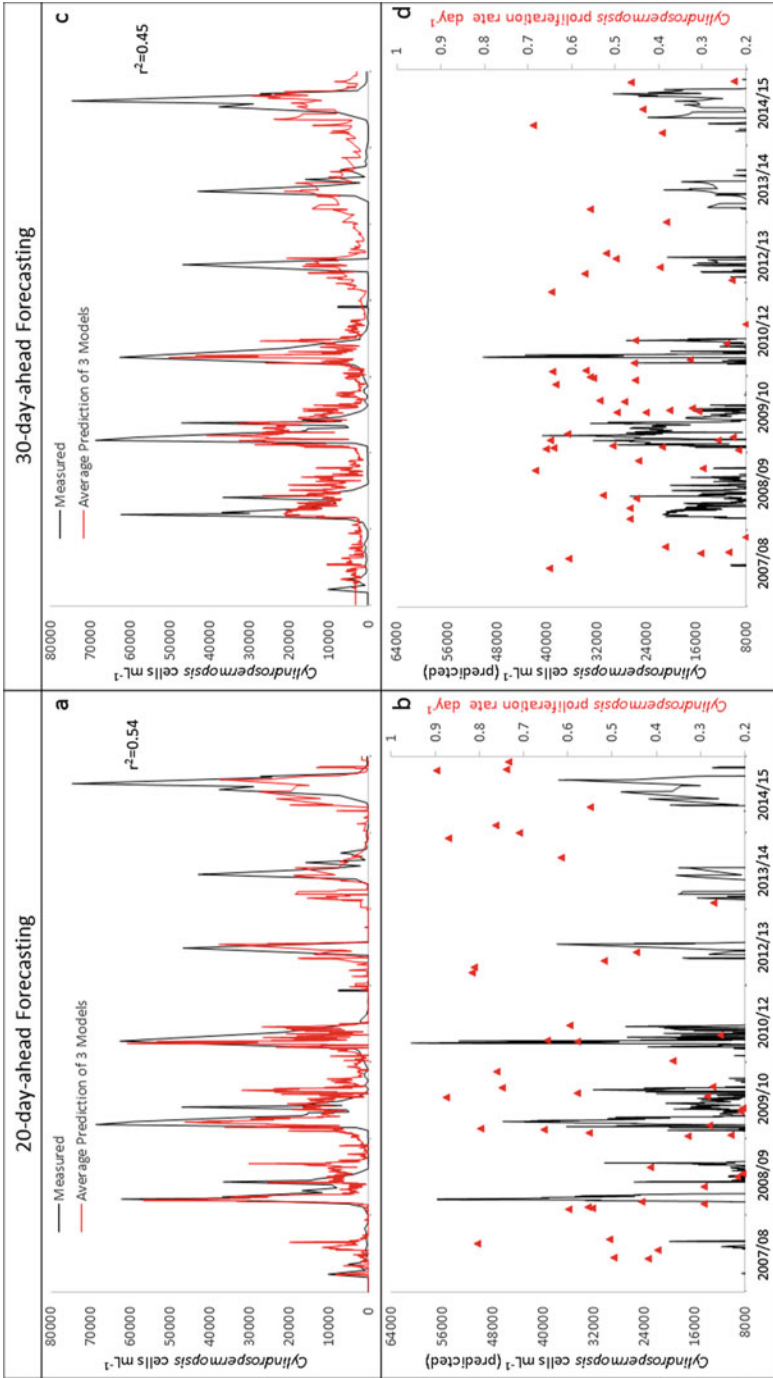
Figure 15.4 illustrates validation results for 20- and 30-day-ahead forecasts of *C. raciborskii* that were averaged from the three best models for each case. The averaged models accurately predict the timing of fast population growth of the observed bloom events between 2008 and 2015 but underestimate magnitudes for most of the bloom events (Fig. 15.4a, c). However, both the 20- and 30-day-ahead forecasts accurately predict cell division rates greater than 0.2 day<sup>-1</sup> and magnitudes higher than 8000 cells mL<sup>-1</sup> that according to Fig. 15.1 would cause alarm.

The case study of Lake Wivenhoe leads to the following conclusions:

- Models developed for 10- to 30-day-ahead forecasting of *C. raciborskii* predicted observed bloom events even though magnitudes of observed cell numbers were often underestimated.
- The water temperature above which *C. raciborskii* grows fastest in Lake Wivenhoe appears to be 26.1 °C, which matches findings by Briand et al. (2002). The thresholds also suggest that *C. raciborskii* is tolerant to higher electrical conductivity up to 382 μS cm<sup>-1</sup> as previously suggested by Briand et al. (2002) and Moisander et al. (2008).
- Forecasting models for *C. raciborskii* in Lake Wivenhoe are solely driven by electronically-measurable *in situ* water quality data.



**Fig. 15.3** 10-day-ahead forecasting models for *Cylindrospermopsis* in Lake Wivenhoe from 2007 to 2015 (flood year 2011 not included). *Cross-validation*: IF-THEN-ELSE model (a), validation result (b), separation of high and low cell numbers by IF-conditions; *Split-sample validation*: IF-THEN-ELSE model (d), validation results (e), (f). (Recknagel et al. 2017)



**Fig. 15.4** Forecasting of *Cylindrospermopsis* in Lake Wivernhoe from 2007 to 2015 (flood year 2011 not included). 20-day-ahead forecasting: cross-validation of 3 best models (a), cell concentrations and cell division rates predicted by three best models (b) 30-day-ahead forecasting: Cross-validation of three best models (c), cell concentrations and cell division rates predicted by 3 best models (d) (Recknagel et al. 2017)



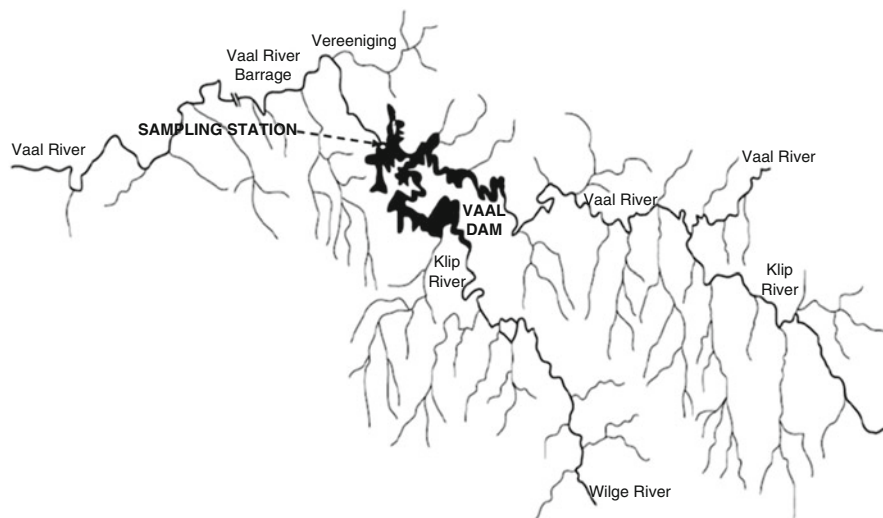


Fig. 15.5 Location of the Vaal Reservoir

### 15.2.2 Cyanotoxin Microcystins in Lake Vaal (South Africa)

The warm-monomictic and mesotrophic Vaal Reservoir (Fig. 15.5) is South Africa's largest drinking water reservoir and is located approximately 150 km south of Johannesburg. It has a catchment area of 38,500 km<sup>2</sup>, a maximum depth of 47 m, a surface area of 320 km<sup>2</sup> and a maximum volume of 2.61 million ML.

The toxic microcystins that are produced by *Microcystis* spp. detrimentally affects aquatic biodiversity, animals and human health (Carmichael 1994). The maximum health limit of microcystins concentrations in drinking water has been defined by the World Health Organization (WHO 2003) as 1 µg L<sup>-1</sup>. Climate and water quality conditions of the Vaal Reservoir favour outbreaks of *Microcystis* spp. blooms and can lead to increased concentrations of microcystins in the reservoir (Conradie and Barnard 2012). Thus, the monitoring and control of cyanobacteria blooms is a high priority for Rand Water ([www.randwater.co.za](http://www.randwater.co.za)), the water authority responsible for the management of the Vaal Reservoir.

In-situ water quality data and total microcystin concentrations measured from 2002 through 2015 (Table 15.2) were used to develop forecasting models for microcystins for 10 to 30 days ahead by HEA. Resulting models were tested for their suitability to be applied for early warning of microcystins concentrations that exceed 1 µg L<sup>-1</sup> in Vaal Reservoir. Such models would directly target the threat from cyanobacteria blooms posed by cyanotoxins rather than forecasting nontoxic cyanobacteria populations.

Figure 15.6 documents cross- and split-sample validation for 10-day-ahead forecasting models of microcystins. Both models fail to predict a minor peak



**Table 15.2** Limnological data of Vaal Reservoir

Water quality variables	Units	2002–2015
		Mean/Min/Max
Physical-chemical parameters		
WT (Temperature)	°C	17.5/8.8/26
TURB (Turbidity)	NTU	57.1/8.6/141
pH		7.7/5.8/10.7
DO (dissolved oxygen)	mg L <sup>-1</sup>	7.4/3.2/11.1
EC (electrical conductivity)	mS cm <sup>-1</sup>	19.6/13.9/55
Biological parameters		
Chlorophyll-a	µg L <sup>-1</sup>	11.9/0.67/101
Total microcystin	µg L <sup>-1</sup>	0.42/0.1/5

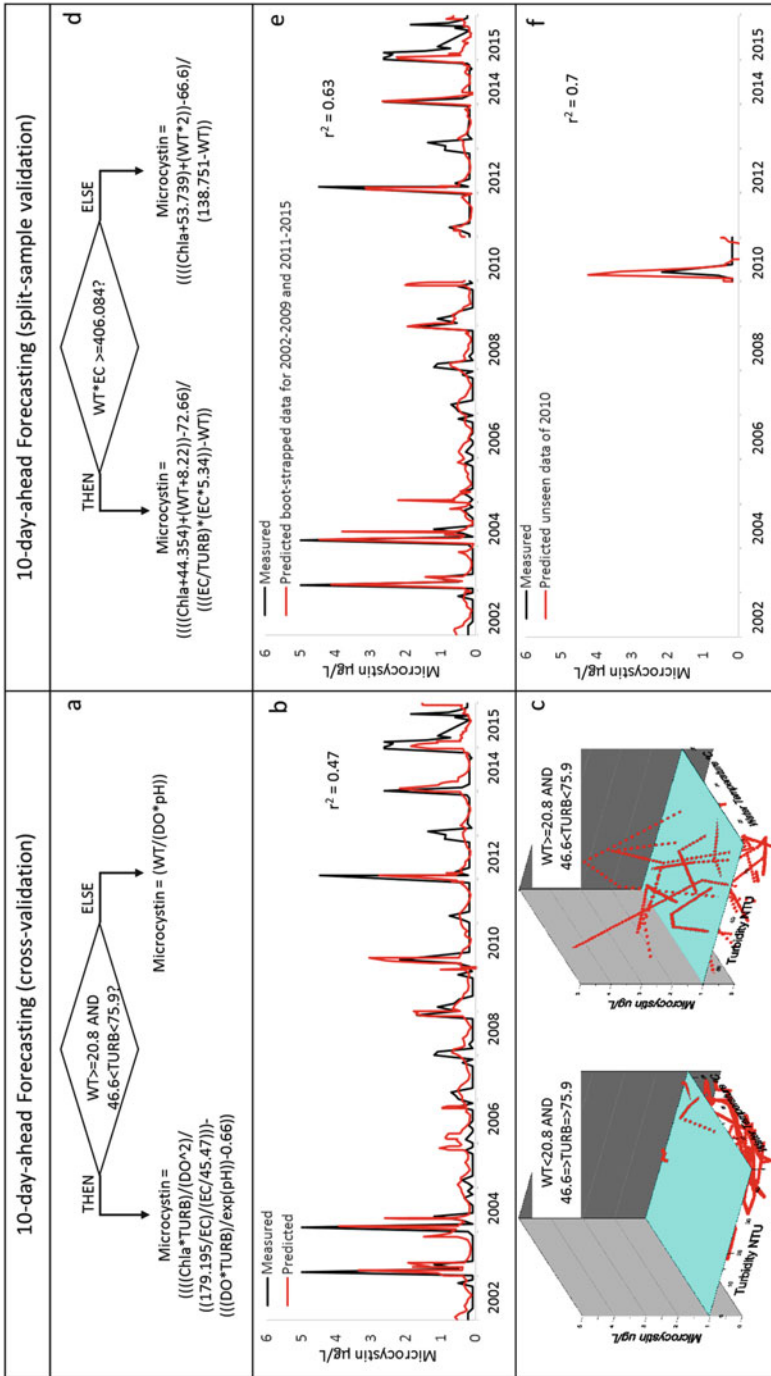
event in 2013 but accurately forecast the major peak events in 2003, 2004, 2010, 2012, 2014 and 2015 in relation to correct timing and concentrations of microcystins greater than 1 µg L<sup>-1</sup> (Fig. 15.6b, d, e). The threshold conditions of the model in Fig. 15.6a suggest that water temperatures greater than 20.8 °C and turbidity ranging between 46 and 75 NTU were indicative of high microcystins concentrations in the Vaal Reservoir. This finding reflects the fact that highest microcystins concentrations can be expected during the collapse of a *Microcystis* bloom typically occurring at warm water temperatures and causing low transparency. Figure 15.6c illustrates how these thresholds separate microcystins concentrations above and below 1 µg L<sup>-1</sup> as a prerequisite for the model's forecasting performance.

Figure 15.7 documents results averaged from the 3 best models for forecasting microcystins concentrations in the Vaal Reservoir for 20- and 30-day-ahead. The models accurately forecast major peak events in 2003, 2004 and 2011 with microcystins concentrations above 1 µg L<sup>-1</sup> (Fig. 15.7a, c). Figure 15.7b, d reveal daily proliferation rates greater than 0.2 day<sup>-1</sup> before and during events of significantly increased microcystins concentrations.

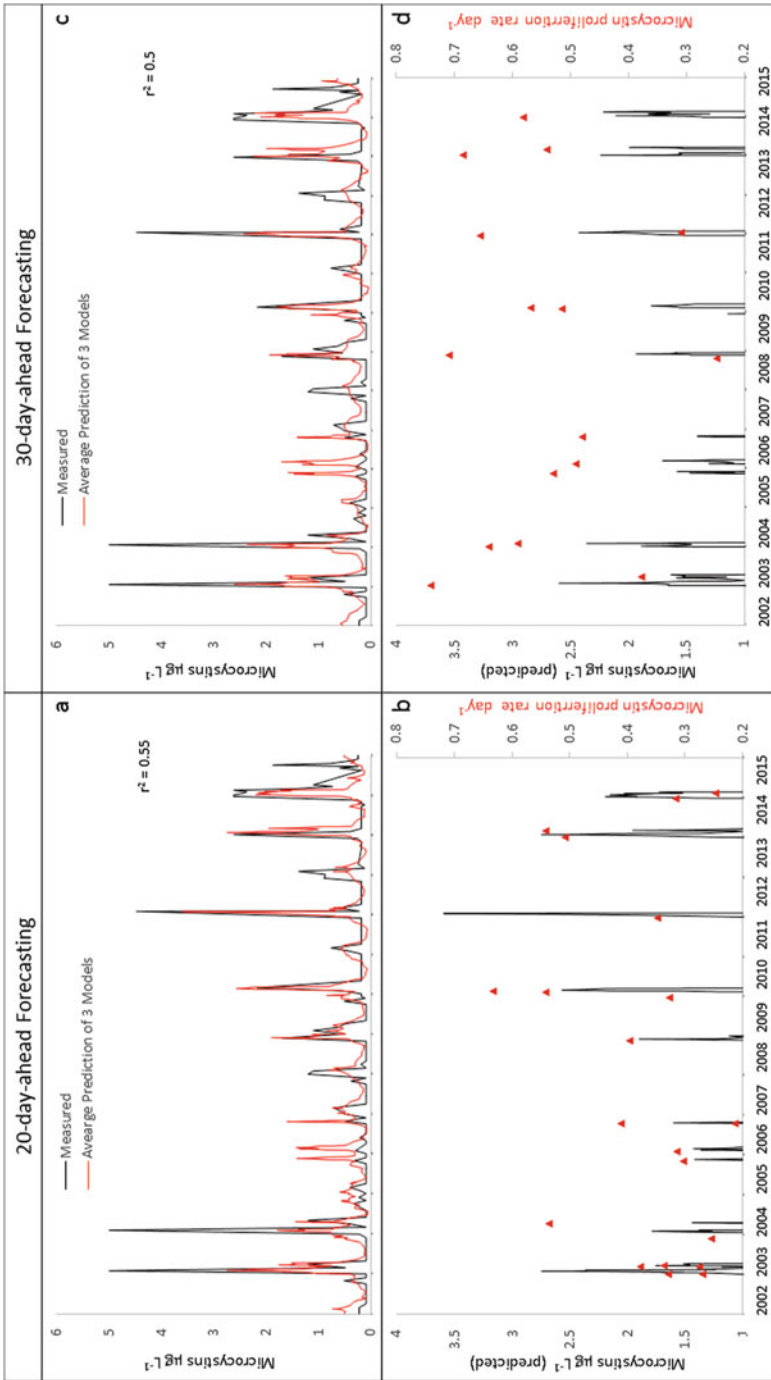
The Vaal Reservoir case study leads to the following conclusions:

- Models for 10- to 30-day-ahead forecasting of microcystins concentrations in the Vaal Reservoir prove to be valid for early warning of events that exceed 1 µg L<sup>-1</sup>.
- IF-conditions of the models reveal water quality conditions under which concentrations of microcystins are typically rising in Vaal Reservoir, and suggest that water temperatures greater than 20 °C and turbidity ranging between 46 and 75 NTU may be indicative for such events.
- Forecasting models for microcystins in the Vaal Reservoir neither require costly cyanobacteria cell counts nor nutrient measurements but are solely driven by electronically-measurable *in situ* water quality data.

Although there is a highly complex synergy among environmental and climate factors on one hand, and the sudden proliferation of cyanobacteria and -toxins in freshwaters on the other hand, the two case studies demonstrated that inferential



**Fig. 15.6** 10-day-ahead forecasting models for microcystins in the Vaal Reservoir from 2002 to 2005. *Cross-validation*: IF-THEN-ELSE model (**a**), validation result (**b**), separation of microcystins concentration below and above  $1 \mu\text{g L}^{-1}$  by IF-conditions (**c**) *Split-sample validation*: IF-THEN-ELSE model (**d**), validation results (**e**), (**f**) (Recknagel et al. 2017)



**Fig. 15.7** Forecasting models for microcystins in Vaal Reservoir from 2002 to 2015. 20-day-ahead forecasting: cross-validation (a) (dotted line  $1 \mu\text{g L}^{-1}$ ), forecasted concentration gradients (b); 30-day-ahead forecasting: cross-validation (c) (dotted line  $1 \mu\text{g L}^{-1}$ ), forecasted daily concentrations and proliferation rates (d) (Recknagel et al. 2017)

models based on HEA can accurately forecast short-term (10 to 30 day) temporal patterns in cyanobacteria and -toxins. Findings indicate the possibility to forecast cyanobacteria blooms in real-time mode solely based on online water quality data monitored by multi-sensor data loggers. More detailed results of the two case studies are documented in Recknagel et al. (2017).

## 15.3 Early Warning of HABs Based on Remotely-Sensed Data

### 15.3.1 Earth Observation of Water Quality Parameters

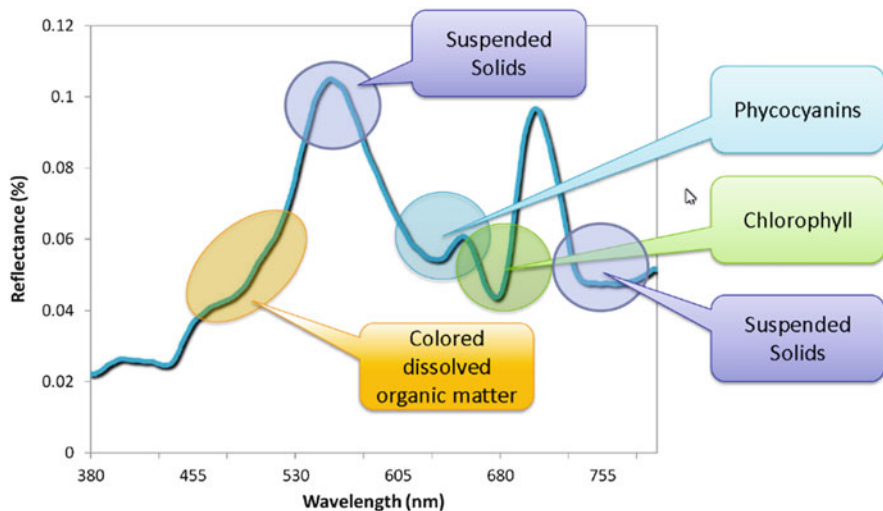
Forecasts of harmful algal blooms in lakes have historically been based on single- or multi-site observations of water quality parameters. This is usually done by selecting one or more locations in a lake that are assumed to represent the whole system. Of course, high spatial and temporal variability can lead to significant over- or underestimation of parameter values. To overcome this, field campaigns are usually necessary to record spatial and temporal patterns in a limited number of sites. Satellite remote sensing has an advantage over traditional field-monitoring methods, as it can provide a picture across the entire lake surface. The last decades of development in monitoring technology, i.e. increased spatial and spectral resolution, has enhanced the applicability of remote sensing for monitoring lakes. In addition to bathymetric and large-scale surface water mapping (Mueller et al. 2015), efforts have been undertaken to monitor algal blooms in inland waters (e.g., Klemas 2012; Odermatt et al. 2010).

Earth observations enable scientists to quantify the light environment of a water column and derive water quality parameters (for an overview, see e.g., Dekker and Hestir 2012). Chlorophyll is the most widely used index of water quality and nutrient status. Other pigments, i.e. cyano-phycoyanin and cyano-phycoerithrin can be used to discriminate cyanobacteria. Coloured dissolved organic matter can be used to estimate carbon content in aquatic systems, and total suspended matter allows one to derive the underwater light environment. For hydrodynamic modelling, vertical light attenuation in the water column can also be derived from remote sensing data. Observation of harmful algal blooms in water bodies is based mainly on the absorption/reflectance characteristics of algal pigments. Although focussing, by nature, only on the surface layers, remote sensing can provide an invaluable source of information for the spatio-temporal distribution of water quality parameters and cyanobacteria blooms. However, challenges for the observation of blooms in lakes remain, e.g., spatial and temporal resolution, cloud cover, atmospheric correction, etc. (Dekker and Hestir 2012; Dörnhöfer and Oppelt 2016; Palmer et al. 2015 and references therein).

As ocean observation was the first application of remote sensing in aquatic systems, focus was given to the well-known absorption characteristics of

chlorophyll-a to derive phytoplankton biomass characteristics on large scales. Satellite optical instruments are available for observation of specific bands of chlorophyll absorption around 440 nm and 681 nm. However, inland waters pose a more complex problem, as many water constituents absorb light at different bands across the entire visible spectrum (Fig. 15.8). Optical instruments in different satellites resolve different bands across the whole spectrum (see Table 15.3) which might not coincide or overlap with specific bands of chlorophyll absorption. Water quality parameters like chlorophyll content can only be estimated using ratios between discrete numbers of available spectral bands.

The observation of lake water using satellite optical sensors depends on the spatial, temporal and spectral resolution of the sensors and the availability of high resolution in-situ data for algorithm validation. The long-running Landsat series of satellites (Landsat 1 started in 1972, Landsat 7 has been in orbit since 1999) has a good spatial resolution of 30–79 m, but lacks a high spectral band resolution. The newer satellite in this series, Landsat 8, has a better spectral resolution but is of limited use for the observation of harmful algal blooms due to a long revisit time of about 16 days. Nevertheless, as the Landsat image archive consists of several decades of data, this can provide useful information on historic changes in water quality parameters. Two other sensors are often used for water quality information, MODIS (Moderate resolution imaging spectrometer) and MERIS (Medium resolution imaging spectrometer) have higher spectral resolution (see Table 15.3) but their spatial resolution of 300–1000 m limits their use to larger lakes (e.g. Lunetta et al. 2015). The Sentinel-2 mission (launched in 2015; a second satellite will be launched in 2017) offers new opportunities for water quality observations having optical sensors with 13 bands in the visible, near infrared, and short wave infrared part of the spectrum and spatial resolutions of 10, 20, and



**Fig. 15.8** Characteristic absorption spectrum of inland waters

**Table 15.3** Characteristics of selected satellite sensor systems and their usability for water quality observation [modified after Dekker and Hestir (2012)]

Satellite sensor system		Pixel size	(400–1000 nm)	Spectral bands	Revisit cycle	Suitability
		(m)		(d)	Chlorophyll	Cyano pigments
Current	MODIS	1000	9	1	Highly suited	Potential
	MODIS	250–500	2	1	Not suitable	Not suitable
	MERIS & OCM-2	300	15	2–3	Highly suited	Highly suited
	Landsat	30	4	16	Potential	Potential
	Rapideye	6.5	5	1.5	Potential	Potential
	Worldview-2	2	8	On demand	Suited	Suited
Future	Sentinel-3	300	21	1	Highly suited	Highly suited
	HySpIRI	60	60	19	Highly suited	Highly suited

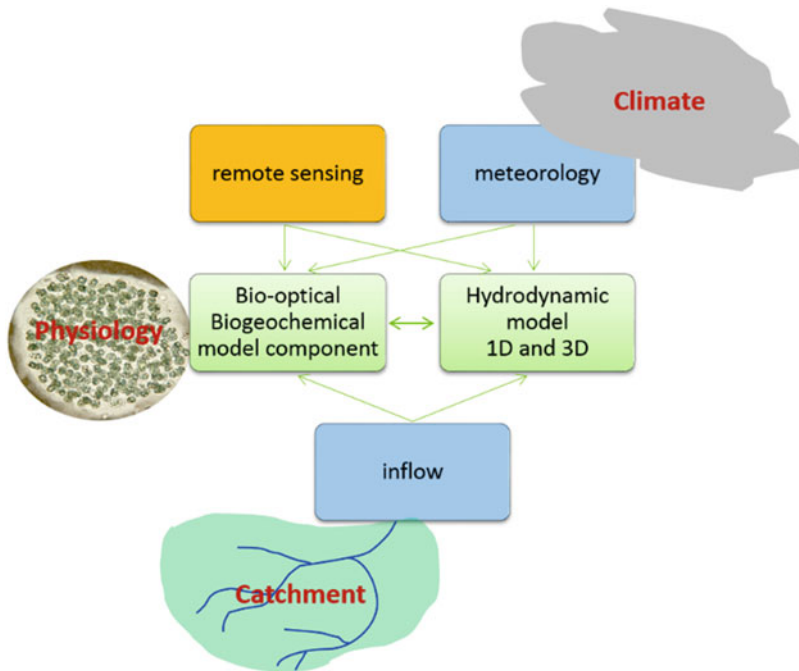
60 m. Their combined revisiting time will be 5 days. The use of such new satellite systems for the interpretation of water quality parameters will enhance our capability to predict harmful algal blooms, but further research is needed to generate suitable algorithms to match satellite information with high-resolution in-situ measurements (Manzo et al. 2015; Toming et al. 2016). In the future we will see hyperspectral satellite missions, e.g., Hyperspectral Infrared Imager or HySpIRI, with several hundreds of spectral bands and spatial resolution of 60 m but low temporal resolution of 19 days. This system will be effective for detecting seasonal changes (Hestir et al. 2015) and can serve as a basis for process-based models covering the intermediate time between overpasses. Furthermore, it will enable discrimination of cyanobacteria species depending on their pigment composition (Kudela et al. 2015).

Mapping of cyanobacteria blooms requires the aforementioned, multispectral or hyperspectral resolution sensors to derive information on cyanobacteria-specific pigments. When such information is not available, chlorophyll can be used as proxy as long as cyanobacteria are the dominant species. It will even result in a better estimation than cyano-phyco cyanin from current sensors due to its sensitivity to detection (Stumpf et al. 2016). For lakes with a known record of cyanobacteria blooms such an approach is permissible, i.e. in-situ data are required to legitimize the remotely sensed information. However, a chlorophyll-based estimation for concentration of cyanobacteria and related cyanotoxins tends to overestimate cyanotoxin content (Loftin et al. 2016).

### 15.3.2 Forecasting HABs Using Earth Observations

Classical, process-based models rely on massive data on water quality parameters at multiple sites for calibration and validation. This includes knowledge of physiological characteristics of dominant species, in-situ data in the lake, as well as drivers like meteorological data over the lake and inflow characteristics to the lake (Fig. 15.9). Earth observation using satellite remote sensing provides spatially extensive information on certain water quality parameters reflecting bio-optical properties in the surface waters. These snapshots, depending on cloud-free conditions and overpass timing, allow for calibration of hydrodynamic models as well as bio-optical models. The development of such data assimilation systems for inland waters, combining process-based models—usually 2D horizontal or 3D hydrodynamic-biogeochemical models—are currently being investigated. While monitoring of algal bloom using remote sensing is broadly used, only a few real-world applications exist for coupling with process-based models for forecasting.

For the shallow Mantua Superior Lake, Italy, Pinardi et al. (2015) verified the simulation results for wind driven transport of chlorophyll-a concentration in the lake using a 3D hydrodynamic model compared to chlorophyll data derived from



**Fig. 15.9** Conceptual framework for combining remote sensing with modelling of harmful algal bloom [after Jöhnk et al. (2016)].

airborne and satellite remote sensed data. However, this application is restricted to assessing pure transport processes as it does not simulate algal growth.

Ongoing work at the artificial, urban Lake Burley Griffin, Australia, is developing a framework to integrate remote sensing data (Fig. 15.10) with a 3D hydrodynamic-biogeochemical model (Jöhnk et al. 2016) and in-situ measurements of bio-optical properties (Cherukuru et al. 2017). The availability of high spatial resolution images (Worldview-2 satellite, 2m resolution) for coloured dissolved organic matter (CDOM), non-algal particulate matter (NAP), and chlorophyll-a (Fig. 15.10) will allow scientists to forecast the development of cyanobacteria blooms on a timescale of 7 days based on initial remote sensing data and simulations driven by meteorological weather forecast.

For large lakes the spatial limitation of current satellite sensors is of minimal concern. Based on data from a geostationary satellite—Geostationary Ocean Color Imager (GOCI) launched by Korea—high temporal resolution images were used to follow algal blooms in Lake Taihu, China, with an hourly resolution (Huang et al. 2015). Basis for such a highly resolved series are the availability of in-situ data as prerequisite for calibrating the retrieval algorithms and cloud-free conditions.

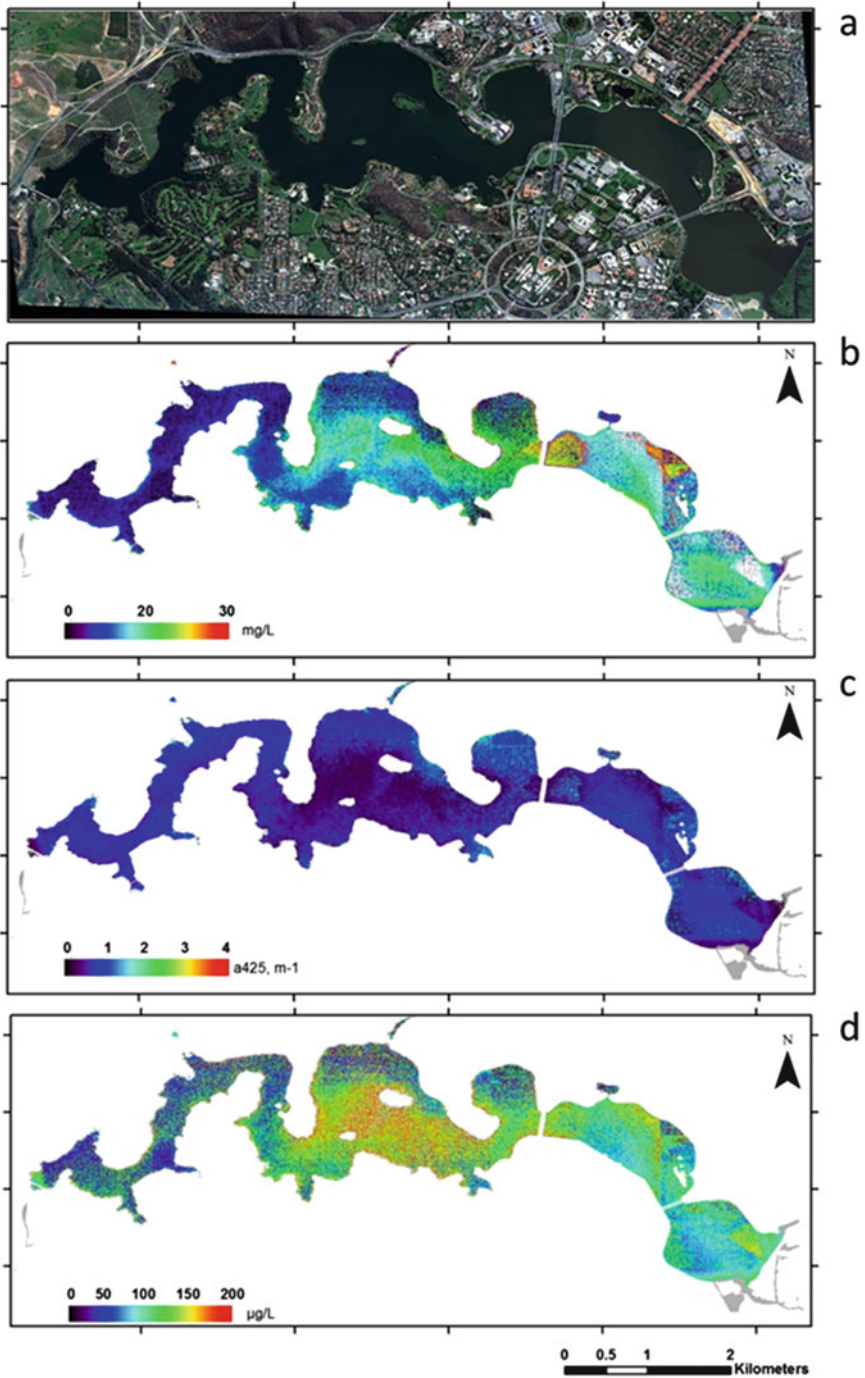
Lake Erie, USA, is another example of a large lake, where hydrodynamic models were used to predict seasonally occurring cyanobacteria blooms. A two-dimensional, vertically integrated circulation model was used to forecast the trajectory of blooms (Wynne et al. 2011). Combined with in-situ monitoring and remote sensing data (MERIS) this led to the development of an online system for bloom dynamics in Lake Erie (Wynne et al. 2013). This system developed by the NOAA's Great Lakes Environmental Research Laboratory is presently the only operational harmful algal bloom forecasting system known to us (Western Lake Erie HAB tracker: [https://www.glerl.noaa.gov/res/HABs\\_and\\_Hypoxia/habTracker.html](https://www.glerl.noaa.gov/res/HABs_and_Hypoxia/habTracker.html)).

The integration of remote sensing data with process-based models to forecast harmful algal blooms in inland waters is currently an active research area. With the launch of new, multispectral satellite sensors it can be expected that this field will rapidly expand and lead to new forecasting systems. It should be stressed, that process based models and algorithms to derive water quality data from satellite imagery need large databases of in-situ measurements to have predictive power. Thus, even with remote sensing allowing for a large-scale picture, field data are still necessary to ground truth Earth observation based models.

## 15.4 Conclusions

Operational forecasting and early warning of sudden occurring, unfavourable changes in ecosystems is prerequisite for prevention and mitigation of high economic and ecological costs. Novel computational and sensor technology becomes available to efficiently monitor and forecast such events like outbreaks of harmful cyanobacteria blooms.





**Fig. 15.10** Water quality parameters for Lake Burley Griffin, Australia, derived from Worldview 2 satellite. (a) True colour image for 17. March 2010, (b) CDOM—coloured dissolved organic matter characterized by a specific wavelength, (c) NAP—non algal particulate matter, (d) chlorophyll-a

Inferential models derived from *in situ* water quality data by evolutionary computation have been demonstrated to achieve up to 30-day-ahead forecasts of fast-growing concentrations of cyanobacteria cells and cyanotoxins in drinking water reservoirs. These models can be developed for different species of cyanobacteria and different cyanotoxins, and allow real-time early warning driven by online *in situ* water quality data monitored by multi-probe data loggers.

Future research focuses on generalising models of cyanobacteria and cyanotoxin species derived from lakes with similar climate, topology and trophic states. Libraries of species specific models categorised for lakes with similar properties will allow to share models across lakes with similar properties but insufficient historical data for on-site modelling.

Remote sensing allows to monitor spatio-temporal distribution of water quality parameters and cyanobacteria blooms based on sufficient spatial, temporal and spectral resolution of the sensors, and the availability of suitable algorithms to match satellite information with high-resolution in-situ measurements. Future research focuses on forecasting seasonal trajectories of harmful algal blooms by combining in-situ monitoring and remote sensing data with hydrodynamic models. By deriving vertical light attenuation in the water column from remote sensing data, hydrodynamic models will be enabled to predict seasonally occurring cyanobacteria blooms.

**Acknowledgements** Friedrich Recknagel wishes to thank Seqwater (Australia) and RAND Water (South Africa) for making available high quality limnological data of Lake Wivenhoe and Vaal Dam. He also is grateful for the support of his research by the Australian Research Council (Project Number LP0990453).

## References

- Bouvy M, Falcao D, Marinho M et al (2000) Occurrence of *Cylindrospermopsis* (Cyanobacteria) in 39 Brazilian tropical reservoirs during the 1998 drought. *Aquat Microb Ecol* 23:13–27
- Briand JF, Robillot C, Quiblier-Lloberas C et al (2002) Environmental context of *Cylindrospermopsis raciborskii* (Cyanobacteria) blooms in a shallow pond in France. *Water Res* 36:3183–3192
- Cao H, Recknagel F, Orr PT (2014) Parameter optimisation algorithms for evolving rule models applied to freshwater ecosystem. *IEEE Trans Evol Comput* 18:793–806
- Carmichael WW (1994) The toxins of cyanobacteria. *Sci Am* 270(1):78–86
- Cherukuru N, Malthus TJ, Sherman BS et al (2017) Optical response associated with changing summer biogeochemical conditions in a turbid lake. *Limnologia*. doi:[10.1016/j.limno.2017.01.009](https://doi.org/10.1016/j.limno.2017.01.009)
- Conradie RC, Barnard S (2012) The dynamics of toxic *Microcystis* strains and microcystin production in two hypertrophic South African reservoirs. *Harmful Algae* 20:1–10
- Dekker AG, Hestir EL (2012) Evaluating the feasibility of systematic inland water quality monitoring with satellite remote sensing, CSIRO, Canberra: Water for a Healthy Country National Research Flagship, 116 p
- Dörnhöfer K, Oepelt N (2016) Remote sensing for lake research and monitoring – recent advances. *Ecol Indic* 64:105–122

- Hawkins PR, Runnegar MTC, Jackson ARB et al (1985) Severe hepatotoxicity caused by the tropical cyanobacterium (blue-green alga) *Cylindrospermopsis raciborskii* (Woloszynska) Seenaya and Subba Raju isolated from a domestic supply reservoir. *Appl Environ Microbiol* 50:1292–1295
- Hestir EL, Brando VE, Bresciani M et al (2015) Measuring freshwater aquatic ecosystems: the need for a hyperspectral global mapping satellite mission. *Rem Sens Environ* 167:181–195
- Huang C, Shi K, Yang H et al (2015) Satellite observation of hourly dynamic characteristics of algae with Geostationary Ocean Color Imager (GOCI) data in Lake Taihu. *Rem Sens Environ* 159:278–287
- Huber V, Wagner C, Gerten D et al (2012) To bloom or not to bloom: contrasting responses of cyanobacteria to recent heat waves explained by critical thresholds of abiotic drivers. *Oecologia* 169:245–256
- Jöhnk KD, Cherukuru N, Anstee J et al (2016) Model-data assimilation framework for harmful algal bloom (CyanoHAB) prediction in inland waters on a continental scale. In: Webb JA et al (eds) *Proceedings of the 11th international symposium on ecohydraulics*. Melbourne, Australia, 7–12 Feb 2016. The University of Melbourne, ISBN:978 0 7340 5339 8
- Klemas V (2012) Remote sensing of algal blooms: an overview with case studies. *J Coastal Res* 28:34–43
- Kudela RM, Palacios SL, Austerberry DC et al (2015) Application of hyperspectral remote sensing to cyanobacterial blooms in inland waters. *Rem Sens Environ* 167:196–205
- Loftin KA, Graham JL, Hilborn ED et al (2016) Cyanotoxins in inland lakes of the United States: occurrence and potential recreational health risks in the EPA National Lakes Assessment 2007. *Harmful Algae* 56:77–90. doi:10.1016/j.hal.2016.04.001
- Lunetta RS, Schaeffer BA, Stumpf RP et al (2015) Evaluation of cyanobacteria cell count detection derived from MERIS imagery across the eastern USA. *Rem Sens Environ* 157:24–34
- Manzo C, Bresciani M, Giardino C et al (2015) Sensitivity analysis of a bio-optical model for Italian lakes focused on Landsat-8, Sentinel-2 and Sentinel-3. *Eur J Rem Sens* 48:17–32
- Matthews MW, Odermatt D (2015) Improved algorithm for routine monitoring of cyanobacteria and eutrophication in inland and near-coastal waters. *Rem Sens Environ* 156:374–382
- Moisander PH, Paerl HW, Zehr JP (2008) Effects of inorganic nitrogen on taxa specific cyanobacterial growth and *nifH* expression in a subtropical estuary. *Limnol Oceanogr* 53:2519–2532
- Mueller N, Lewis A, Roberts D et al (2015) Water observations from space: mapping surface water from 25 years of Landsat imagery across Australia. *Rem Sens Environ* 174:341–352
- Odermatt D, Giardino C, Heege T (2010) Chlorophyll retrieval with MERIS Case-2-Regional in perialpine lakes. *Rem Sens Environ* 114:607–617
- Orr PT, Rasmussen P, Burford MA et al (2010) Evaluation of quantitative real-time PCR to characterise spatial and temporal variations in cyanobacteria, *Cylindrospermopsis raciborskii* (Woloszynska) Seenaya et Subba Raju and *cylindrospermopsis* concentrations in three subtropical Australian reservoirs. *Harmful Algae* 9:243–254
- Palmer SC, Kutser T, Hunter PD (2015) Remote sensing of inland waters: challenges, progress and future directions. *Rem Sens Environ* 157:1–8. doi:10.1016/j.rse.2014.09.021
- Pinardi M, Fenocchi A, Giardino C et al (2015) Assessing potential algal blooms in a shallow fluvial lake by combining hydrodynamic modelling and remote-sensed images. *Water* 7:1921–1942
- Recknagel F, Ostrovsky I, Cao H et al (2014a) Hybrid evolutionary computation quantifies environmental thresholds for recurrent outbreaks of population density. *Ecol Inform* 24:85–89
- Recknagel F, Orr P, Cao H (2014b) Inductive reasoning and forecasting of population dynamics of *Cylindrospermopsis raciborskii* in three sub-tropical reservoirs by evolutionary computation. *Harmful Algae* 31:26–34
- Recknagel F, Orr P, Bartkow M et al (2017) Early warning of limit-exceeding concentrations of cyanobacteria and cyanotoxins in drinking water reservoirs by inferential modeling. *Harmful Algae* (in press)

- Scheffer M, Bascompte J, Brock WA et al (2009) Early-warning signals for critical transitions. *Nature* 461:53–59
- Stumpf RP, Davis TW, Wynne TT et al (2016) Challenges for mapping cyanotoxin patterns from remote sensing of cyanobacteria. *Harmful Algae* 54:160–173
- Toming K, Kutser T, Laas A et al (2016) First experiences in mapping lake water quality parameters with Sentinel-2 MSI imagery. *Rem Sens* 8:640
- World Health Organisation (WHO) (2003) Guidelines for safe recreational waters. Volume 1: coastal and fresh waters. WHO Publishing, Geneva, Switzerland
- Wynne TT, Stumpf RP, Tomlinson MC et al (2011) Estimating cyanobacterial bloom transport by coupling remotely sensed imagery and a hydrodynamic model. *Ecol Appl* 21:2709–2721
- Wynne TT, Stumpf RP, Tomlinson MC et al (2013) Evolution of a cyanobacterial bloom forecast system in western Lake Erie: development and initial evaluation. *J Great Lakes Res* 39:90–99
- Ye L, Cai Q, Zhang M et al (2014) Real-time observations, early warning and forecasting phytoplankton blooms by integrating in situ observations, online sondes and hybrid evolutionary algorithms. *Ecol Inform* 22:44–51

# Chapter 16

## Strategic Forecasting in Ecology by Inferential and Process-Based Models

Friedrich Recknagel, George Arhonditsis, Dong-Kyun Kim,  
and Hong Hanh Nguyen

**Abstract** Long-term forecasts are crucial for successful preventative and restorative management in ecology, and therefore require valid forecasting models. However, the validity of models is restricted by their scope and their inherent uncertainties.

This chapter discusses benefits of ensemble modelling in order to strengthen the validity and reliability of long-term forecasts. An ensemble of inferential models is demonstrated to overcome the limited scope of a single model for forecasting population dynamics of the cyanobacterium *Microcystis* in response to adaptive flow management of the River Nakdong (South Korea). Ensembles of alternative process-based models based on model averaging are examined to decrease uncertainties of single models when applied to determine the Remedial Action Plan for eutrophication control of Hamilton Harbour (Canada) and global warming effects on the phytoplankton community of Lake Engelsholm (Denmark). An ensemble of the complementary models SWAT and SALMO is applied to the catchment-reservoir system Millbrook (Australia) to overcome limitations of the scope of the two individual models. Results indicate that both, complex catchment-specific and lake-specific processes need to be considered in order to realistically forecast spatial cascading effects between catchments and lakes under the influence of prospective land use and climate changes.

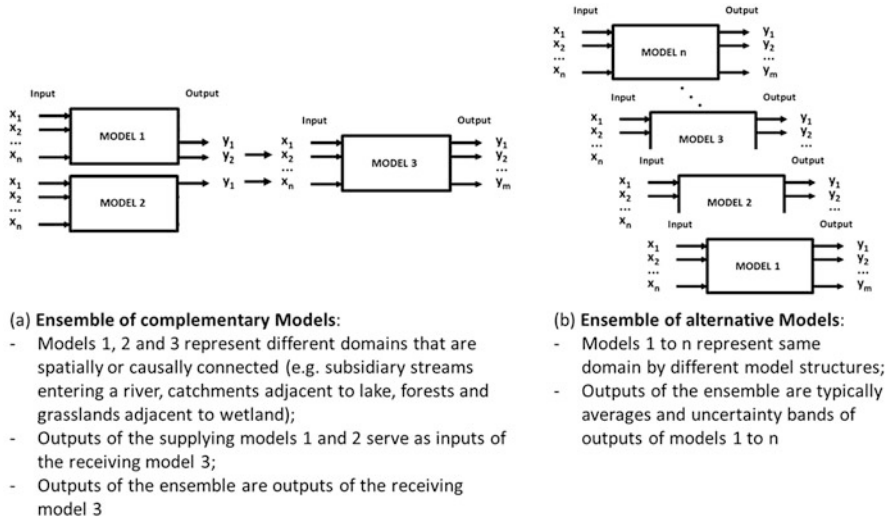
### 16.1 Introduction

Strategic or long-term forecasting is traditionally considered the domain of process-based models that simulate ‘what-if’ scenarios by running the process equations with scenario-specific parameter and input settings. Resulting state trajectories

---

F. Recknagel (✉) • H.H. Nguyen  
University of Adelaide, Adelaide, SA, Australia  
e-mail: [friedrich.recknagel@adelaide.edu.au](mailto:friedrich.recknagel@adelaide.edu.au); [hanh.nguyen@adelaide.edu.au](mailto:hanh.nguyen@adelaide.edu.au)

G. Arhonditsis • D.-K. Kim  
University of Toronto Scarborough, Scarborough, ON, Canada  
e-mail: [georgea@utsc.utoronto.ca](mailto:georgea@utsc.utoronto.ca); [dkkim1004@gmail.com](mailto:dkkim1004@gmail.com)



**Fig. 16.1** Rationale of ensembles of complementary (a) and alternative models (b)

display the likely scenario effect. However, scenario results from single process-based models may be limited by two factors: (1) their inherent uncertainty, and (2) their scope. Model ensembles as illustrated in Fig. 16.1 are one possible approach to addressing these limiting factors. Ensembles of alternative models may mitigate single model uncertainty (e.g., Ramin et al. 2012; Trolle et al. 2014), and ensembles of complementary models may extend the scope of single models.

By contrast, scenario analysis by inferential models may be limited by the fact that they lack the mechanistic foundation, and are directly driven by predictor variables. Again, ensembles of inferential models may address this limiting factor and would allow information to be cascaded between complementary models, and enable the simulation of nutrient cycles and community dynamics (e.g., Recknagel et al. 2014) as a prerequisite for scenario analysis.

This chapter discusses novel approaches for scenario analysis based on ensembles of inferential and process-based models.

## 16.2 Scenario Analysis by Inferential Models

This case study is based on an ensemble of inferential models that have been developed from hybrid evolutionary algorithms HEA (Cao et al. 2014, see also Chap. 9). The ensemble is applied to test the hypothesis that seasonally altered flow regimes in the River Nakdong (Korea) can be used to control population growth of *Microcystis aeruginosa* by changing water residence time and water quality. In

order to test the hypothesis, two flow regimes were created from historical data that maintain the base flow above the threshold of  $350 \text{ m}^3 \text{ s}^{-1}$  in winter, and limit the peak flow in summer to  $700 \text{ m}^3 \text{ s}^{-1}$ . These flow regimes can be practically managed by maintaining the recommended flow level during the dry winter season by releasing additional water from adjacent dams, and by maintaining the recommended peak flow limit in summer by releasing less water from dams (e.g. Jeong et al. 2007; Hong et al. 2014).

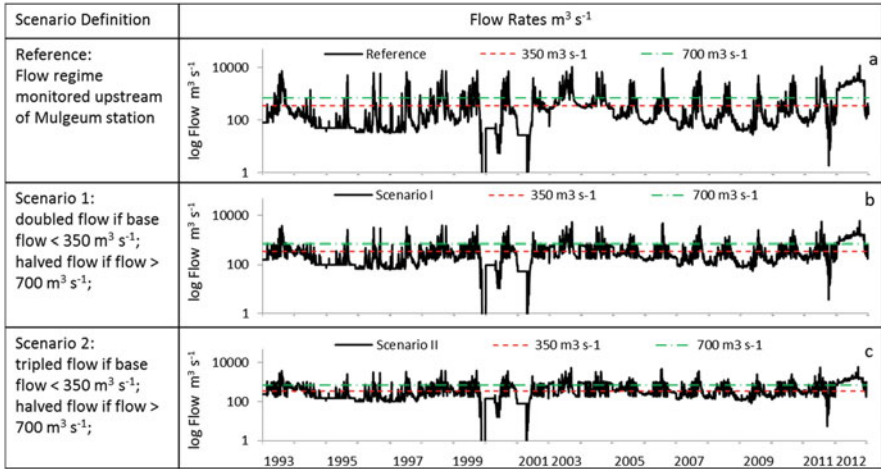
The Nakdong River stretches 525 km across South Korea with a catchment area of  $23,380 \text{ km}^2$  and is regulated by dams that supply irrigation and drinking water to adjacent communities. The river has a history of phytoplankton proliferation dominated by cyanobacteria in summer and diatoms in winter (Ha et al. 1999, 2003; see also LTER Nakdong River in Chap. 20). Weekly to monthly water quality data and algae cell counts monitored at Mulgeum Station of the Nakdong River from 1993 to 2012 (see Table 16.1) have been linearly interpolated for daily time steps before modeling *Microcystis aeruginosa* using the hybrid evolutionary algorithm HEA.

Scenarios have been based on the observation that  $350 \text{ m}^3 \text{ s}^{-1}$  is the flow threshold above which chlorophyll *a* concentrations decline significantly (Hong et al. 2014). Accordingly, scenario 1 assumes a twofold increase and scenario 2 assumes a threefold increase of flow that is below the threshold, whereby flow rates exceeding  $700 \text{ m}^3 \text{ s}^{-1}$  ( $\approx 83$ rd percentile of the river flow) have been halved (Fig. 16.2).

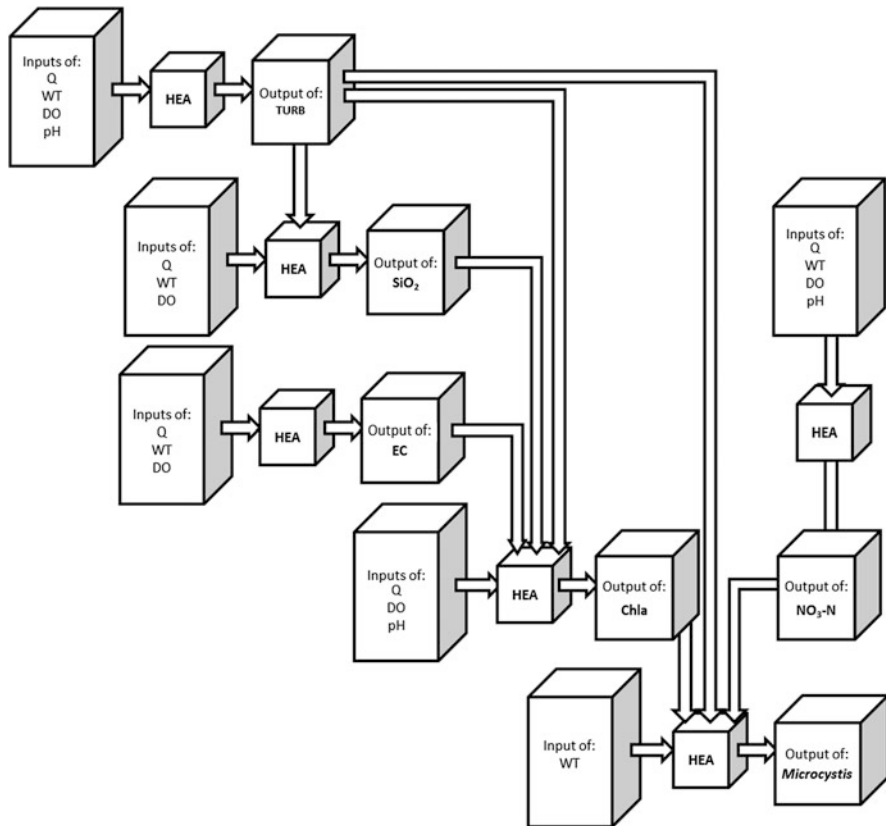
Figure 16.3 illustrates the model ensemble that has been designed for analysing alternative flow scenarios for the cyanobacterium *Microcystis* based on the best-performing model for 5-day-ahead forecasts of *Microcystis* (see Fig. 16.4a). The model achieved a coefficient of determination  $r^2 = 0.94$  and identified the water

**Table 16.1** Statistical summary of limnological data from Nakdong River monitored from 1993 to 2012

Variable	Unit	Min	Max	Mean	SD	Coefficient of variance (%)
Water temperature WT	°C	0	34.4	16.4	8.7	53
Dissolved oxygen DO	mg L <sup>-1</sup>	2.5	24.7	10.9	3.9	36
pH		6.27	10.73	8.24	0.82	10
Electrical conductivity EC	µS cm <sup>-1</sup>	10	670	292	116	40
Turbidity TURB	NTU	1.6	648.0	17.4	43.1	248
Flow rate Q	m <sup>3</sup> s <sup>-1</sup>	0.1	11,996.9	527.9	1009.9	191
Nitrate NO <sub>3</sub> -N	mg L <sup>-1</sup>	0.05	5.62	2.54	0.86	34
Phosphate PO <sub>4</sub> -P	µg L <sup>-1</sup>	2	1114	56	53	94
Silica SiO <sub>2</sub>	mg L <sup>-1</sup>	0.01	21.64	5.41	3.90	72
Chlorophyll <i>a</i> Chl- <i>a</i>	µg L <sup>-1</sup>	0.4	1035.0	34.9	61.0	175
<i>Microcystis aeruginosa</i>	cells mL <sup>-1</sup>	0	9,500,837	49,420	479,907	971



**Fig. 16.2** Definition of flow scenarios 1 and 2 in relation to the reference flow regime monitored at Mulgeum Station of the Nakdong River from 1993 to 2012 (Recknagel et al. 2017)



**Fig. 16.3** Model ensemble for simulating flow scenarios for *Microcystis* (Recknagel et al. 2017)



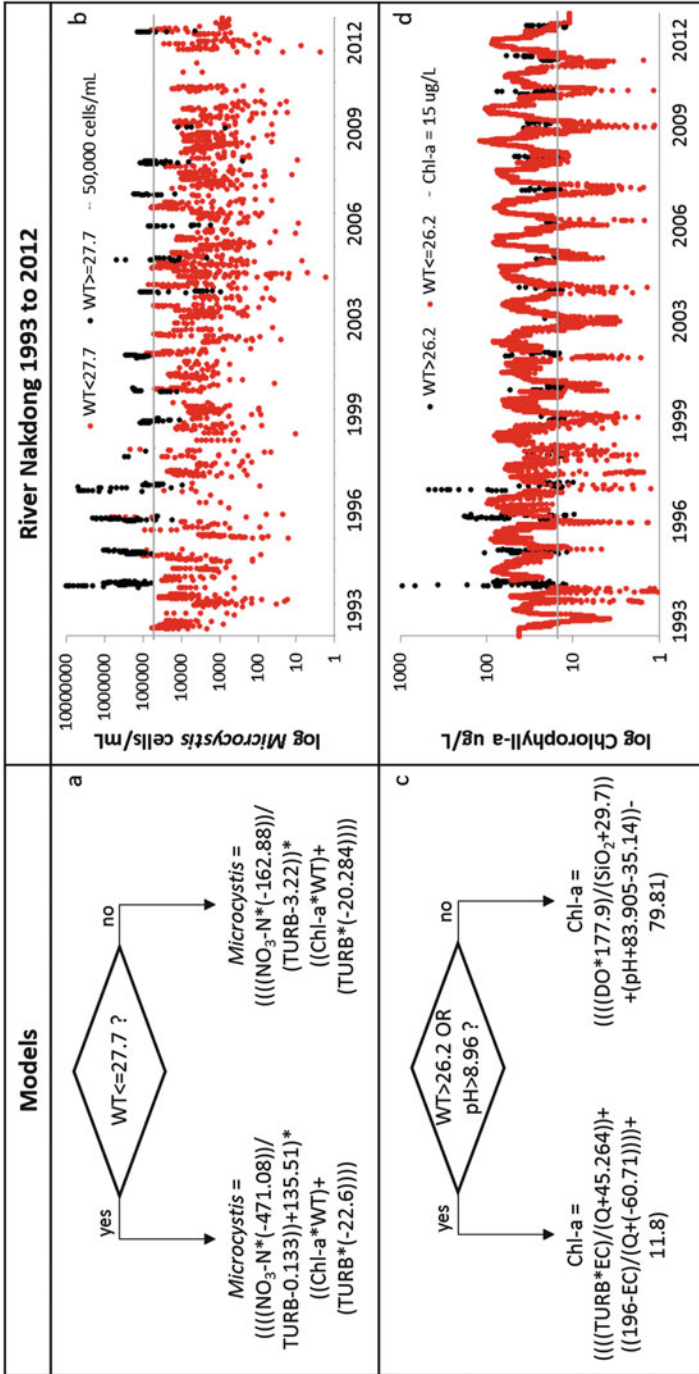


Fig. 16.4 Illustration of IF-THEN-ELSE models and threshold conditions for *Microcystis* (a, b) and chlorophyll-a (c, d) (Recknagel et al. 2017, modified)

temperature of 27.7°C as threshold above which the model forecasts high population density of *Microcystis* of greater than 50,000 cells mL<sup>-1</sup> (see Fig. 16.4b). The temperature threshold corresponds well with literature findings suggesting that *Microcystis* tend to have optimum growth rates above 25 °C (e.g. Robarts and Zohary 1987).

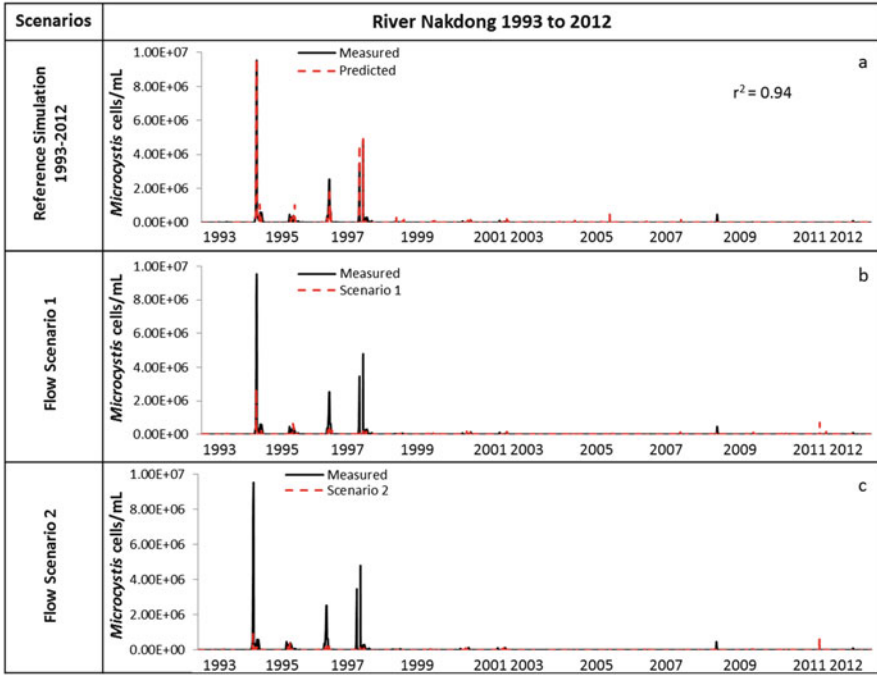
The predictor variable water temperature (WT) of the *Microcystis* model is considered to be least affected by flow and has therefore been maintained unchanged for the scenario analysis. However, the predictor variables turbidity (TURB), nitrate (NO<sub>3</sub>-N), and chlorophyll *a* (Chl-*a*) are expected to be flow-dependent. They have been represented by separate forecasting models (see Table 16.2 and Fig. 16.4c), and incorporated into the model ensemble for *Microcystis* (Fig. 16.3). The model for TURB included WT, dissolved oxygen (DO) and pH as predictor variables which were considered to be flow-independent and therefore maintained unchanged for the scenario analysis. The model for NO<sub>3</sub>-N was based on the predictor variables WT, DO and pH which remained unchanged for the scenario analysis. However, the Chl-*a* model included the flow-dependent predictor variables electrical conductivity (EC) and silica (SiO<sub>2</sub>) for which forecasting models have also been developed (see Table 16.2).

The Chl-*a* model achieved an  $r^2 = 0.54$  (Fig. 16.4c) suggesting water temperatures greater than 26.2 °C and pH values greater than 9.6 as threshold conditions for forecasting Chl-*a* concentrations greater than 15 µg L<sup>-1</sup> (see Fig. 16.4d). Both thresholds indicate that Chl-*a* in the River Nakdong is seasonally correlated with the cyanobacterium *Microcystis*, which is dominating in summer at optimum water temperatures greater than 25 °C (see above), and causing pH values to rise above 8.5 (Reynolds 2006).

Whilst bloom events of *Microcystis* with more than 50,000 cells ml<sup>-1</sup> occurred frequently in River Nakdong, major blooms in 1994, 1996 and 1997 exceeding

**Table 16.2** Documentation of models for turbidity TURB, electrical conductivity EC, phosphate PO<sub>4</sub>-P, nitrate NO<sub>3</sub>-N and silica SiO<sub>2</sub>

Models	$r^2$
IF(pH/371.7)*Q<=72 THEN <b>TURB</b> = ((-25.85/(WT-33.48)+((Q/15.3)+84.66)/DO)) ELSE <b>TURB</b> = (((-489.97/(WT-17.15))/(WT-23.88)-433.49/(WT-31.65))	0.31
IF (DO>=19.5 AND Q>=46.53) OR (Q>=30.15 AND (pH-Q> -28.61)) THEN <b>EC</b> = (451.96-((-16.32-(WT*(-0.536)))*ln((Q-88.43)))) ELSE <b>EC</b> = (410.95-((18.05-(WT*(-0.536)))*ln((Q-44.83))))	0.54
IF (TURB<8.2 OR Q<=136.5) OR (WT<28.7 OR (TURB<97.6 AND TURB>=48.5)) THEN <b>PO<sub>4</sub>-P</b> = ((3.2/(246.9-(WT*27.4)))+54.3-(41.22/(TURB-88.185))) ELSE <b>PO<sub>4</sub>-P</b> = (((TURB/0.36)-DO)+62.89)	0.37
IF pH>=9.1 OR Q<76.7 THEN <b>NO<sub>3</sub>-N</b> = (((3.364-(WT/19.24))-0.052)-((31.93/(25.987-Q))/107.28)) ELSE <b>NO<sub>3</sub>-N</b> = (((3.127-(WT/29.1))-(37.6/(Q+DO)))-((1.84/(7.01-pH))/173.48))	0.33
IF Q>395.35 THEN <b>SiO<sub>2</sub></b> = ln(((WT*TURB)*(WT*TURB))*((-208.55/DO)/Q)) ELSE <b>SiO<sub>2</sub></b> = ln(((Q*0.03)*(Q*0.102)+(277.9/(WT-0.367))))	0.37



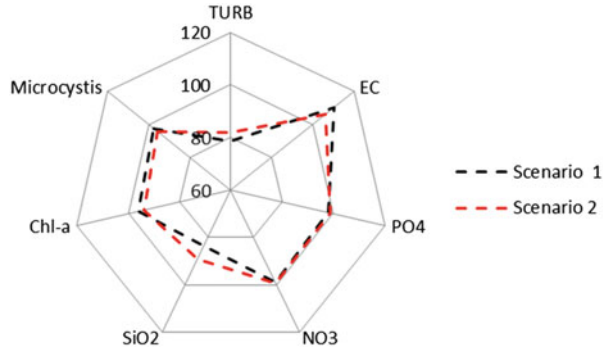
**Fig. 16.5** 5-day-ahead forecasting of *Microcystis* in River Nakdong from 1993 to 2012 based on: (a) reference flow (Fig. 16.2a), (b) flow scenario 1 (Fig. 16.2b), (c) flow scenario 2 (Fig. 16.2c) (Recknagel et al. 2017)

1 million cells  $\text{ml}^{-1}$  (Fig. 16.5a) were of particular interest and have been forecasted accurately in terms of timing and magnitude by the model documented in Fig. 16.4a. The scenario analysis by means of the model ensemble in Fig. 16.3 predicted a 70% lower magnitude of the *Microcystis* bloom in 1994 and the prevention of major bloom events in 1996 and 1997 (Fig. 16.5b) by the flow regime 1 (see Fig. 16.2b) and the prevention of all 3 bloom events (Fig. 16.5c) by the flow regime 2 (Fig. 16.2c).

Figure 16.6 represents results of the scenario analysis in terms of percentage of average change of water quality parameters in response to the two flow scenarios forecasted by the models for TURB, EC,  $\text{PO}_4\text{-P}$ ,  $\text{NO}_3\text{-N}$  and  $\text{SiO}_2$  as well as the models in Fig. 16.4. It illustrates that EC increased up to 110% under the influence of the flow scenarios while concentrations of  $\text{PO}_4$ ,  $\text{NO}_3$  and  $\text{SiO}_2$  decreased. Turbidity decreased most significantly to almost 80% whilst Chl-a diminished up to 96%. Figure 16.6 provides further evidence that altered flow regimes can prevent extreme algal blooms in River Nakdong reflected by decline of the average population density of *Microcystis* to 96%.

In summary, the case study has demonstrated that ensembles of inferential models allow scenario analysis of complex systems such as population dynamics

**Fig. 16.6** Comparison of percentage changes of water quality parameters and phytoplankton forecasted in response to scenarios 1 and 2 (Recknagel et al. 2017, modified)



of the cyanobacterium *Microcystis* in response to river flow regimes. River flow influences *Microcystis* growth not only directly by changing water residence time, but also indirectly by altering predictor variables such as turbidity, conductivity, nutrient and chlorophyll-a concentrations. It therefore proved to be sensible to model these ‘indirect’ predictor variables affected by flow first before they feed into the *Microcystis* model in combination with the ‘direct’ predictor variable flow. The so-designed model ensemble enabled cascading effects of changed flow regimes to be simulated through the network of predictor variables for *Microcystis*. Results of the scenario analysis suggest that managed flow is a viable option for controlling blooms of *Microcystis* in the River Nakdong. The full study including a model ensemble for predicting winter blooms of *Stephanodiscus* has been documented in Recknagel et al. (2017).

### 16.3 Scenario Analysis by Process-Based Models

Recognizing that there is no true model of an ecological system, but rather several adequate descriptions of different conceptual basis and structure, simulation libraries and ensemble modeling may mitigate uncertainty inherent in the model selection process. Environmental management decisions relying on a single inadequate model can introduce bias and uncertainty that is much larger than the error stemming from the erroneous choice of model parameter values (Neuman 2003). Basing ecological forecasts on a single mathematical model implies that a valid alternative model may be rejected (or omitted) from the decision making process (Type I model error), but also that projections can potentially result from a flawed mathematical construct that was not rejected in an earlier stage (Type II model error).

The simulation library SALMO-OO (Recknagel et al. 2008a) is an object-oriented implementation of the lake model SALMO (see Chap. 10) that provides access to alternative process representations for algal growth and grazing as well as zooplankton growth and mortality adopted from Arhonditsis and Brett (2005),

Hongping and Jianyi (2002), and Park et al. (1974), which are integrated into the SALMO framework. Keeping the selection of these process representations optional, the aim is to give SALMO-OO structural flexibility in order to assemble “best-fit” model structures for particular applications (Recknagel et al. 2008b).

In this section we demonstrate how ensembles of alternative and consecutive models can improve the validity of scenario analyses.

### ***16.3.1 Ensemble of Alternative Models Based on Bayesian Model Averaging***

Bayesian Model Averaging (BMA) is a technique designed to integrate across many different competing models, thereby incorporating the uncertainty about the optimal model for any given exercise into the inference drawn about parameters and predictions (Raftery et al. 2005). Thus, rather than picking the single “best-fit” model to predict future system responses, we can use Bayesian model averaging to provide a weighted average of the forecasts from different models (Hoeting et al. 1999). In weather forecasting, BMA has offered a strategy for statistical post-processing of ensemble outputs, thereby achieving lower predictive error and sharper predictive probability density functions (Bao et al. 2010; Sloughter et al. 2007, 2010). In the context of eutrophication, despite the increasing number of process-based modeling studies that have adopted uncertainty analysis techniques (Arhonditsis et al. 2007, 2008a, b; Law et al. 2009; Ramin et al. 2011; McDonald et al. 2012), there is an overwhelming gap in the literature of ensemble approaches to guide risk assessment. Moreover, there has been little focus on the benefits of basing ecological forecasts on combinations of process-based models, and practically no discussion on the ways that outputs of mathematical models with multiple endpoints (state variables) and derived quantities (process rates) can be objectively integrated into a single averaged prediction.

Ramin et al. (2012) recently examined the potential benefits for model-based environmental management when a combination of models of different complexity is being used. In particular, predictions drawn from a simple plankton model were synthesized with those provided by a complex ecosystem model in order to guide the water quality criteria setting process in Hamilton Harbour (Ontario, Canada; see Chap. 11). The former (simpler) model accounted for the basic processes underlying the interplay among phosphate, detritus, and generic phytoplankton and zooplankton state variables, such as phosphate uptake, grazing, metabolic losses, phosphorus recycling, and sedimentation (Fig. 16.7). The complex eutrophication model reproduced the interactions among a generic phytoplankton group, a “cyanobacteria-like” phytoplankton, and zooplankton with the nitrogen and phosphorus cycles. The latter model also considered a dynamic causal association between nutrient release rates from the sediments and particulate fluxes from the water column.

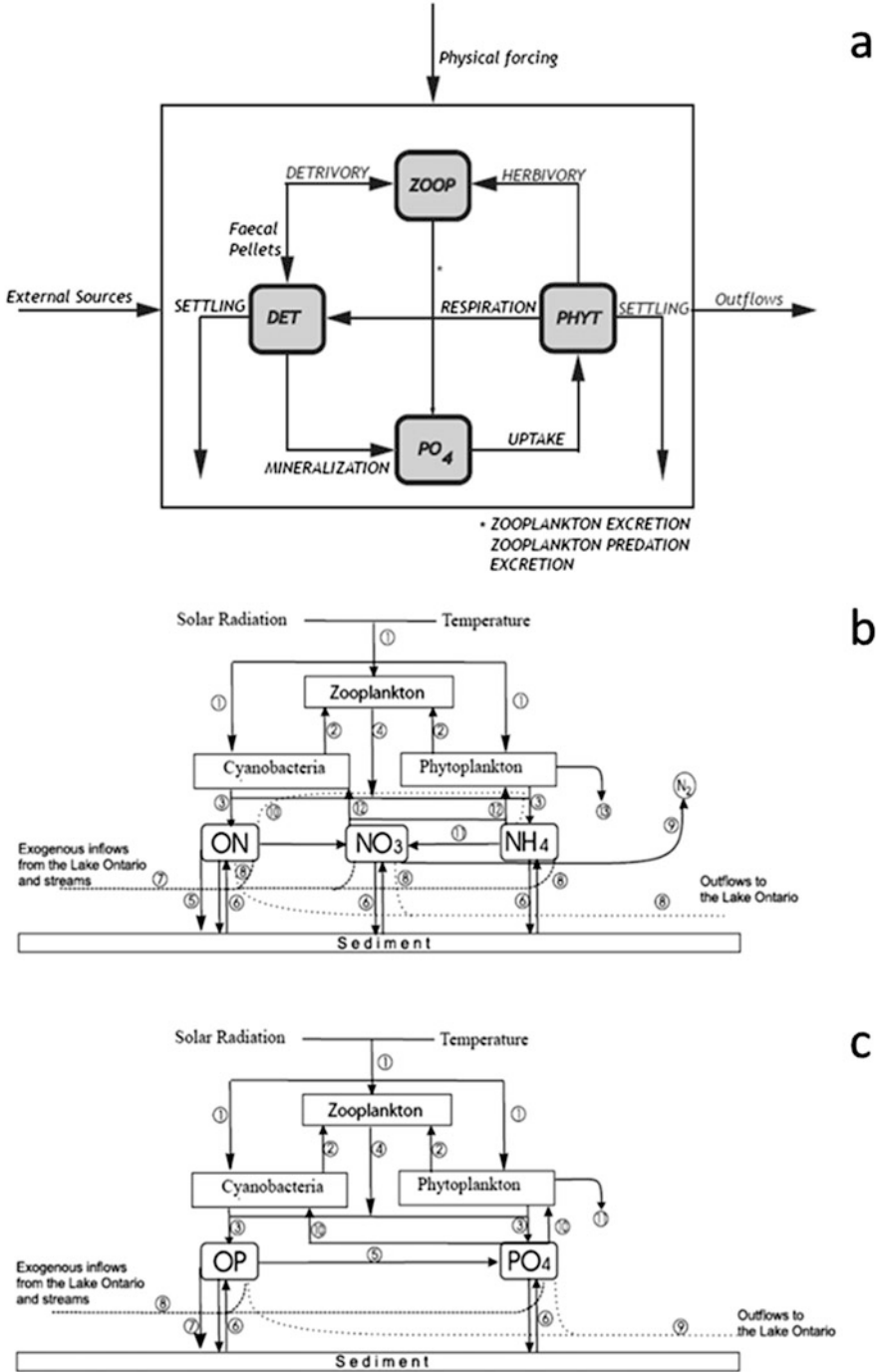


Fig. 16.7 The flow diagram of the phosphate ( $PO_4$ )–phytoplankton ( $PHYT$ )–zooplankton ( $ZOOP$ )–detritus or particulate phosphorus ( $DET$ ), also referred to as NPZD model (a). The nitrogen biogeochemical cycle of the model: (1) external forcing to phytoplankton growth

Using a Bayesian framework (see also Chap. 11), the two ecological models were calibrated independently against the water quality conditions currently prevailing in the Hamilton Harbour. A sequence of realizations from the posterior distribution of the models was obtained using Markov chain Monte Carlo (MCMC) simulations (Gilks et al. 1998). Quantifying model performance in terms of the magnitude of the structural (or process) error terms for each state variable with available calibration data indicated that the complex model outperformed the simple one (Ramin et al. 2012, see Table 16.3). In a post-hoc sensitivity analysis test, the ability of the latter model to support predictions outside its calibration domain was examined against the empirical relationships among annual phosphorus loading, summer total phosphorus (TP), and chlorophyll *a* concentrations historically recorded in the Hamilton Harbour (Ramin et al. 2012). However, because of the uncertainty of the year-specific loading values from the early 1990s, when the system was hyper-eutrophic, a predictive validation exercise to examine the credibility of the model to reproduce year-to-year variations was not undertaken, and thus the likelihood of overfitting the data with the complex model is not entirely ruled out (Gudimov et al. 2010; Ramin et al. 2011).

**Table 16.3** Markov chain Monte Carlo estimates of the mean values and standard deviations (*SD*) of the model structural (or process) error for the different state variables of the two eutrophication models

Parameters	Simple (NPZD)		Complex	
	Mean	SD	Mean	SD
$\sigma_{PO_4epi}$	1.732	0.457	0.287	0.101
$\sigma_{PHYTepi}$	205.5	52.51	–	–
$\sigma_{ZOOPEpi}$	55.47	19.96	13.86	3.466
$\sigma_{DET/OPepi}$	1.083	0.421	2.221	0.599
$\sigma_{NH_4epi}$	–	–	48.56	11.71
$\sigma_{NO_3epi}$	–	–	240.5	66.52
$\sigma_{CYAepi}$	–	–	111.7	29.76
$\sigma_{NONCYAepi}$	–	–	52.32	14.71
$\sigma_{PO_4hypo}$	2.261	0.533	0.834	0.218
$\sigma_{DET/OPhypo}$	3.794	0.898	2.212	0.554
$\sigma_{NH_4hypo}$	–	–	14.81	5.33
$\sigma_{NO_3hypo}$	–	–	268.2	64.39



**Fig. 16.7** (continued) (temperature, solar radiation); (2) zooplankton grazing; (3) phytoplankton basal metabolism excreted as  $NH_4$  (Ammonium) and  $ON$  (Organic Nitrogen); (4) zooplankton basal metabolism excreted as  $NH_4$  and  $ON$ ; (5) settling of particles; (6) water sediment  $NO_3$  (Nitrate),  $NH_4$ , and  $ON$  exchanges; (7) exogenous inflows of  $NO_3$ ,  $NH_4$ , and  $ON$ ; (8) outflows of  $NO_3$ ,  $NH_4$ , and  $ON$ ; (9)  $NO_3$  sinks due to denitrification; (10)  $ON$  mineralization; (11) nitrification; and (12) phytoplankton uptake (b). The phosphorus biogeochemical cycle of the model: (1) external forcing to phytoplankton growth (temperature, solar radiation); (2) zooplankton grazing; (3) phytoplankton basal metabolism excreted as  $PO_4$  (Phosphate) and  $OP$  (Organic Phosphorus); (4) zooplankton basal metabolism excreted as  $PO_4$  and  $OP$ ; (5)  $OP$  mineralization; (6) water sediment  $PO_4$  and  $OP$  exchanges; (7) settling of particles; (8) exogenous inflows of  $PO_4$  and  $OP$ ; and (9) outflows of  $PO_4$  and  $OP$  (c)

Furthermore, drawing parallels between the parameter posteriors of the two models, the study identified their similarities with respect to the ecological characterization of the planktonic food web of the studied system. In particular, the generic phytoplankton group in both models was assigned high maximum phytoplankton growth rates ( $>2 \text{ day}^{-1}$ ), fast response to light availability (i.e., half saturation light intensity for phytoplankton  $<140 \text{ MJ m}^{-2} \text{ day}^{-1}$ ), fast phosphorus kinetics (i.e., half saturation constant  $<10 \text{ } \mu\text{g P L}^{-1}$ ), and high maximum uptake rates ( $0.02 \text{ } \mu\text{g P L}^{-1} \text{ day}^{-1}$ ). Likewise, the updating of the two models resulted in similar zooplankton grazing ( $\approx 0.5 \text{ day}^{-1}$ ) and mortality rates ( $0.11\text{--}0.15 \text{ day}^{-1}$ ) as well as sedimentation rates of particulate matter ( $>0.4 \text{ m day}^{-1}$ ), and relative importance of the two factors that determine the illumination of the water column, i.e., the light extinction due to chlorophyll *a* ( $0.02\text{--}0.03 \text{ L } \mu\text{g chl}a^{-1} \text{ m}^{-1}$ ), and the background light attenuation ( $\approx 0.2 \text{ m}^{-1}$ ).

After the calibration exercise, the MCMC estimates of the mean and standard deviation parameter values along with their covariance structure were used to update the two models (Gelman et al. 2013). The updated models provided the basis for long-term forecasting through a series of posterior simulations aiming to examine the compliance of the system with targeted water quality standards,  $20 \text{ } \mu\text{g TP L}^{-1}$  and  $10 \text{ } \mu\text{g chl}a \text{ L}^{-1}$ , under reduced nutrient loading conditions (Ramin et al. 2011). Predictions from the two models were also combined to obtain averaged forecasts from the two ecological characterizations of the system. One of the critical decisions when considering models of different complexity involves the selection of the averaging scheme to synthesize their predictions (Lindström et al. 2015). Ramin et al. (2012) opted for a strategy that considers performance over all the model endpoints rather than the subset of state variables included in both models or the variables more closely related to the environmental management problem at hand. Thus, the adopted strategy used the respective mean process error values as weights in a weighted model average:

$$w_{ij} = \frac{\sum_{k=1}^{MC} \frac{\sigma_{ijk}}{\bar{Y}_j}}{MC} \quad (16.1)$$

$$w_{Mi} = \frac{m}{\sum_{j=1}^m w_{ij}} \quad (16.2)$$

$$\overline{TP} = \sum_{i=1}^l w_{Mi} TP_{Mi} \quad \overline{chl}a = \sum_{i=1}^l w_{Mi} chla_{Mi} \quad (16.3)$$

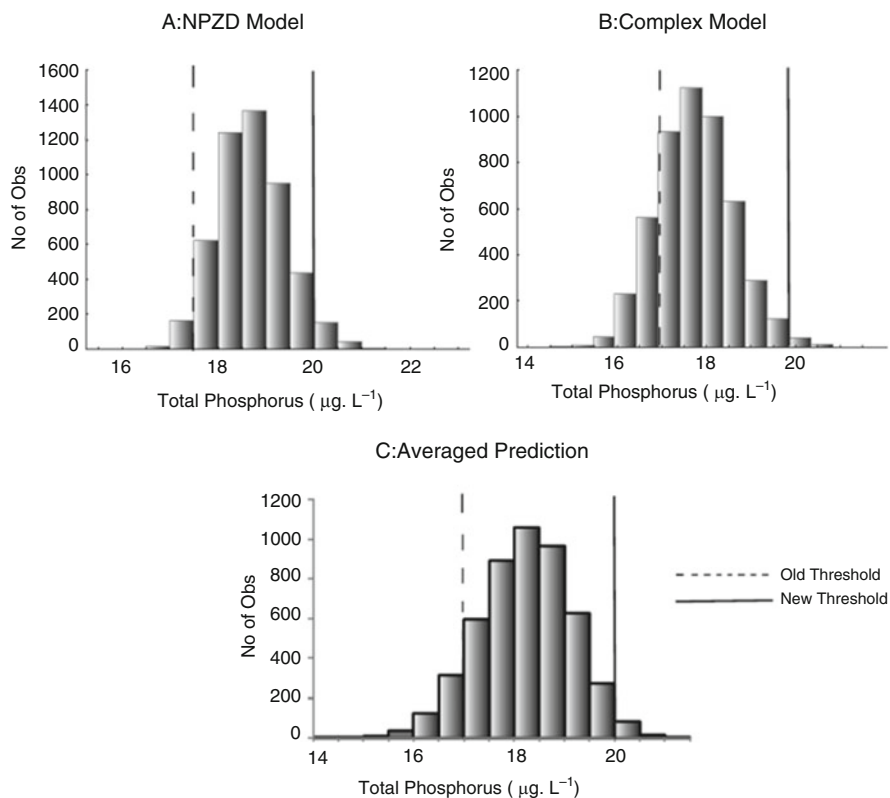
where:  $l$  represents the number of models considered in this analysis ( $l = 2$ );  $m$  corresponds to the number of state variables  $j$  of the model  $M_i$  for which data are available ( $m = 6$  or  $11$ );  $MC$  is the total number of MCMC runs sampled to form the model posteriors;  $\sigma_{ijk}$  denotes the model structural error for the state variable  $j$  of the model  $M_i$  as sampled from the MCMC run  $k$ ;  $\bar{Y}_j$  represents the annual observed



average for the variable  $j$ ,  $TP_{Mi}$  and  $chl a_{Mi}$  are the total phosphorus and chlorophyll  $a$  predictions from the individual models weighted by the corresponding weights  $w_{Mi}$  to obtain the averaged predictions  $\overline{TP}$  and  $\overline{chl a}$ . This weighting scheme entails the risk of downplaying the impact of the best performing model for a particular variable, but also reflects the notion that all models integrated in an ensemble ecological forecast should demonstrate balanced performance over their entire structure. In particular, this approach aims to penalize the likelihood of calibration bias, whereby the maximization of the fit for a specific state variable (e.g., phytoplankton biomass, dissolved oxygen) may be accompanied by high error for other state variables (herbivorous zooplankton biomass, nutrient concentrations), and thus to avoid forecasts founded on models with misleadingly high weights that downplay fundamentally flawed ecological structures (Franks 1995; Arhonditsis and Brett 2004).

Regarding the nutrient loading scenario examined with the two updated models, both the simple model ( $18.7 \pm 0.7 \mu\text{g TP L}^{-1}$ ) and the complex one ( $17.8 \pm 0.9 \mu\text{g TP L}^{-1}$ ) predicted that the average TP concentrations during the summer stratified period will fall below the level of  $20 \mu\text{g TP L}^{-1}$ , if the exogenous phosphorus loading is reduced to  $142 \text{ kg day}^{-1}$  (Fig. 16.8a, b). The complete agreement between the two forecasts for total phosphorus is also reflected in their averaged prediction (Fig. 16.8c). Both models also predict that the epilimnetic chlorophyll  $a$  concentrations will fall below the threshold level of  $10 \mu\text{g chl a L}^{-1}$  (Fig. 16.9a, b). Nonetheless, the simple model appears to support more optimistic predictions with respect to phytoplankton response to the reduced ambient TP concentrations relative to the complex one. Consequently, the averaged predictive distribution for chlorophyll  $a$  demonstrates a distinct bimodal pattern with a primary mode at  $7.5 \mu\text{g chl a L}^{-1}$ , reflecting the greater weight placed on the complex model, and a secondary peak at  $5.1 \mu\text{g chl a L}^{-1}$ , associated with the simple one (Fig. 16.9c). The fact that both models predict the achievability of the water quality standard related to the mean chlorophyll  $a$  concentrations ( $<10 \mu\text{g L}^{-1}$ ) in the Hamilton Harbour is certainly encouraging; nonetheless, the more conservative predictions of the complex ecosystem model invite investigation of the factors that could be driving this discrepancy.

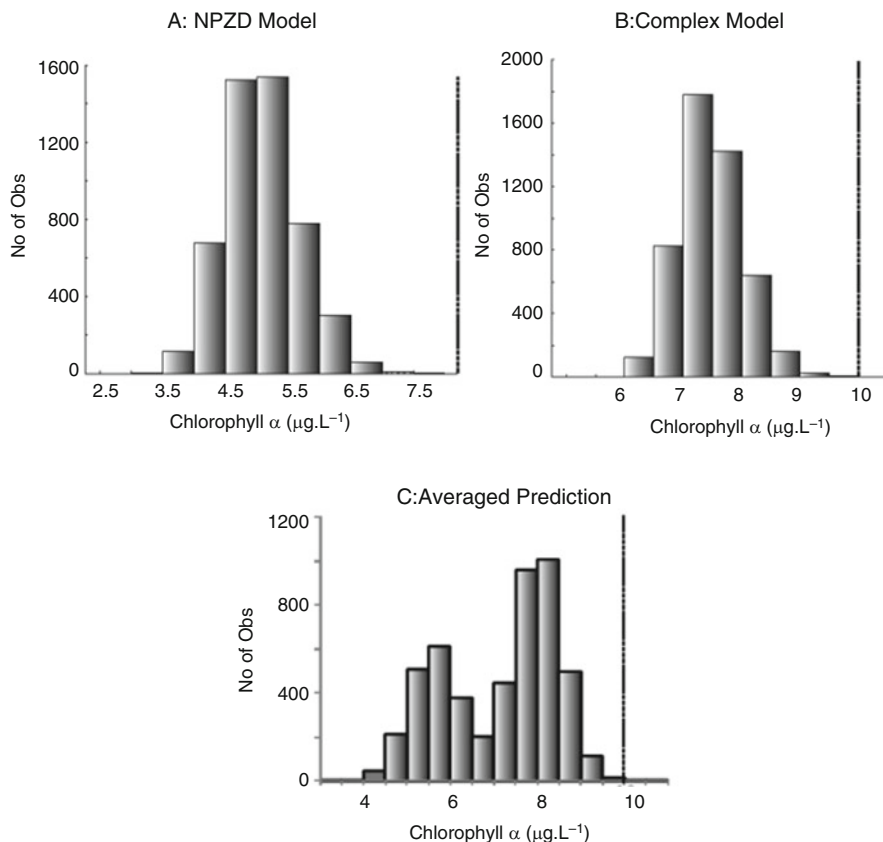
To this end, one of the major structural differences between the two models lies in the way the nutrient fluxes from the sediments are treated, i.e., a user-specified temperature-dependent phosphorus flux rate vis-à-vis a dynamic characterization of the phosphorus release as a function of the particulate sedimentation and burial rates (Ramin et al. 2011). The simple model predicts that the sediments contribute approximately  $1.1 \text{ mg P m}^2 \text{ day}^{-1}$  into the overlying water column, whereas the same fluxes are increased to  $2.0 \text{ mg P m}^2 \text{ day}^{-1}$  with the complex model. Empirical evidence from the system suggests that upward diffusive phosphate fluxes in the Harbour are closer to the latter estimate, as they can reach the level of  $1.7 \text{ mg m}^2 \text{ day}^{-1}$  (Azcue et al. 1998). Under the reduced nutrient loading scenario, the dynamic nature of the sediment response with the complex model decreases the phosphorus release at the level of  $1.5 \text{ mg m}^2 \text{ day}^{-1}$ , which is still higher than the flux used to force the simple model. Using a temperature-dependent phosphorus



**Fig. 16.8** Predictions of the epilimnetic summer total phosphorus concentrations, under the proposed nutrient loading reductions by the Hamilton Harbour Remedial Action Plan (RAP), based on the two eutrophication models (a, b) and their averaged predictions (c). Old threshold refers to the 17 µg TP L<sup>-1</sup> standard, while the new delisting criterion sets the water quality target at 20 µg TP L<sup>-1</sup> (Gudimov et al. 2010, 2011; Ramin et al. 2011, 2012)

release rate to reproduce the sediment-water column interactions, which is then treated as an inverse problem (i.e., data for the dependent variables are used to specify the values of model parameters), likely oversimplifies this facet of the ecosystem functioning. Thus, the discrepancy between the two models pinpoints a structural weakness in the simple model and also highlights the importance of embracing more sophisticated modeling strategies to shed light on the sediment diagenesis processes in the Hamilton Harbour (Gudimov et al. 2016). Bearing in mind that the Occam's razor suggests a shift towards simpler theories until simplicity can be gradually traded for increased predictive capacity (Jaynes 1994), the consideration of more than one model for environmental management problems can be particularly useful. This practice offers an opportunity to identify areas where extra complexity should be invoked and knowledge gaps that can be critical for increasing the credibility of our ecological forecasts.

The Hamilton Harbour case study in principle reflects our lack of confidence in mechanistic modeling to support reliable "real-time" forecasts. After the training



**Fig. 16.9** Predictions of the epilimnetic summer chlorophyll *a* concentrations, under the proposed nutrient loading reductions by the Hamilton Harbour RAP, based on the two eutrophication models (a, b) and their averaged predictions (c)

and validation phase, we typically opt for analysis of long-term ecological scenarios aiming to address questions of the type “*What would happen if...?*” while the derived projections are conditioned on the model assumptions, residual error, and/or associated uncertainty. The novel feature of this study is the explicit recognition of the uncertainty pertaining to the selection of the optimal model structure for a specific environmental management problem.

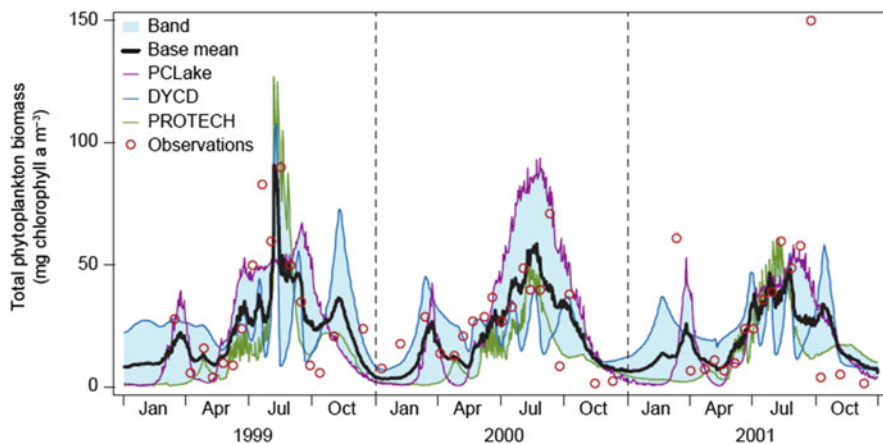
### 16.3.2 Ensemble of Alternative Models for Simulating Phytoplankton Response to Climate Change

Trolle et al. (2014) applied a model ensemble to test the hypothesis that the corresponding mean predictions (derived as the average of daily outputs of the

ensemble members) can provide a better working model compared with any individual process-based model (Gneiting and Raftery 2005). The case study for testing this hypothesis was Lake Engelsholm in Denmark—a shallow eutrophic lake surrounded by a catchment area (15.2 km<sup>2</sup>) that consists mainly of cultivated arable land, forested hills, and scattered dwellings. The ensemble tool comprised three process-based models: PCLake (Janse 1997), PROTECH (Elliott et al. 2010) and DYRESM-CAEDYM (Hamilton and Schladow 1997).

The calibration of PCLake and DYRESM-CAEDYM was conducted independently by adjusting parameters related to intracellular nutrient storage, maximum potential growth rates for phytoplankton and zooplankton grazing rates. PROTECH and DYRESM-CAEDYM were also subject to iterative adjustments of the release rates of nutrients from the bottom sediments, whereas PCLake reflects a dynamic sediment nutrient pool in which the nutrient reflux from the sediments is related dynamically to the biogeochemical processes of the water column (Trolle et al. 2014). In general, Trolle et al. (2014) found that the ensemble mean predictions were superior to any of the individual models used in reproducing both day-by-day and monthly mean total phytoplankton biomass for the entire 1999–2001 study period (Fig. 16.10). The same study further noted that the differences in phytoplankton biomass simulated by the three models (blue shaded zone in Fig. 16.10) were higher when biomass peaks during spring and summer months, whereas their predictions appear to converge with respect to the timing of low biomass, a period also known as the clear-water phase between spring and summer blooms.

Capitalizing upon these predictive discrepancies as well as the conceptual differences of the three models, Trolle et al. (2014) examined climate change scenarios, reflecting a 1.5, 3 and 5 °C warming, and two increased nutrient loading



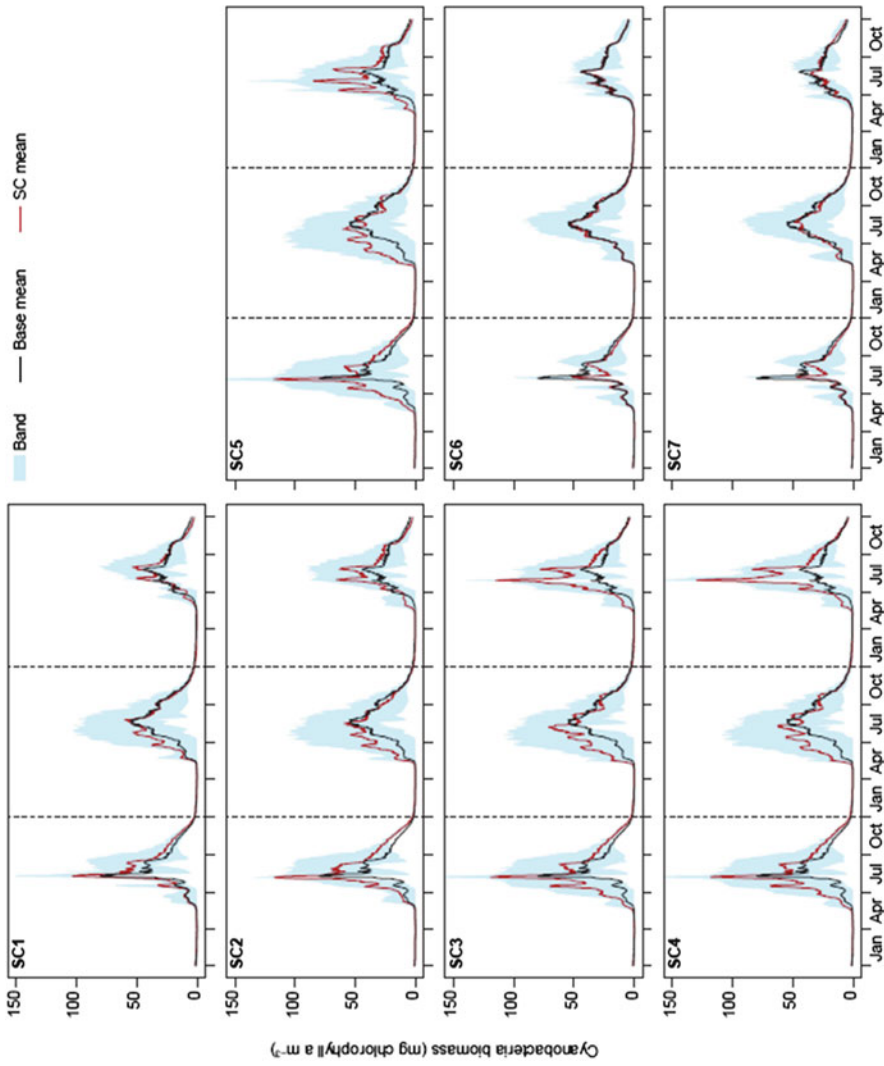
**Fig. 16.10** Calibration (1999–2000) and validation (2001) of PCLake (purple line), DYRESM-CAEDYM or DYCD (blue line), and PROTECH (green line) relative to observed phytoplankton dynamics (red circles) in Lake Engelsholm. The blue shaded “Band” represents the total range (maximum/minimum) of the three models and the thick black “Mean” line represents the ensemble mean of all three models

**Table 16.4** Potential future climate and nutrient load scenarios relative to base scenario (years 1999–2001). These scenarios were used to project the response of Lake Engelsholm with the model ensemble comprising three process-based models

	Scenario details	Daily temperature change relative to base (°C)	Increase in total nitrogen and phosphorus loads relative to base (%)
Scenario 1	Indicative of warming by year 2050	1.5	0
Scenario 2	Indicative of warming by year 2100	3	0
Scenario 3	Indicative of high warming by year 2100	5	0
Scenario 4	Indicative of high warming and increased precipitation by year 2100	5	+5
Scenario 5	Indicative of high warming and highly increased precipitation by year 2100	5	+15
Scenario 6	Nutrient loading increase by 5%	0	+5
Scenario 7	Nutrient loading increase by 15%	0	+15

regimes (Table 16.4). These simulations showed that overall phytoplankton biomass is likely to increase, and cyanobacteria will become a more dominant group of the phytoplankton assemblages with warmer climate (Fig. 16.11). In particular, it was predicted that future climate warming may cause an increase in the average number of days per year when cyanobacteria biomass exceeds the World Health Organization recommended limits, from 8 to 23 days per year, even with conservative scenarios of air temperature increase. In the model simulations, the pattern of cyanobacteria dominance was triggered not only through direct influence of temperature on growth rate, but also indirectly through changes in water column stability and/or nutrient transformation rates (Trolle et al. 2014).

Similar to the Hamilton Harbour study (Ramin et al. 2012), Trolle et al. (2014) used the uncertainty underlying model predictions to pinpoint directions for optimizing the structure of water quality models through process reformulation (e.g., exclusion/inclusion of highly uncertain/missing ecological mechanisms) or refinement of the spatial resolution. For example, the largest uncertainty for the scenarios that combined climate warming and increased external nutrient loads, relative to the scenarios with warming alone, were primarily attributed to: (1) the conceptual differences in the way the three models handle the interplay between nutrients in bottom sediments and the overlying water column, and (2) several other structural differences regarding the critical mechanisms that shape phytoplankton dynamics, such as the explicit representation or not of cyanobacterial nitrogen fixation. Another stark difference was the dramatic oscillations in phytoplankton biomass simulated by DYRESM-CAEDYM in the summer periods relative to the other two



**Fig. 16.11** Response of simulated cyanobacteria biomass to climate change scenarios (SC) in Lake Engelholm. The ensemble mean simulation of cyanobacteria biomass (expressed as chlorophyll  $a$  concentrations) for the seven scenarios (red lines) presented in Table 16.3. Blue shaded band represents the uncertainty range of the three models and black line corresponds to the ensemble mean from the base simulation

models. The latter pattern was associated with the detailed high-frequency hydrodynamics, which can influence phytoplankton dynamics in daily (or even sub-daily) scales and vertical distributions, thereby resulting in greater output variability relative to the simplified physical environment postulated by PCLake and PROTECH. A second plausible explanation for the emergence of these dynamic phytoplankton behaviours could have been the explicit consideration of phytoplankton-zooplankton interactions in DYRESM-CAEDYM, instead of the implicit accommodation through a simple mortality rate on phytoplankton (Trolle et al. 2014).

As with any modeling exercise, an important mechanism for further improving the reliability of strategic forecasting with model ensembles is to test them against observation data that truly reflect future conditions (Refsgaard et al. 2014). While this may seem a key challenge in the context of climate change, the multi-model ensemble approach provides greater confidence and will likely become commonplace methodology in the future, as it enables increased robustness of model projections and scenario uncertainty estimation due to differences in model structures. The only difference between the two case studies presented in this chapter is that the work by Ramin et al. (2012) propagates both within- (initial conditions, parametric error) and among- (structural) model uncertainty through the ecological forecasts used to guide future management and planning.

### ***16.3.3 Ensemble of Complementary Models for Simulating Climate and Land Use Effects on Catchment-Reservoir Systems***

Drinking water reservoirs are typically designed to store surface run-off water from upstream catchments. Therefore, both water quantity and quality of reservoirs are largely determined by impacts of climate and land uses on soils and vegetation in catchments. The concept of external nutrient loadings by Vollenweider (1976) was the first attempt to take these catchment-reservoir relationships explicitly into account by classifying the trophic state of reservoirs depending on phosphorus loadings from the catchment. Meanwhile, process-based catchment models such as SWAT (Arnold et al. 2012) are available that can simulate nutrient loadings at daily time-steps, and lake models such as SALMO (see Chap. 10) that can simulate in-lake nutrient cycling and plankton dynamics in response to external nutrient loadings.

This case study applies the model ensemble SWAT-SALMO to the semi-arid Millbrook catchment-reservoir system in South Australia to demonstrate benefits of simulating spatial-cascading effects of climate and land-use changes in the long-term. The Millbrook catchment covers an area of 361 km<sup>2</sup> and is characterized by multiple land uses including orchards, vineyards and residential areas (see Fig. 16.12), that over time undergo changes driven by demographic and economic

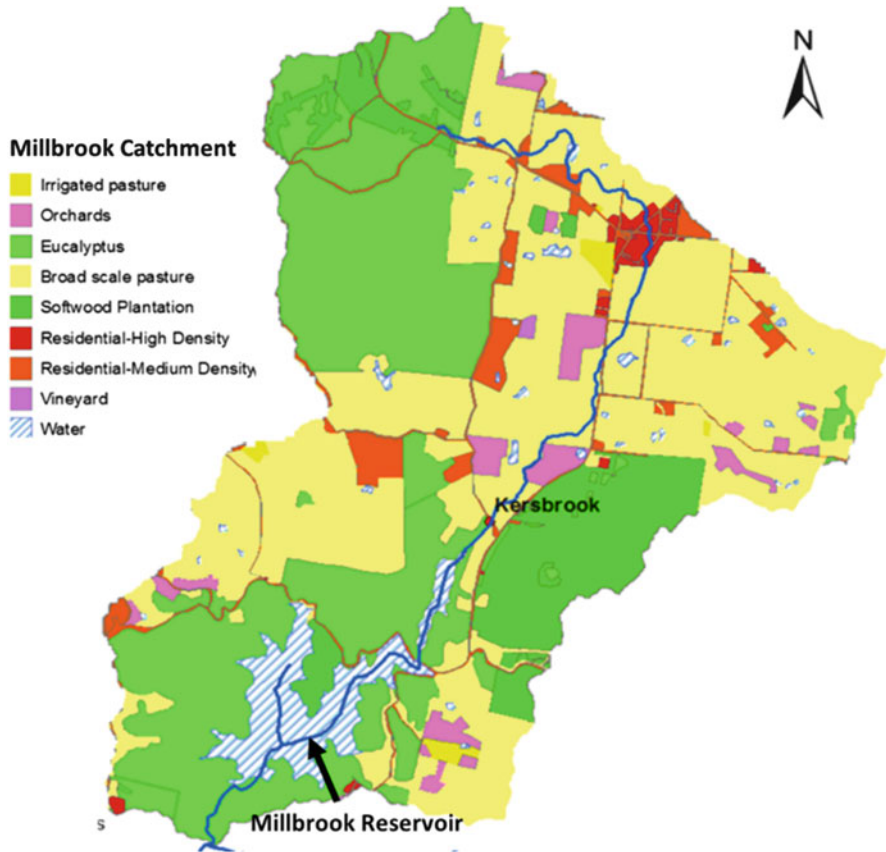


Fig. 16.12 Millbrook catchment-reservoir system (South Australia)

development. The Millbrook reservoir (Fig. 16.12) has a volume of 16,000 ML and a surface area of 176 ha, and contributes approximately 16% of the drinking water supply for Adelaide, the capital of South Australia. The reservoir is equipped with an aerator that is operated during the summer months (December through March) in order to prevent thermal and oxygenic stratification of the water body.

The model ensemble SWAT-SALMO (Fig. 16.13) has been applied as follows:

### Step 1

Calibration and validation of the model SWAT for the Millbrook catchment from 2008 to 2012 based on the catchment-specific digital elevation model (DEM), soil and land-use maps, and meteorological data, as well as stream flow and nutrient data. Figure 16.14 displays validation results for the simulated flow and concentrations of nitrate and phosphate at daily time steps that entered the Millbrook reservoir from 2008 to 2012. It shows that despite seasonal overestimation of flow and nitrate, the overall results satisfactorily matched observed data as



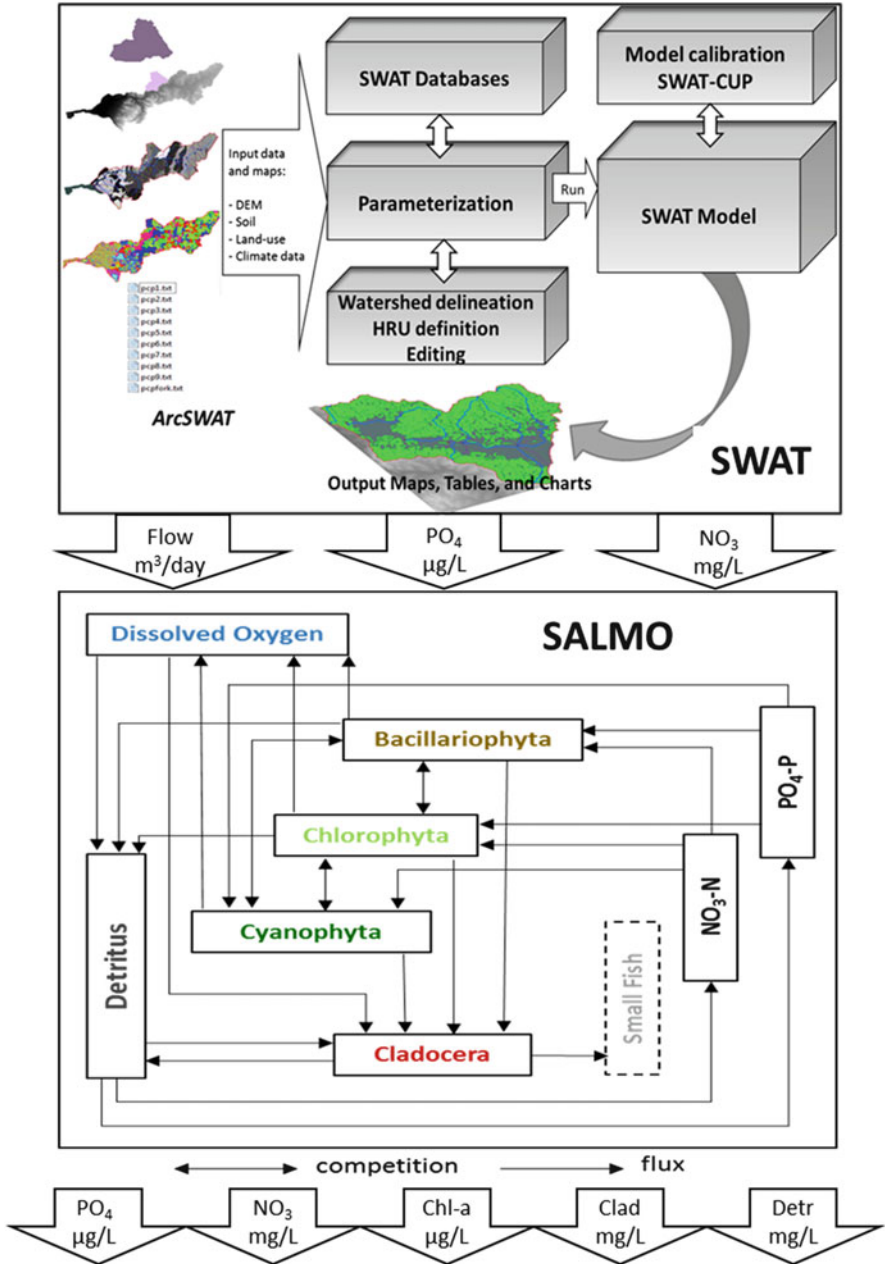
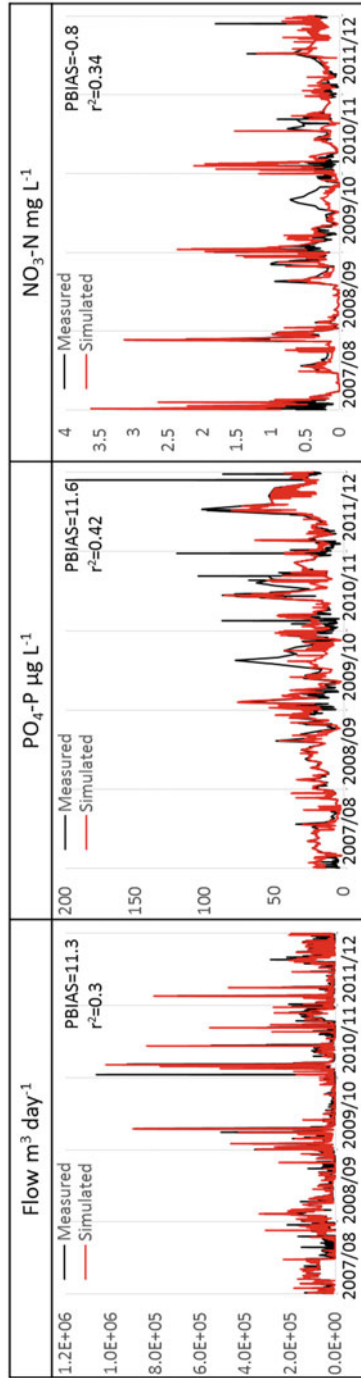


Fig. 16.13 Model ensemble SWAT-SALMO



**Fig. 16.14** SWAT simulation of flow,  $PO_4-P$ - and  $NO_3-N$ - concentrations in the Millbrook catchment entering the Millbrook reservoir from 2008 to 2012

indicated by the percentage bias (PBIAS) values that were well above the criteria of  $\pm 70\%$  as recommended by Moriasi et al. (2007), and justified the use of the SWAT simulated outputs as inputs for SALMO.

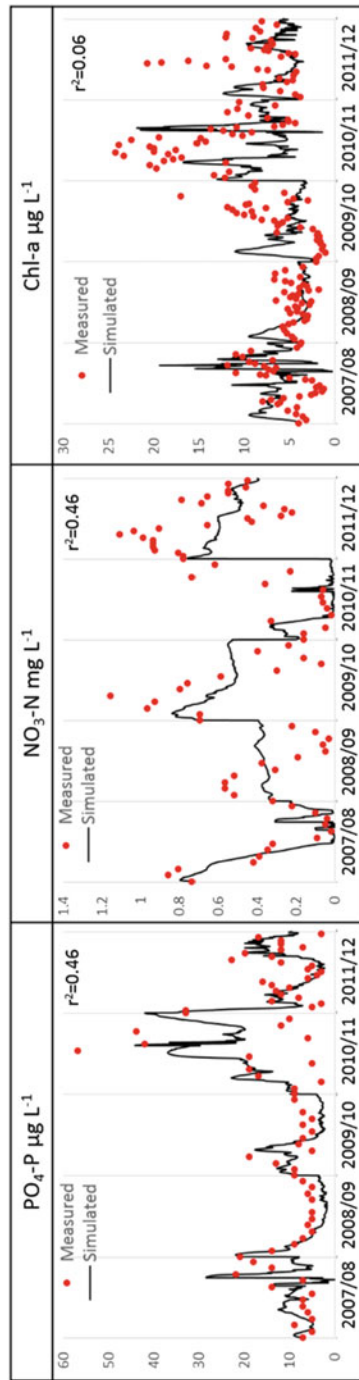
### Step 2

Calibration and validation of the model SALMO for the artificially de-stratified Millbrook reservoir from 2008 to 2012 based on: daily phosphate and nitrate loadings provided by SWAT, daily volumes, mixing depths, solar radiation and water temperature. Figure 16.15 displays validation results for simulated phosphate, nitrate and Chl-a concentrations that match seasonal trends of observed data but differ year by year. For more details about SALMO see Chap. 10.

### Step 3

Using SWAT to simulate flow and nutrient concentrations that enter the Millbrook reservoir based on following scenarios:

- (1) Restricting import of external river water to the Millbrook catchment by 50%.  
This scenario has been designed to test the hypothesis that a 50% reduced import of external river water may lower nutrient loads to the Millbrook reservoir, and may mitigate eutrophication effects from future land and climate changes.
- (2) Replacing 50% of pasture areas in the Millbrook catchment by residential areas.  
This scenario takes into account likely effects of on-going population growth assuming that in the forthcoming 30 years up to 50% of current pasture land will be converted to residential areas, and may impact water quality of the Millbrook reservoir.
- (3) Imposing effects of global warming on the Millbrook Catchment-Reservoir system as projected for the upcoming 30 years by the global climate models (GCM) of the 5th IPCC Report (IPCC5 2014).  
This scenario utilises daily rainfall and air temperature data provided by GIWR (2015) that were forecasted and calibrated for different regions of South Australia until 2100 by means of global climate models (GCM) from the 5th IPCC Report (IPCC5 2014). The GCM produced 100 stochastic replicates of climate data until 2100 both for “low” emission  $4.5 \text{ W/m}^2$  and “high” emission  $8.5 \text{ W/m}^2$ . However, only one replicate that corresponded to the median of projected total precipitation for the period between 2006 and 2100 was selected for scenario (3) utilising data for the “high” emission case represented as RCP 8.5.
- (4) Combining scenarios (1) and (3)
- (5) Combining scenarios (2) and (3)
- (6) Combining scenarios (2) and (3) with thermal stratification of the reservoir.  
This scenario investigates the impact of prospective land use changes and global warming on water quality if the reservoir is thermally stratified during summer.



**Fig. 16.15** SALMO simulation of NO<sub>3</sub><sup>-</sup>, PO<sub>4</sub><sup>-</sup> and chl-a concentrations in the Millbrook reservoir from 2008 to 2012 driven by nutrient loadings simulated by SWAT (Fig. 16.15)

#### Step 4

Using SALMO to simulate phosphate, nitrate and chlorophyll-a concentrations in the Millbrook reservoir based on the SWAT outputs from scenarios (1) to (6).

Results of scenarios (1) to (6) have been illustrated in Figs. 16.16 and 16.17 and summarised in Table 16.5 for the period from July 2008 to June 2009 that experienced dry conditions (so-called ‘dry year’) and the period from July 2010 to June 2012 that experienced wet conditions (so-called ‘wet year’). Figure 16.18 illustrates differences in air and water temperatures of these 2 years before and after global warming simulations.

Scenario (1) confirms observations that the external river water carries higher phosphate concentrations than the natural catchment water. Therefore, a 50% reduced import of river water is expected to lower phosphate loads to the reservoir, and consequently phosphate and chlorophyll-a concentrations within the reservoir during the ‘wet year’, but may have only minor effects during the ‘dry year’.

Scenario (2) suggests a slightly increased flow from the catchment driven by extended impervious residential areas. As a result, it enriches phosphate and nitrate concentrations in the reservoir leading to slightly higher chlorophyll-a during the ‘wet year’ only.

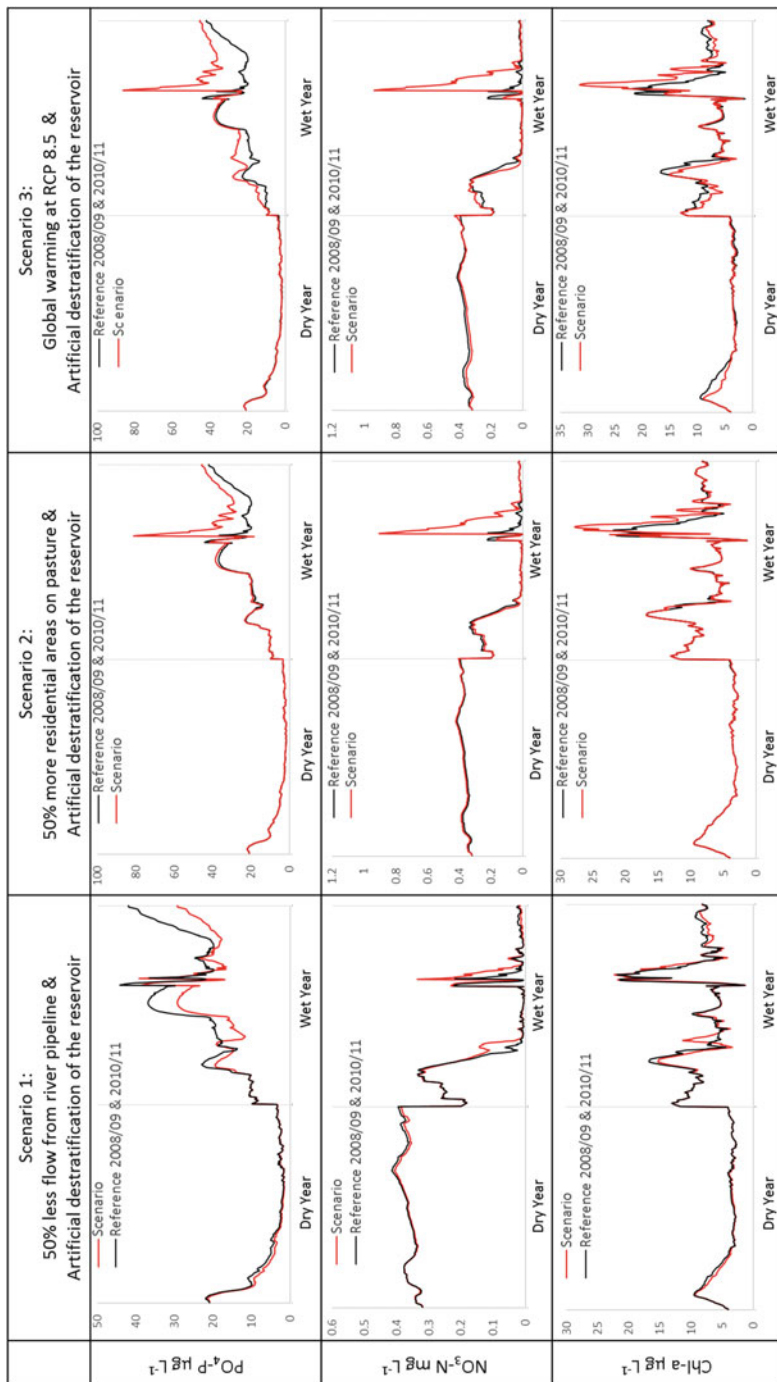
The prospective global warming over the next 30 years as simulated by scenario (3) is likely to affect flow from the catchment by less and more sporadic rainfall. However, resulting flow may carry higher phosphate concentrations driven by more intense microbial and photochemical decomposition of organic matter under the influence of higher air temperature and UVB light. These effects together with reservoir water that is on average warmer by 1.98 °C (see Fig. 16.18) will stimulate algal growth particularly during the ‘wet year’ as reflected by an increased chlorophyll-a concentration of 0.2% during summer.

Scenario (4) suggests that a stimulation of algal growth by global warming as forecasted by scenario (3) can partially be mitigated by lowering the phosphate load from the Millbrook catchment by a 50% reduced import of external river water.

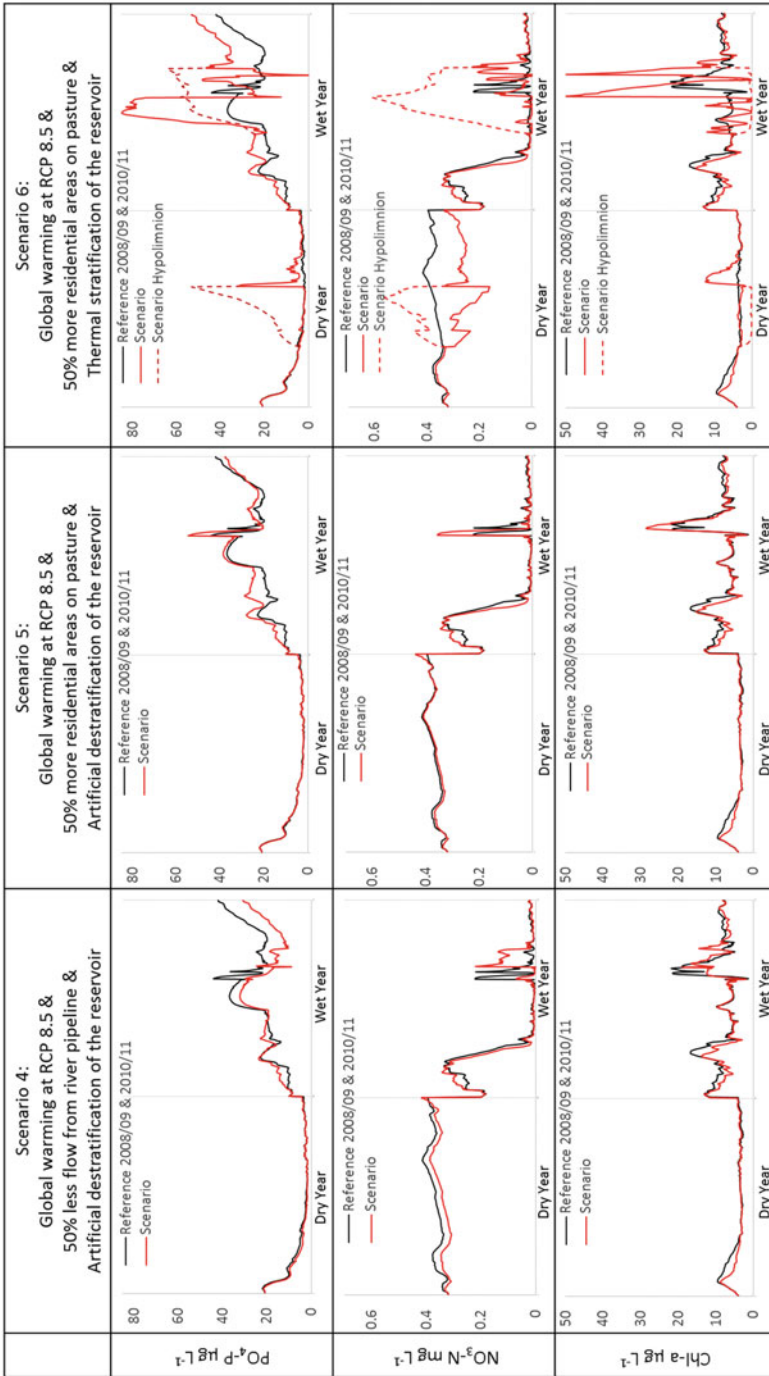
Scenario (5) demonstrates that combined effects of extending residential areas and global warming as anticipated for the next 30 years will increase both nitrate and phosphate loadings from the catchment resulting in higher nutrient and chlorophyll-a concentrations in the reservoir and posing the risk of cyanobacteria blooms.

Since the Millbrook reservoir had been artificially destratified during the period of this case study with the aim to control not only physical-chemical processes such as internal phosphate loads from anaerobic sediments but also algal growth by limiting light contact and buoyancy (e.g. Cooke et al. 2005), the interpretation of these five scenarios must take into account successful mitigation effects by this control measure. These mitigation effects become also evident by unusual seasonal dynamics of algal biomass simulated by chlorophyll-a.

Scenario (6) indicates that the reservoir would face severe eutrophication effects from prospective land use and climate changes if its water body would be thermally stratified. This is reflected in particular by estimated PO<sub>4</sub>-P concentrations that would increase by 21% in dry years and 49% in wet years, and chl-a concentrations



**Fig. 16.16** SWAT-SALMO simulation results for the Millbrook reservoir of the scenarios: 50% less flow from river pipeline (left column), 50% more residential areas on pasture (middle column), and Global warming at RCP8.5 (right column)

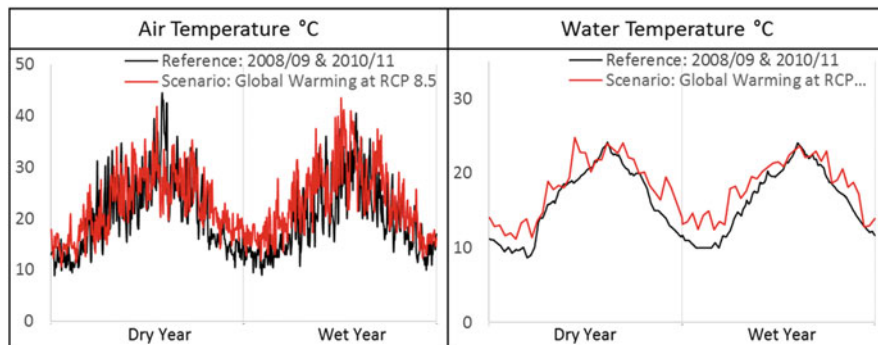


**Fig. 16.17** SWAT-SALMO simulation results for the Millbrook reservoir of the scenarios: 50% less flow & global warming (*left column*), 50% more residential areas & global warming (*middle column*), and 50% more residential areas & global warming & thermal stratification of the reservoir (*right column*)

**Table 16.5** Summary of the simulation results for the six scenarios

Scenarios	Change in flow		Change in inflow PO <sub>4</sub> -P concentrations		Change in inflow NO <sub>3</sub> -N concentrations		Change in in-lake PO <sub>4</sub> -P concentrations (summer months)		Change in in-lake NO <sub>3</sub> -N concentrations (summer months)		Change in in-lake Chl-a concentrations (summer months)	
	Dry year	Wet year	Dry year	Wet year	Dry year	Wet year	Dry year	Wet year	Dry year	Wet year	Dry year	Wet year
(1) 50% less flow from river pipeline	-33.3%	-19.5%	-5.1%	-10.1%	7.7%	-5.2%	0%	-0.17%	0.02%	0.14%	-0.04%	-1.22%
(2) 50% more residential areas on pasture	1.1%	3.2%	-1%	-9.3%	2.6%	-4.7%	-0.1%	0.19%	-0.01%	0.45%	0.02%	9.52%
(3) Global warming at RCP 8.5	-1.8%	-0.5%	4%	8.2%	-11.2%	2.8%	0.01%	0.33%	-0.03%	0.47%	0.02%	13.59%
(4) Global warming at RCP 8.5 & 50% less flow from river pipeline	-33.7%	-17.2%	1.5%	-3.9%	-10.1%	-5.3%	-0.1%	-0.15%	-0.06%	0.03%	-0.01%	-11.35%
(5) Global warming at RCP 8.5 & 50% more residential areas on pasture	1.5%	4.2%	2%	0.2%	-15.2%	1.4%	0.01%	0.12%	-0.01%	0.05%	0.02%	8.06%
(6) Global warming at RCP 8.5 & 50% more residential areas on pasture & thermal stratification of reservoir	1.5%	4.2%	2%	0.2%	-15.2%	1.4%	21.16%	49.29%	-21.9%	8.28%	6.76%	52.43%





**Fig. 16.18** Air temperature (*left column*), and water temperature (*right column*) of the ‘dry year’ and the ‘wet year’ before and after global warming simulations

that would increase by 6.8% in dry years and 52.4% in wet years. The results of scenario (6) clearly approve the precautionary measure by SA Water to prevent thermal stratification of the reservoir by artificial mixing since the 1990s as prerequisite for sustainable water supply in future.

Overall, this case study has demonstrated that complex scenarios such as assessing impacts of human population growth and global warming on eutrophication of lakes by far exceed the scope of a single lake model. To make such scenario analyses relevant and credible: (1) ensembles of complementary models are required that reflect both key processes in catchments as well as those in reservoirs; and (2) validation of seasonal and inter-annual nutrient and phytoplankton dynamics is required to make forecasted eutrophication effects transparent and justifiable. The full study has been documented in Nguyen et al. (2017).

## 16.4 Concluding Remarks

The use of model ensembles is a promising strategy to improve contemporary ecological forecasting. Ensembles of inferential models can overcome the limitation of a single model that is lacking information transmitting processes, and enable information to cascade between complementary models as required for the simulation of nutrient cycles and community dynamics (e.g. Recknagel et al. 2014, 2017).

Ensembles of alternative process-based models provide not only a framework to improve forecasting validity, but also to compare alternative ecological structures, to challenge existing ecosystem conceptualizations, and to integrate across different (and often conflicting) paradigms (Ramin et al. 2012; Trolle et al. 2014). As previously shown, the discrepancy between the projections of two distinct ecosystem characterizations offers an excellent opportunity to formulate testable hypotheses and identify potentially critical ecological processes/mechanisms under

significantly different external conditions (e.g., climate warming, nutrient loading, invasive species).

Ensembles of complementary process-based models extend the scope of a single model in order to realistically simulate exchange processes between highly-interrelated “open” ecosystems such as catchments and lakes, which are vital to determine their response to such complex scenarios as global warming.

To further improve credibility and acceptance of strategic forecasting, future research should focus on the refinement of the weighting schemes and other performance standards to impartially synthesize predictions of different models (Wilks 2002; Lindström et al. 2015). Specifically, some outstanding challenges involve: (1) the development of ground rules for the features of the calibration and validation domain in order to effectively weight the individual members of model ensembles on the basis of their performance; (2) the inclusion of penalties for model complexity that will allow building ensemble forecasts upon parsimonious models; and (3) performance assessment that does not exclusively consider model endpoints but also examines the plausibility of the underlying ecosystem structures, i.e., biological rates, ecological processes or other derived mass fluxes.

**Acknowledgements** George Arhonditsis wishes to acknowledge the continuous support of his work on model uncertainty analysis from the National Sciences and Engineering Research Council of Canada (Discovery Grants). Friedrich Recknagel expresses his gratitude to the Department of Biological Sciences, Pusan National University (South Korea) for making available limnological data of River Nakdong, and to SA Water (Australia) for providing hydrological and limnological data of the Millbrook catchment-reservoir system.

## References

- Arhonditsis GB, Brett MT (2004) Evaluation of the current state of mechanistic aquatic biogeochemical modeling. *Mar Ecol Prog Ser* 271:13–26
- Arhonditsis GB, Brett MT (2005) Eutrophication model for Lake Washington (USA). Part I. Model description and sensitivity analysis. *Ecol Model* 187:140–178
- Arhonditsis GB, Qian SS, Stow CA et al (2007) Eutrophication risk assessment using Bayesian calibration of process-based models: application to a mesotrophic lake. *Ecol Model* 208:215–229
- Arhonditsis GB, Papantou D, Zhang WT et al (2008a) Bayesian calibration of mechanistic aquatic biogeochemical models and benefits for environmental management. *J Mar Syst* 73:8–30
- Arhonditsis GB, Perhar G, Zhang WT et al (2008b) Addressing equifinality and uncertainty in eutrophication models. *Water Resour Res* 44:W01420
- Arnold JG, Moriasi DN, Gassman PW et al (2012) SWAT: Model use, calibration and validation. *Trans ASABE (Am Soc Agric Biol Eng)* 55:1491–1508
- Azcue JM, Zeman AJ, Mudroch A et al (1998) Assessment of sediment Harbour, Canada. *Water Sci Technol* 37:323–329
- Bao L, Gneiting T, Grimit EP et al (2010) Bias correction and Bayesian model averaging for ensemble forecasts of surface wind direction. *Mon Weather Rev* 138:1811–1821
- Cao H, Recknagel F, Orr PT (2014) Parameter optimisation algorithms for evolving rule models applied to freshwater ecosystem. *IEEE Trans Evol Comp* 18:793–806

- Cooke GD, Welch EB, Peterson S et al (2005) Restoration and management of freshwater lakes, 3rd edn. CRC Press, New York
- Elliott JA, Irish AE, Reynolds CS (2010) Modeling phytoplankton dynamics in fresh waters: affirmation of the PROTECH approach to simulation. *Freshw Res* 3:75–96
- Franks PJS (1995) Coupled physical-biological models in oceanography. *Rev Geophys* 33:1177–1187
- Gelman A, Carlin JB, Stern HS et al (2013) Bayesian data analysis, 3rd edn. Chapman and Hall, New York
- Gilks WR, Richardson S, Spiegelhalter DJ (1998) Markov Chain Monte Carlo in practice. Chapman & Hall/CRC, New York
- GIWR (2015) SA climate ready data for South Australia – a user guide, Goyder Institute for Water Research Occasional Paper No. 15/1, Adelaide, South Australia
- Gneiting T, Raftery AE (2005) Weather forecasting with ensemble methods. *Science* 310:248–249
- Gudimov A, Stremilov S, Ramin M et al (2010) Eutrophication risk assessment in Hamilton Harbour: system analysis and evaluation of nutrient loading scenarios. *J Great Lakes Res* 36:520–539
- Gudimov A, Ramin M, Labencki T et al (2011) Predicting the response of Hamilton Harbour to the nutrient loading reductions, a modeling analysis of the “ecological unknowns”. *J Great Lakes Res* 37:494–506
- Gudimov A, McCulloch J, Chen J et al (2016) Modeling the interplay between deep water oxygen dynamics and sediment diagenesis in a hard-water mesotrophic lake. *Ecol Inform* 31:59–69
- Ha K, Cho E-A, Kim H-W et al (1999) *Microcystis* bloom formation in the lower Nakdong River, South Korea: importance of hydrodynamics and nutrient loading. *Mar Freshwater Res* 50:89–94
- Ha K, Jang M-H, Joo G-J (2003) Winter *Stephanodiscus* bloom development in the Nakdong River regulated by an estuary dam and tributaries. *Hydrobiologia* 506/509:221–227
- Hamilton DP, Schladow SG (1997) Prediction of water quality in lakes and reservoirs, part 1: model description. *Ecol Model* 96:91–110
- Hoeting JA, Madigan D, Raftery AE et al (1999) Bayesian model averaging: a tutorial. *Stat Sci* 14:382–417
- Hong D-G, Jeong K-S, Kim D-K et al (2014) Remedial strategy of algal proliferation in a regulated river system by integrated hydrological control: an evolutionary modeling framework. *Mar Freshw Res* 65:379–395
- Hongping P, Jianyi M (2002) Study on the algal dynamic model for West Lake, Hangzhou. *Ecol Model* 148:67–77
- IPCC5 (2014) Climate change 2014: impacts, adaptation and vulnerability. Working group II contribution to the fifth assessment report of the intergovernmental panel on climate change. Cambridge University Press, Cambridge
- Janse JH (1997) A model of nutrient dynamics in shallow lakes in relation to multiple stable states. *Hydrobiologia* 342–343:1–8
- Jaynes ET (1994) Probability theory: the logic of science. Cambridge University Press, New York
- Jeong K-S, Kim D-K, Joo G-J (2007) Delayed influence of dam storage and discharge on the determination of seasonal proliferations of *Microcystis aeruginosa* and *Stephanodiscus hantzschii* in a regulated river system of the lower Nakdong River (South Korea). *Water Res* 41:1269–1279
- Law T, Zhang W, Zhao J et al (2009) Structural changes in lake functioning induced from nutrient loading and climate variability. *Ecol Model* 220:979–997
- Lindström T, Tildesley M, Webb C (2015) A Bayesian Ensemble Approach for Epidemiological Projections. *PLoS Comput Biol* 11(4):e1004187. doi:10.1371/journal.pcbi.1004187
- McDonald CP, Bennington V, Urban NR et al (2012) 1-D test-bed calibration of a 3-D Lake Superior biogeochemical model. *Ecol Model* 225:115–126
- Moriasi DN, Arnold J, Van Liew M et al (2007) Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. *Trans Am Soc Agric Eng* 50:885–900

- Neuman WL (2003) Maximum likelihood Bayesian averaging of uncertain model predictions. *Stoch Environ Res Risk A* 17:291–305
- Nguyen HH, Recknagel F, Meyer W et al (2017) Modelling the impacts of altered management practices, land use and climate changes on the water quality of the Millbrook catchment-reservoir system in South Australia. *J Environ Manage* [http://ac.els-cdn.com/S0301479717306801/1-s2.0-S0301479717306801-main.pdf?\\_tid=1c0d5a64-8b9e-11e7-b5ee-00000aab0f02&acdnat=1503889865\\_c53ba4804a5f818671d66a8452a07c47](http://ac.els-cdn.com/S0301479717306801/1-s2.0-S0301479717306801-main.pdf?_tid=1c0d5a64-8b9e-11e7-b5ee-00000aab0f02&acdnat=1503889865_c53ba4804a5f818671d66a8452a07c47)
- Park RA et al (1974) A generalised model for simulating lake ecosystems. *Simulation* 23:33–50
- Raftery AE, Gneiting T, Balabdaoui F et al (2005) Using Bayesian model averaging to calibrate forecast ensembles. *Mon Weather Rev* 133:1155–1174
- Ramin M, Stremilov S, Labencki T et al (2011) Integration of mathematical modeling and Bayesian inference for setting water quality criteria in Hamilton Harbour, Ontario, Canada. *Environ Model Softw* 26:337–353
- Ramin M, Labencki T, Boyd D et al (2012) A Bayesian synthesis of predictions from different models for setting water quality criteria. *Ecol Model* 242:127–145
- Recknagel F, van Ginkel C, Cao H et al (2008a) Generic limnological models on the touchstone: testing the lake simulation library SALMO-OO and the rule-based *Microcystis* agent for warm-monomictic hypertrophic lakes in South Africa. *Ecol Model* 215:144–158
- Recknagel F, Cetin LT, Zhang B (2008b) Process-based simulation library SALMO-OO for lake ecosystems. Part 1: object-oriented implementation and validation. *Ecol Inf* 3:170–180
- Recknagel F, Ostrovsky I, Cao H (2014) Model ensemble for the simulation of plankton community dynamics of Lake Kinneret (Israel) induced from in situ predictor variables by evolutionary computation. *Environ Model Softw* 61:380–392
- Recknagel F, Kim D-K, Joo G-J et al (2017) Response of *Microcystis* and *Stephanodiscus* to alternative flow regimes of the regulated River Nakdong (South Korea) quantified by model ensembles based on the hybrid evolutionary algorithm HEA. *River Res Appl*. <http://onlinelibrary.wiley.com/doi/10.1002/rra.3141/full>
- Refsgaard JC, Madsen H, Andréassian V et al (2014) A framework for testing the ability of models to project climate change and its impacts. *Clim Change* 122:271–282
- Reynolds CS (2006) *The ecology of phytoplankton*. Cambridge University Press, Cambridge
- Roberts RD, Zohary T (1987) Temperature effects on photosynthetic capacity, respiration, and growth rates of bloom-forming cyanobacteria. *NZ J Mar Freshw Res* 21:391–399
- Slougher JM, Raftery AE, Gneiting T et al (2007) Probabilistic quantitative precipitation forecasting using Bayesian model averaging. *Mon Weather Rev* 135:3209–3220
- Slougher JM, Gneiting T, Raftery AE (2010) Probabilistic wind speed forecasting using ensembles and Bayesian model averaging. *J Am Stat Assoc* 105:25–35
- Trolle D, Elliott JA, Mooij WM et al (2014) Advancing projections of phytoplankton responses to climate change through ensemble modeling. *Environ Model Softw* 61:371–379
- Vollenweider R (1976) Advances in defining critical loading levels for phosphorus in lake eutrophication. *Mem Inst Ital Idrobiol* 33:53–86
- Wilks DS (2002) Smoothing forecast ensembles with fitted probability distributions. *Q J Roy Meteorol Soc* 128:2821–2836

**Part V**  
**Case Studies**

# Chapter 17

## Biodiversity Informatics

Cynthia S. Parr and Anne E. Thessen

**Abstract** Biodiversity informatics, the application of informatics techniques to biodiversity data, is rooted in physical objects and nomenclatural codes. Through two user stories, one from wildlife conservation and another from agriculture, we demonstrate the importance and process of biodiversity informatics. We discuss the importance and integration of taxonomic names, identification tools, species distributions, phylogenetic trees, traits, associations, the literature, ontologies, controlled vocabularies, standards, and genomics. Despite the plethora of resources, a seamless, biodiversity question and answer engine is still out of reach. The largest impediment to our user stories is the lack of cross-disciplinary infrastructure and the digitized and standardized data to support services. Satisfying our user stories will require additional investment in infrastructure and data that will be a challenge to manage and sustain. This chapter discusses the basic biodiversity informatics concepts that are at the heart of our user stories, and will be the basis of the user stories of the future as society rushes to cope with global environmental change.

### 17.1 Introduction

Biodiversity, the variety of life (Wilson 1999), is more than the charismatic subject of your favorite nature documentaries. The wealth of genetic diversity, the millions of named and unnamed species, and their roles in every ecosystem are vital to sustaining human life on this planet, and our planet would be unrecognizable without them. For this reason, 196 countries have ratified the 1992 Convention on Biological Diversity (CBD 2016), a legally binding agreement that commits them to reversing the loss of these precious natural resources. In this chapter, we present biodiversity informatics as a discipline with deep roots and a promising future.

---

C.S. Parr (✉)

National Agricultural Library, Beltsville, MD, USA

e-mail: [cynthia.parr@ars.usda.gov](mailto:cynthia.parr@ars.usda.gov)

A.E. Thessen

Ronin Institute for Independent Scholarship, Monclair, NJ, USA

The Data Detektiv, Waltham, MA, USA

e-mail: [annethessen@gmail.com](mailto:annethessen@gmail.com)

Through the lens of two specific case studies, we aim to leave the reader with a solid understanding of the key areas of biodiversity informatics as well as some concrete examples of projects, tools, and standards that they may use. We aim to identify areas where more work is needed so that tomorrow's biodiversity informaticians will be inspired to push information technology forward to support science and policy to find real solutions to society's problems.

### ***17.1.1 What Is Biodiversity Informatics***

The terms “Biodiversity” and “informatics” were both coined in the latter half of the twentieth century (Wilson 1988; Widrow et al. 2005) but does that mean that the discipline is new? Certainly, emerging computer technology of the last few decades has enabled collaboratively-built, web-accessible databases and services about biological organisms (Bisby 2000; Heidorn 2011) that have transformed how biodiversity information is shared. Yet it can be argued that the development and sharing of information about biological diversity has very deep roots. Cave paintings of animals and oral traditions about medicinal plants are early examples where humans captured their knowledge of the living world with the intent to communicate that information to present and future audiences. Systematic ways of portraying and naming and organizing and describing organisms are now being translated into the digital world, but the essential questions to be addressed with the information remain the same as in ancient times: what organism is this, where does it live, how does it relate to other organisms, how is it important for us, how do we protect it or protect against it if that's necessary, and what can we learn from it.

Several key concepts are central to biodiversity informatics. These concepts have some parallels in other informatics disciplines but are worth making explicit in the biodiversity context. First, much of biodiversity informatics rests on physical objects that ground the digital information: specimens of organisms. Second, some of these specimens are vouchers for the nomenclature that is used when referring to organisms—as required by the codes that govern the names of organisms. We will talk more about names below, but they are vital to biodiversity informatics, and also in linking information about biodiversity to the other kinds of information about the world. A final key concept is the taxonomic impediment. A recent estimate is that 6.5 million or 86% of all species remain unnamed and undescribed (Mora et al. 2011). The number of experts currently engaged in what is called “primary” or “alpha”-taxonomy may be too small to finish the job, or at least not with current tools; although some disagree (see Costello et al. 2012). Even if ecologists who work with biodiversity use proxies [e.g. representatives, functional groups, guilds, or higher taxonomic groups as in the Madingly model (Purves et al. 2013)], knowledge of the many individual organisms that make up those groups remains essential to a robust ecological understanding of the biosphere.

### 17.1.2 *Our User Story Approach*

In order to demonstrate the importance and process of biodiversity informatics, we will explore user stories involving the impact of climate change on organisms. These user stories were previously described from the perspective of using ontologies to link phenotypes and environments (Thessen et al. 2016). In this chapter, we will describe the existing biodiversity informatics infrastructure these user stories require while highlighting examples the reader can explore further.

Coping with climate change is anticipated to be one of the most challenging aspects of applied biodiversity science, affecting both wildlife conservation and agriculture. Managers want to know “*What species or crop varieties are projected to do well in my location over the next century?*” Addressing this question requires data about species observations, traits, phylogenetics, and genomes. It also requires data infrastructure for curation, management, discovery for digital records, physical specimens, and nomenclatural history. Following are the stories of two users, each of whom is trying to solve their problems in wildlife conservation and plant breeding (modified after Thessen et al. 2016).

#### **User Story 1: Coping with Climate Change in Wildlife Conservation**

Lupita is a park ranger who manages a coastal wildlife sanctuary. Some of the species in her sanctuary are listed as threatened by the IUCN. According to the latest climate change projections, her sanctuary is going to be hotter and wetter in 50 years. She has limited resources to maintain the biodiversity in her sanctuary for the long term. She needs to identify which species might be at risk under the projected future climate regime and consider her options for mitigating that risk.

#### **User Story 2: Coping with Climate Change in Agriculture**

Steve is a scientist working for a seed company. He wants to develop crop hybrids that perform well in the drier, warmer climates predicted for the next 30 years in the region of the country that he serves. He knows that farmers will want to plant crops that are drought tolerant and have high, stable yields. These crops must also be suitable for local soil conditions and sometimes rapidly changing factors such as emerging diseases, invasive pests, and threats to pollinators. He needs to identify promising species and varieties so he can include them in his breeding program.

## 17.2 Meeting the Needs of Biodiversity User Stories

Our user stories require similar types of data, tools, and services. However, the gaps in the existing infrastructure are somewhat different. Below, we discuss the specific needs of each user story and the existing infrastructure in place to fill those needs. For easy reference, standards and ontologies are summarized in Table 17.1. In “Next Steps” (Sect. 17.3) we highlight the gaps as opportunities for future research and technology development.



**Table 17.1** Selected data and metadata standards and ontologies used in biodiversity informatics

Standard	Used Primarily to Describe	References
Darwin Core	Species occurrences	Wieczorek et al. (2012)
Audubon Core	Multimedia	Morris et al. (2013)
Global Genome Biodiversity Network	Biological samples	Droege et al. (2016)
Gene Ontology (GO)	Cell structure and function, genes, proteins, cellular processes	Ashburner et al. (2000)
Relations Ontology (RO)	Organism interactions	RO Project (2016)
Biological Collections Ontology (BCO)	Biological sampling concepts and their relations	Walls et al. (2014)
Phenotype and Trait Ontology	Trait qualities	OBO Technical WG (2016)

### 17.2.1 Taxonomic Names

Observations about organisms have been linked to an organism name for the past several centuries. Most of the observations known to science are attached to scientific, Linnaean names, but vernacular names and, more recently, genetic barcodes have been used. This “taxon identifier” is an important point of integration across data sets and is a common search term used to discover data. Unfortunately, names make poor identifiers, mostly because they are not unique or persistent and are often used ambiguously (Remsen 2016). Problems of several names for one species or one name for multiple species are very common. For example, if Lupita has the common muskrat in her sanctuary, searching for data using *Ondatra zibethicus* will miss information using *Castor zibethicus*, a synonym. How can Lupita and Steve find out about all the synonyms and homonyms they need to find and correctly aggregate existing data to get as complete and accurate a picture as possible of their taxa? To find the answer, we need to understand the nature of taxonomic names.

When taxonomists think they’ve discovered a new species, they publish a description of that species and name it, sometimes depositing a specimen of the new taxon in a museum or herbarium. Species are named and described according to a collection of rules called the codes of nomenclature. There are five: one for animals (Ride et al. 1999), one for plants, fungi, and algae (McNeill et al. 2012), one for cultivated plants (Brickell et al. 2016), one for bacteria (Lapage et al. 1992), and one for viruses (King 2011). This may seem overly complicated, but, these rules attempt to: (1) prevent taxonomists from applying the same name to two different species; (2) mandate that original descriptions are published in such a way that they are widely available; and (3) ground taxonomic descriptions in physical reality by requiring specimens. The rules aren’t perfect. Homonyms are more common than we would like, literature can be difficult to obtain, and specimens can go missing. As unstable as taxonomic names are, they would be much worse

without the rules of nomenclature. The rules of nomenclature have been around for centuries; thus, taxonomic descriptions are very reliant on the printed page and physical objects. Recently, the taxonomic process is adopting more technology through the use of ZooBank (Pyle and Michel 2008), part of the larger Global Names Architecture project (Patterson et al. 2010) and a web-based registry, to make nomenclatural acts machine-readable, instead of solely existing on paper.

There are an estimated 22 million names and 2 million described taxa (Chapman 2009; Patterson 2014). For Lupita and Steve, correctly applying these names can be a daunting task because the rules of nomenclature require that the entire history of a name and specimen be respected. Fortunately, there are tools and services that can help. Nomenclatural authorities and aggregators act as stewards of names, sometimes only for specific taxonomic groups (Table 17.2). They typically have some degree of community buy-in from taxonomists practicing in that area and are accepted as an authority. Some authorities take responsibility for maintaining a list of current, accepted names and their taxonomic synonyms, either by resource-intensive manual curation (e.g., IPNI 2012), curation by committee (e.g., Catalogue of Life, Ruggiero et al. 2015, COL 2016), or an aggregator-enabled, crowd-sourcing approach (e.g., Encyclopedia of Life (EOL), Parr et al. 2014). Only some authorities include vernacular names, surrogates (like strain numbers), or misspellings in their efforts (e.g., EOL, Parr et al. 2014, EOL 2016; uBio, Leary et al. 2007; Global Names Index, Patterson et al. 2010, GNI 2016). The combination of any name string into bundles of synonyms referring to the same taxon are called reconciliation groups (Patterson et al. 2010). Building reconciliation groups can be partially automated using algorithms for fuzzy matching and parsing (see software description in Patterson et al. 2016), but significant manual work is still needed to capture synonym information that resides only in published work and to

**Table 17.2** List of nomenclatural authorities and name aggregators

Authority	Taxonomic Focus	Reference(s)
Catalogue of Life	Lifewide	COL (2016), Roskov et al. (2016)
Encyclopedia of Life	Lifewide	EOL (2016), Parr et al. (2014)
uBio	Lifewide	Leary et al. (2007)
Global Names Index	Lifewide	GNI (2016), Patterson et al. (2010)
Integrated Taxonomic Information System	Lifewide	ITIS (2016)
Taxonomy Database National Center for Biotechnology Information	Lifewide	Federhen (2012), NCBI (2016)
World Register of Marine Species	Lifewide—only marine	WoRMS Editorial Board (2016)
The Interim Register of Marine and Nonmarine Genera	Lifewide to Genus	Rees (2016)

integrate names across databases (Patterson 2014; Patterson et al. 2016). The human curation of names and classifications is a major bottleneck.

Fortunately for Lupita and Steve, valuable services can be built on top of curated classifications and reconciliation groups (e.g., Boyle et al. 2013). Some compelling examples include automated expansion of name searches and name validation (Boyle et al. 2013; Patterson 2014). Many algorithms for finding taxonomic names in data files and text are available (see Thessen et al. 2012 for a review) and some of these tools can automatically parse the found name string and return the current name based on a user-defined authority such as the Global Names Recognition and Discovery tools and services (GNRD 2016). In order for Steve and Lupita to find and correctly aggregate all the data they need to make their decisions, they need to know all the names that have been used to describe the taxa they are interested in. They can use resources like Catalogue of Life and IPNI to find the taxonomic synonyms and homonyms and EOL and uBio to find the vernacular names. This list of names, divided into reconciliation groups, can be used to discover and aggregate the occurrence, trait, genetic, and phylogenetic data they need to answer their questions. Thus, names form the basis of any biodiversity informatics task.

### ***17.2.2 Identification Tools***

People often need to identify an unknown organism, and biodiversity informaticists build tools to help. Lupita and her colleagues might use these identification tools as they inventory the species that live in their park. Steve's team may need to identify new crop pests or pollinators. In recent years there has been a shift in identification technology from identification keys focused primarily on morphological characteristics and aimed at experts, to DNA barcoding and image recognition that can assist both experts and non-experts.

Taxonomists historically published identification keys in books and scholarly articles in the form of dichotomous keys that take a reader through a prescribed set of decisions between two answers ("couplets"). With the advent of computers, software has been developed so that keys are matrix-based (rows for species and columns for their characteristics) and interactive (users can check off characteristics in any order). Many of these platforms allow users to develop their own keys; examples include DELTA—DEscription Language for TAXonomy (DELTA-Intkey; Dallwitz 2010), Lucid (Lucidcentral 2016), and the IDnature guides at DiscoverLife (DiscoverLife 2016). Some keys are designed for offline use, but many are online. Notably, keys often focus on characteristics or traits that will distinguish organisms from each other, not on characteristics that may be useful for phylogenetic or evolutionary analysis or modeling of ecological processes. While digital interactive keys remain useful, critics argue that they have not been widely used or usable outside a narrow audience of taxonomists (see review by Walter and

Winterton 2007). Given the taxonomic impediment mentioned above, more scalable and broadly usable approaches to identification are needed.

Molecular methods such as DNA barcoding have been developed for rapid identification (Hebert et al. 2003). DNA barcoding involves sequencing short stretches of DNA that are expected to be unique at the species level, and comparing the sequences to a reference library of these sequences from known organisms. Unlike traditional keys, it is not necessary to have a mostly complete specimen or a good look at the unknown organism with observable morphological characteristics. A sample with degraded DNA fragments often suffices. Thus, while DNA barcoding can help scientists and their citizen collaborators confirm their identifications of their organisms of study (Shen et al. 2013) or rapidly inventory inhabitants of a study area (e.g., Miller et al. 2016), it is also used by customs agents to determine if imported goods are made from threatened species regulated by CITES (Staats et al. 2016). It can be used by students who send in or even analyze samples from their schoolyards (Santschi et al. 2013) or from restaurant meals (e.g., “SushiGate”, Wong and Hanner 2008). This technique only works if the sequences can be compared to a well-curated database of expertly identified reference sequences (Barcode of Life Datasystem, BOLD, Ratnasingham and Hebert 2007). If barcode sequences are not found in BOLD or produce ambiguous results, this may indicate problems with the database or the methods (Collins and Cruickshank 2012; Lis et al. 2016) or scientists may conclude that a species new to science has been found and merits naming (reviewed in Miller et al. 2016).

Computer vision is another growing area for biodiversity identification (additional references in Thessen 2016). LeafSnap, for example, is a mobile phone application that citizen scientists in North America can use to identify trees by taking a photo of a leaf (Kumar et al. 2012). The image is segmented (separated from the background) and the shape of the leaf is compared to known leaf shapes. New approaches to machine classification of biodiversity images are tested in an annual event called LifeCLEF (Joly et al. 2015).

Finally, both book-based and online field guides remain popular resources for identification (Farnsworth et al. 2013). While the “unsung” artists and authors of paper-based field guides deserve more recognition, geographic biases in field guide publication and sales may have wide-ranging impacts (Holt 2016). Can free, online guides even the playing field? Resources such as Scratchpads (Smith et al. 2009), EOL, and iNaturalist have large numbers of images, maps, and descriptions that can be used for identification all over the world. EOL and iNaturalist have teamed up to make it easy for people to create their own field guides (California Academy of Sciences 2016). The Cornell Laboratory of Ornithology is developing Merlin to support both interactive question-and-answer identification and image recognition (Cornell University 2016a). The increasing availability of online resources means that even searches with a few keywords on Google Images can assist in identification.

Even if Lupita and Steve do not use these resources themselves, they can take advantage of crowd-sourcing and post their images of unknown organisms on a variety of platforms. A growing community of both professionals and citizen

scientists will bring their expertise and their online searching skills to provide identifications.

### ***17.2.3 Occurrence Data and Species Distributions***

Biodiversity informatics helps answer the question, “What lives here?” Lupita needs a list of organisms in her park before she can determine which of them is threatened by changing conditions. Or, a different formulation of the question might be “What is the geographic distribution of this organism?” Steve could use maps indicating where his crop variety of interest grows to establish the typical soil and weather it needs to do well, then he could look for other varieties or crop relatives that have similar distributions or habitat preferences. Because it is impossible to conduct up-to-date inventories at all places at all times, we must rely on samples of data and algorithms to determine an estimated distribution. Many biodiversity informaticists are building systems to collect, manage, aggregate, serve, and analyze these primary occurrence data, and the maps that result from them.

Occurrence data may start in field notebooks and specimen tags on physical specimens. Explorers like Lewis and Clark or biologists like Darwin collected specimens during their journeys and logged basic information about them in the field. As those specimens were deposited in museums, they would have been catalogued individually or in groups with metadata: information about who collected what, where, and when. Modern digital museum catalogs aim for georeferenced, time-stamped specimen records. This is much easier now that collectors use Global Positioning Systems (GPS) and specimen data is “born digital”, but software such as BioGeoBIF (Hill et al. 2009) and GEOLocate (Rios and Bart 2010) can retroactively estimate coordinates (and uncertainty estimates) for specimens that previously had only vague locations handwritten on tags. Large scale efforts are now underway to digitize the specimen data associated with physical specimens (e.g., iDigBio, Page et al. 2015), and increasingly these efforts use crowd-sourcing (Ellwood et al. 2015).

Biodiversity informaticists have developed digital museum catalog software such as BRAHMS (University of Oxford 2016), EMu (Axiell Group 2016), Specify (Specify Software Project 2016), Digital Information System for Natural History Collections (DINA; DINA Consortium 2016), and Symbiota (Gries et al. 2014). In addition to providing ready access to the specimen-based occurrence information, this software typically helps collection managers manage specimen loans and reports about their collections.

Occurrence data also results from observations that are not associated with museum specimens. These observations may be associated with sound or video recordings (e.g., Cornell University’s Macaulay Library; Cornell University 2016b), or with photographs (e.g., California Academy of Sciences 2016), or with no voucher at all (e.g., eBird; Audubon and Cornell Lab of Ornithology 2016).

Some occurrence data can be derived from nucleotide sequence data in GenBank (reviewed in Gratton et al. 2016). With the advent of citizen science and accelerating improvements in digital and sensor technology it is now possible to generate vast amounts of timely occurrence data. For example, wildlife biologists and citizens set up camera traps that are triggered by the motion of passing wildlife and capture large numbers of images (Fig. 17.1). These assist in understanding what organisms live in an area and when and how they are active.

Darwin Core is a mature, widely-used standard for the sharing of biodiversity occurrence data (Wieczorek et al. 2012). Darwin Core is a product of the Biodiversity Information Standards organization, known by the acronym TDWG due to its historical name, Taxonomic Databases Working Group (TDWG 2016). Darwin Core is used by the Global Biodiversity Information Facility (GBIF 2016) to create its central index of more than 600 million occurrences, which in turn is used by mapping projects such as Biodiversity Information Serving our Nation (BISON; USGS 2016), Map of Life (Jetz et al. 2012; MOL 2016), and AquaMaps (Kaschner et al. 2016).

Researchers can use standardized digital occurrence data to estimate the distributions of individual species, better understand their habitat preferences, and generate maps of species richness. A variety of approaches used for species distribution modeling are reviewed in Peterson et al. (2015). Ecological niche modeling uses correlations with environmental parameters available for each occurrence point (or lack of a point), or our current understanding of a species niche, or both. Map of Life (Jetz et al. 2012) uses absence information to better account for knowledge that a species was not observed during inventory, not just knowledge that it was observed. Process or hybrid distribution modeling integrates niche and dispersal characteristics and sometimes includes interactions with other species (reviewed in Evans et al. 2016). Many of these approaches will be useful to Steve and Lupita as they attempt to forecast species distributions under future climate change scenarios.

Occurrence data are used in many other ways (see GBIF Science Committee 2016 for a recent review). Simple estimates of range size from GBIF data were recently combined with protected areas maps (ProtectedPlanet 2016) and the PRE-DICTS (Projecting Responses of Ecological Diversity In Changing Terrestrial Systems) database to conclude that species diversity is higher inside of protected areas than in matched areas outside them (Gray et al. 2016). The National Phenology Network (Schwartz et al. 2012) collects additional information about seasonal events such as flowering or migration so that researchers can extend their understanding of distributions to include timing impacts of climate change.

Lupita may use species distribution information to prioritize her resources to assist species that have very limited ranges, or to ensure the quality of habitat critical for migration. She may be able to argue that she should have more resources because more precious diversity is in her park compared to others. Steve, on the other hand, may be able to identify promising species or varieties with current distributions that do well in the forecasted climate. But knowing species distributions will not be enough—this is just the next step on the path to solutions.

The screenshot shows the eMAMMAL website interface. At the top, there is a navigation bar with the following items: Home, About, View Photos, Explore Projects, Browse Data, Resources, and a search bar. Below the navigation bar, there is a search form with the following fields: Common Species Name (with a dropdown arrow), Project, Subproject, From (Date), To (Date), and Region (with a dropdown menu showing '- Any -'). A blue 'Apply' button is located at the bottom right of the search form. Below the search form, there is a grid of six camera trap images. Each image is accompanied by a project name and social media sharing icons (Facebook, Twitter, Print). The projects shown are: North Carolina's Candid Critters Project (three images), Smithsonian Borneo Mammal Survey at LEWS (three images), and Urban to Wild Project (one image).

Fig. 17.1 Search page for eMammal (Smithsonian Institution 2016a), a repository for camera trap images from projects around the world. Each image has geospatial and time stamp information that can be used to analyze activity or compare species diversities



### ***17.2.4 Phylogenetic Trees***

For both of our user stories it may be helpful to identify the nearest relatives of the relevant organisms and their traits. For example, wild relatives of crops likely have phenotypic variation that could be leveraged to improve qualities of cultivated species. Relatives of species in the national park may live in climates similar to the projected future of the park and may or may not have traits that help them thrive. In both cases, in order to estimate whether the species can adapt (in a genetic sense) quickly enough to the rate of change of the environment it is critical to understand how related species have changed over short and long time scales. Evolutionary biologists called biological systematists construct phylogenetic trees, or hypotheses of evolutionary relationships among organisms to address such questions. They use algorithms and tools developed by evolutionary informaticists (reviewed in Parr et al. 2012).

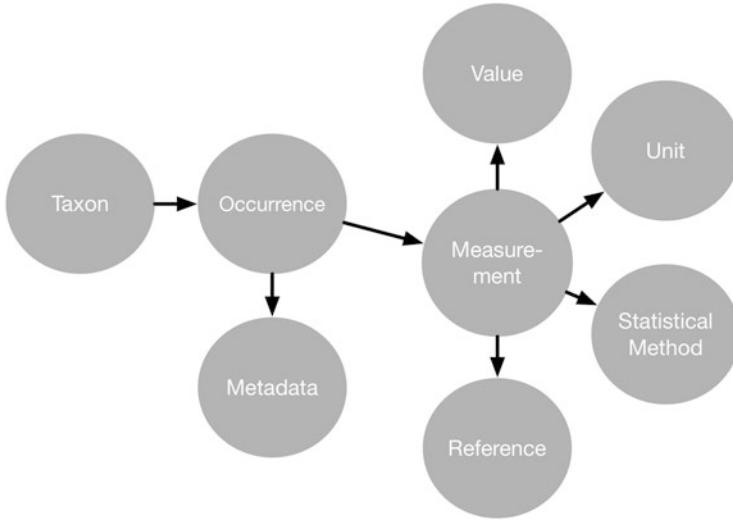
Ideally, we would refer Steve and Lupita to a database with the single true Tree of Life. After all, living organisms today are thought to descend from a common ancestor and systematists have been working for hundreds of years to understand that history of speciation and extinction since life began. Even if we had already named all the organisms that ever existed (we haven't, as noted earlier) it is still a challenging proposition to correctly arrange all of these organisms in a structure that we know now would not be a simple tree, given endosymbiosis (Archibald 2015), gene transfer (Nakhleh 2013), and hybridization (natural or by breeders).

If Steve or Lupita wanted to find phylogenetic trees for their species of interest, they could search the literature and hope to find studies with exactly the species they are looking for. The repository TreeBASE (The Phyloinformatics Research Foundation, Inc. 2016; Sanderson et al. 1994) was designed to make it easier to find published trees. Today, we would refer Steve and Lupita to Open Tree of Life (2016), which aims to synthesize our current knowledge across the entire tree of life (Hinchliff et al. 2015).

### ***17.2.5 Taxa and Their Traits***

All organisms have traits, but the exact definition of "trait" is unclear. In the context of our user stories, traits can be any characteristic of an organism, whether it describes life history, eating habits, morphology, habitat, etc. Taxa are defined by their traits (with the exception of some newly discovered prokaryotic taxa described by a gene sequence). What Steve and Lupita are really interested in are the traits of the taxa in their care. They need to identify specific traits that make an organism better able to adapt to a given environment and what traits might make an organism more vulnerable to change. Then, they need to look for those traits in their organisms of interest.

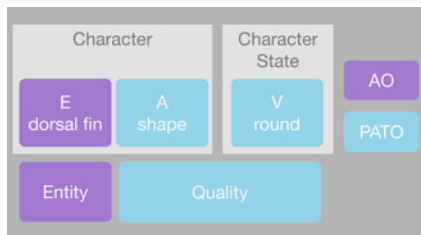




**Fig. 17.2** Data model for trait-related concepts

Trait data are modeled by trait:value pairs that can also be referred to as measurement:value pairs, character:character state pairs, or phenotypes. An example of a trait:value pair would be flower color:purple or habitat:marine. Two important repositories for trait information are the Encyclopedia of Life TraitBank (Parr et al. 2016) and Phenoscope (Phenoscope 2016a). TraitBank contains over 11 million records about more than 330 traits from over 1.7 million taxa. TraitBank uses a data model based on the Darwin Core Archive star schema with occurrences as the basis of record (Fig. 17.2). An extension file is provided that contains measurements. Every occurrence can have multiple measurements. The measurement can be about a taxon or about the occurrence. Measurements are presented with values, units, statistical methods, and a reference. This model can accommodate data from individual organisms or summary statistics about a taxon. The data in TraitBank come from a variety of sources including online databases, data mining, and the published literature. Phenoscope contains nearly 10,000 records about over 5000 taxa, all fish. The data in Phenoscope are modeled as character:character state pairs called “EQ formalisms” (Fig. 17.3) (Mabee et al. 2007). The data in Phenoscope are from the published literature. An important difference between TraitBank and Phenoscope is that the latter has semantic reasoning enabled (Balhoff et al. 2011).

Both repositories use ontologies and controlled vocabularies to standardize the measurement:value and character:character state pairs. The biodiversity literature is plagued with inconsistent term usage (e.g., Yoder et al. 2010), so normalizing and standardizing terms is an important challenge for biodiversity data aggregation. Several ontologies exist that can be used to describe organism traits and metadata (see Table 1 in Thessen et al. 2016), but there are still many necessary terms that are



**Fig. 17.3** Use of the Phenotype and Trait Ontology (PATO) and anatomy ontologies (AO, these may differ across organisms) to capture morphological traits of organisms as entity:quality statements, or formalisms. Modified from Phenoscope (2016b)

not part of any ontology or vocabulary. Many communities have not agreed on a set of terms and their definitions, but where ontologies are well developed, machine-readable species descriptions can be used (Balhoff et al. 2013).

Steve and Lupita need to know the traits of the organisms they work with in order to understand why an organism might do well or poorly in different circumstances. The data model for linking traits to taxa has been demonstrated with semantic species descriptions (Balhoff et al. 2013), but forming a relationship between traits (phenotypes) and environmental conditions has not been formalized (discussed in detail by Thessen et al. 2016). TraitBank and Phenoscope can be easily queried via web services or a search box. While both repositories combined contain many thousands of records, the majority of biodiversity knowledge is still only accessible through human-readable text. Both of our users could get a start on gathering the trait data they need using TraitBank and Phenoscope, but would need to turn to the literature to complete their research. They would likely have to interpret the meaning of the original authors in order to standardize the terms used by different researchers. Thus, if Steve or Lupita wanted to find the traits in common to organisms that live in a specific environment, they would not be able to use a fully automated query, but would need a workflow with a manual component.

### 17.2.6 Digital Biodiversity Literature

Lupita or Steve will likely consult online versions of the peer-reviewed literature, new and old, in their daily activities. Nearly all modern research papers are now published online. But it is especially important that the scholarly literature on biodiversity be digitally available in the most useful forms possible, even if it is hundreds of years old. Why? Because taxonomists must review and cite original descriptions whenever they want to describe new species or change nomenclature based on new information. Moreover, knowledge of rarely encountered species may only be found in the older literature. Finally, even for well-studied organisms there is a treasure trove of information locked in the text of scholarly literature.

While online scientific journals are common now and open access is increasing (Solomon et al. 2013), bringing the older paper-based literature to the World Wide Web is a huge effort. The Biodiversity Heritage Library (BHL 2016) is the result of scanning, performing optical character recognition on, and indexing scientific names in millions of pages by a world-wide consortium of institutions (Gwinn and Rinaldo 2009). BHL already includes over 50 million pages that range from the fifteenth century through today.

The effort to scan the world's biodiversity literature is not yet complete, but researchers are already mining it for information. For example, CharaParser was developed to process natural language sentences in morphological descriptions using machine learning algorithms to produce structured character data (Cui 2012). Other examples of the use of text mining to extract the kind of information that Lupita and Steve need from the literature will be described in the following sections.

### 17.2.7 *Species Associations*

An interaction between two species can impact more than the organisms directly involved (e.g., Schmitz et al. 2000). Organisms can switch from being omnivorous to either entirely carnivorous or herbivorous depending on the food species present. Keystone species can have a profound effect on an ecosystem through their associations, often resulting in a completely different ecosystem in their absence. Climate change, the driver of both user stories, is known to alter trophic, competitive, parasitic, and mutualistic interactions (Tylianakis et al. 2008). Field studies have shown that species interactions can strongly influence an ecosystem's response to climate change (Suttle et al. 2007). It is not enough to understand the traits of organisms, Steve and Lupita must also understand how these traits interact to result in an integrated, functioning ecosystem.

Most of the interaction data that Steve and Lupita need are present only in the published literature and are not machine-readable; however, several efforts are underway to digitize this information. The Global Biotic Interaction database (GloBI) contains over two million associations between nearly 130,000 taxa (Poelen et al. 2014). The data in GloBI are modeled simply as Taxon:Interaction:Taxon. The taxa are recorded using an identifier from an accepted aggregator or taxonomic authority. The interactions are described using terms from the Relations Ontology (RO Project 2016). Every interaction statement is related to a citable reference and can be tagged with georeferencing metadata. The Gulf of Mexico Species Interactions (GoMexSI) database is similar to GloBI, but focuses on trophic interactions in the Gulf of Mexico derived primarily from surveys of gut contents (Simons et al. 2013). Alternatively, mangal.io stores whole networks rather than individual interaction observations (Poisot et al. 2015). It is true that networks can be built from many individual observations of interactions that originate from the

same place; GloBI and GoMexSI focus on the individual interaction while mangal.io focuses on observations of the entire network.

Unfortunately for Steve and Lupita, there is no easy way to know how a change in an individual interaction will affect an entire ecosystem aside from empirical study. Methods for using the contents of GloBI, GoMexSI, and mangal.io for predicting ecosystem-level responses to changes in species associations are being developed through predictive models (e.g., Tarnecki et al. 2016). The transition from a table-based, taxonomically-organized data schema to the graph-based, system-organized data schema (represented by GloBI, GoMexSI, and mangal.io) makes it much easier for Steve and Lupita to find out how taxa are interconnected and thus can give them a place to start their own investigations.

### ***17.2.8 Ontologies and Controlled Vocabularies***

A controlled vocabulary is a standard list of terms (often with definitions) that a community of users has agreed to use, for example, in a metadata record or in a database. An ontology is similar, but also relates the terms to each other and is represented in a machine-readable format. A controlled vocabulary facilitates human communication and standardization, while an ontology facilitates machine understanding. Machine-readability is important for scaling up analyses by removing manual components of an analytical or data management task.

Both controlled vocabularies and ontologies enable unambiguous term usage within a community, but an ontology can also enable reasoning, a type of machine learning (Jensen and Bork 2010). A simple reasoning task would be to tell a machine, using ontologies, that: (1) all birds have wings; and (2) a robin is a bird. The machine could “learn” that: (3) robins have wings. This is a very simple form of the technology used by IBM’s Watson (Gliozzo et al. 2013), for example. Ontologies are a very important part of machine-readable data models and text mining algorithms in biodiversity in support of taxonomic, phylogenetic, and evolutionary biology studies (Cui 2012; Balhoff et al. 2010, 2011, 2013; Midford et al. 2013; Dececchi et al. 2015; Manda et al. 2015; Walls et al. 2014; Chawuthai et al. 2016).

Ontologies used in biodiversity are typically bottom-up, community efforts with varying degrees of coordination between them and within them. OBOFoundry and BioPortal (Noy et al. 2009) are important nuclei for ontology development and maintenance in biodiversity (Smith et al. 2007). Building and maintaining an ontology can be difficult work requiring dedicated staff to address user needs (e.g., requesting new terms, modifying existing terms) and outreach efforts to gain community support and consensus (e.g., term definitions, Seltmann et al. 2013), which can be the most difficult part of ontology development. Some ontologies, such as the Gene Ontology, have developed their own tools for ontology search and browse (AmiGO, Carbon et al. 2009), but OntoBee is available (Xiang et al. 2011). Many ontologies use github (GitHub Inc. 2016) and sourceforge (Slashdot Media 2016) as issue trackers for managing requests for new terms.

Many ontologies that were built by different communities overlap to varying degrees and thus require alignment (Smith et al. 2007). Despite the increase in analytical and reasoning power that an ontology affords us, significant human effort is required to do the initial building of the ontology and then support the maintenance. This human effort is the largest bottleneck in the ontological process. In addition, many aspects of biodiversity science are arguably incompatible with the logical constructs of ontology (Franz 2010).

Fortunately, users of data do not need to be ontology experts to take advantage of their reasoning power. An ideal system would provide semantic technology “under the hood” through a user friendly interface. Several methods have been proposed for biodiversity repositories to implement semantic technology (Malaverri et al. 2009; Lapp et al. 2011; Amanqui et al. 2014; Stucky et al. 2014) and some repositories (e.g., Phenoscope, BioHub, Read et al. 2016) already use semantic reasoning in this way; however, much of the data Steve and Lupita need do not yet have a strong ontological backbone. In biodiversity science for now, many of the connections a machine could make automatically using an ontology have to be made manually—with a few exceptions that are supported by well-developed anatomical ontologies (e.g., Yoder et al. 2010).

### ***17.2.9 Biodiversity Genomics***

Steve will be very interested in any genomic and phenotypic information on strains and crop wild relatives he might use in his breeding program. A full review of molecular biology informatics is out of scope for this chapter, but several efforts are of particular interest to biodiversity informaticists.

Germplasm repositories (sometimes called “genebanks,” not to be confused with GenBank) exist all over the world to preserve seeds and other living genetic material, largely for the world’s crops. GRIN-Global (Germplasm Repository Information Network, The GRIN-Global Project 2016) provides access to information not only on these physical accessions but on their known properties including genes and trait information. After using GRIN-Global to identify candidates for his breeding program Steve can use the system to contact researchers and obtain material. Meanwhile, the Global Genome Initiative (Smithsonian Institution 2016b) is gathering samples from across the tree of life, not just those of agricultural significance, that will be suitable for full-genome sequencing. In most but not all cases, type specimens are too old to have suitable DNA so fresh material needs to be gathered. Information about those samples and their suitability for sequencing is available at [the Global Genome Biodiversity Network](#) (GGBN 2011+, Droege et al. 2014). Both GGBN and GBIF already use GGBN’s proposed TDWG standard (Droege et al. 2016). The Biological Collections Ontology (Walls et al. 2014) provides appropriate semantics in this area.

While some model and crop organisms have large enough genomics communities that can each sustain their own repository and analysis platform, for the long tail

of biodiversity there are shared collaborative tools. For example, the i5K Workspace provides tools for annotating the i5K Consortium's planned 5000 arthropod genomes (Poelchau et al. 2015). Genomic projects are described and aggregated at the Genomes Online Database (Mukherjee et al. 2016).

### 17.3 Next Steps

Our users, Steve and Lupita, have a challenging task in front of them. In order to answer their questions, they need to use data and infrastructure from many different sources. Some of the data they need may not be digital, may not exist, or may be behind a paywall. Despite the plethora of repositories, ontologies, standards, vocabularies, and web services discussed above, a seamless, biodiversity question and answer engine is still out of reach. Infrastructure, especially data formats that are compatible across domains, remain a large gap. Bridges across existing infrastructure are patchy. Where infrastructure is in place, the data to fuel tools and services are incomplete. The management and sustainability of data and infrastructure remain a huge challenge, as documented in the Global Biodiversity Informatics Outlook (Hobern et al. 2013). The absence of a universal system of unambiguous, unique identifiers is a major impediment to biodiversity science because without proper identifiers automated linking of data and properly attributing work become prohibitive (Page 2008). Satisfying our user stories will require additional investment in infrastructure and data digitization.

The largest impediment to our user stories is the lack of digitized and standardized data. This is an important bottleneck because digitization often has manual steps and standardization requires custom solutions for heterogeneous data sets. One promising solution is the development of techniques that use machine learning and semantic technology to automatically extract data from human-readable formats (Thessen et al. 2012), infer missing data (Dececchi et al. 2015), and perform automated QA/QC on existing data (OCR repair). These methods have not yet reached their full potential due to the need for additional development and slow uptake by domain scientists. Data can also be missing due to the general undersampling of most environments and taxa (Hortal et al. 2015) and the lengthy analysis time most samples require. Expansion of citizen science to include data acquisition and analysis can help increase throughput (Theobald et al. 2015; Chandler et al. 2017) in addition to investing in machine learning algorithms that can infer data in undersampled areas (Thessen 2016). Promising techniques for analyzing remote sensing imagery are starting to close the gap between what can be learned from observation on the ground and from monitoring of large areas by satellites and aircraft (Gillison et al. 2016). Significant investments in automated methods and enabling citizen scientist participation are necessary to provide the biodiversity informatics infrastructure with the data it needs to serve users.

Finding and integrating data gathered from experiments and resulting from management actions is difficult both in biodiversity and agricultural science. Both

Lupita and Steve will benefit if they can learn from many small experiments. Interpreting GBIF data is difficult without knowing which of the vast number of data collection protocols were used to collect it. While there are agricultural thesauri that include some research techniques and management concepts (e.g., National Agricultural Library Thesaurus and Glossary, USDA 2016) they were developed primarily for the annotation of literature, not data. The Experimental Factor Ontology (EFO) was developed to describe experimental variables, but was developed for molecular biology (Malone et al. 2010). The Ontology for Biomedical Investigations (OBI) is widely used, but was initially developed for biomedical clinical investigations (Bandrowski et al. 2016). The EXPO ontology was introduced in 2006, but is not widely used (Soldatova and King 2006). The Parasite Experiment Ontology (PEO) is limited to parasites (Cross et al. 2011). Both EFO and OBI have the potential to be applied within biodiversity-focused experiments. Similarly, the ICASA data standard (White et al. 2013) has been developed for describing crop experiments; it could be tested more broadly, further developed as an ontology, and used as a model for biodiversity experiments.

## 17.4 Conclusion

In our focus on users like Steve and Lupita, we have not covered all areas of biodiversity informatics. We did not discuss the development of tools or standards to support policy makers or intergovernmental monitoring of the Convention on Biological Diversity (see the Essential Biodiversity Variables, Pereira et al. 2013). We did not suggest how biodiversity information could be included in decision support tools for water and agricultural resource management. This will be important for ensuring a healthy biosphere while feeding the nine billion people expected to be on our planet in 2050 (Godfray et al. 2010). We barely touched on the rise of data-intensive biodiversity science (Kelling et al. 2009).

However, the basic biodiversity informatics concepts presented in this chapter will be at the heart of those use cases, just as they were important for Lupita and Steve. The tremendous growth in biodiversity informatics over the last few decades suggests there is cause for optimism.

## References

- Amanqui FK, Serique KJ, Cardoso SD et al (2014) Improving biodiversity data retrieval through semantic search and ontologies. Paper presented at the 2014 IEEE/WIC/ACM international joint conferences on web intelligence and intelligent agent technologies, University of Warsaw, 11–14 Aug 2014
- Archibald JM (2015) Endosymbiosis and eukaryotic cell evolution. *Curr Biol* 25(19):R911–R921
- Ashburner M, Ball CA, Blake JA et al (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 25(1):25–29

- Audubon and Cornell Lab of Ornithology (2016) eBird. <http://ebird.org/content/ebird/>. Accessed 21 Nov 2016
- Axiell Group (2016) Emu: transforming data into knowledge. <https://emu.kesoftware.com>. Accessed 21 Nov 2016
- Balhoff JP, Dahdul WM, Kothari CR et al (2010) Phenex: ontological annotation of phenotypic diversity. *PLoS One* 5:e10500
- Balhoff JP, Dahdul WM, Lapp H (2011) Employing reasoning within the Phenoscape knowledgebase. Proceedings of the international conference on biomedical ontology (ICBO), University at Buffalo, 28–30 July 2011, p 230. [http://icbo.buffalo.edu/ICBO-2011\\_Proceedings.pdf](http://icbo.buffalo.edu/ICBO-2011_Proceedings.pdf)
- Balhoff JP, Mikó I, Yoder MJ et al (2013) A semantic model for species description applied to the ensign wasps (Hymenoptera: Evaniidae) of New Caledonia. *Syst Biol* 62:639–659
- Bandrowski A, Brinkman R, Brochhausen M et al (2016) The ontology for biomedical investigations. *PLoS One* 11:e0154556. doi:10.1371/journal.pone.0154556
- Biodiversity Heritage Library (BHL) (2016) BHL: Biodiversity Heritage Library. <http://www.biodiversitylibrary.org>. Accessed 22 Nov 2016
- Bisby FA (2000) The quiet revolution: biodiversity informatics and the internet. *Science* 289:2309–2312
- Boyle B, Hopkins N, Lu Z (2013) The taxonomic name resolution service: an online tool for automated standardization of plant names. *BMC Bioinform* 14:16
- Brickell DC, Alexander C, Cubey JJ et al (eds) (2016) International code of nomenclature for cultivated plants, 9th edn. Belgium, International Society of Horticultural Science
- California Academy of Sciences (2016) Welcome to iNaturalist.org Guides! <http://www.inaturalist.org/guides/>. Accessed 21 Nov 2016
- Carbon S, Ireland A, Mungall CJ et al (2009) AmiGO: online access to ontology and annotation data. *Bioinformatics* 25:288–289
- Convention on Biological Diversity (CBD) (2016) Convention on biological diversity. <https://www.cbd.int>. Accessed 20 Nov 2016
- Chandler M, See L, Copas K et al (2017) Contribution of citizen science towards international biodiversity monitoring. *Biol Conserv*. doi:10.1016/j.biocon.2016.09.004
- Chapman AD (2009) Numbers of living species in Australia and the world report. Commonwealth of Australia, Department of the Environment and Water Resources, Canberra. <http://www.environment.gov.au/biodiversity/abrs/publications/other/species-numbers/index.html>. Accessed 4 Dec 2016
- Chawuthai R, Takeda H, Wuwongse V et al (2016) Presenting and preserving the change in taxonomic knowledge for linked data. *Semant Web* 7(6):589–616. doi:10.3233/SW-150192
- Catalogue of Life (COL) (2016) <http://www.catalogueoflife.org/>. Accessed 20 Nov 2016
- Collins RA, Cruickshank RH (2012) The seven deadly sins of DNA barcoding. *Mol Ecol Resour* 13(6):969–975. doi:10.1111/1755-0998.12046
- Cornell University (2016a) The Cornell Lab: Merlin. <http://merlin.allaboutbirds.org>. Accessed 21 Nov 2016
- Cornell University (2016b) The Cornell Lab of Ornithology Macaulay Library. <http://macaulaylibrary.org>. Accessed 21 Nov 2016
- Costello MJ, Wilson S, Houlding B (2012) Predicting total global species richness using rates of species description and estimates of taxonomic effort. *Syst Biol* 61(5):871–883. <http://doi.org/10.1093/sysbio/syr080>
- Cross V, Stroe C, Hu X et al (2011) Aligning the parasite experiment ontology and the ontology for biomedical investigations using AgreementMaker. Proceedings of International Conference on Biomedical Ontology (ICBO), University at Buffalo, 28–30 July 2011, pp 125–131. [http://icbo.buffalo.edu/ICBO-2011\\_Proceedings.pdf](http://icbo.buffalo.edu/ICBO-2011_Proceedings.pdf). Accessed 4 Dec 2016
- Cui H (2012) CharaParser for fine-grained semantic annotation of organism morphological descriptions. *J Am Soc Inf Sci Technol* 63:738–754



- Dallwitz MJ (2010) Overview of the DELTA System. <http://delta-intkey.com/www/overview.htm>. Accessed 21 Nov 2016
- Decechi TA, Balhoff JP, Lapp H (2015) Toward synthesizing our knowledge of morphology: using ontologies and machine reasoning to extract presence/absence evolutionary phenotypes across studies. *Syst Biol* 64:936–952
- DINA Consortium (2016) Welcome to the DINA project! <http://www.dina-project.net>. Accessed 21 Nov 2016
- DiscoverLife (2016) IDnature guides. <http://discoverlife.org/mp/20q>. Accessed 21 Nov 2016
- Droege G, Barker K, Astrin JJ et al (2014) The Global Genome Biodiversity Network (GGBN) Data Portal. *Nucleic Acids Res* 42:D607–D612
- Droege G, Barker K, Seborg O et al (2016) The Global Genome Biodiversity Network (GGBN) data standard specification. *Database* 2016: baw125
- Ellwood ER, Dunckel BA, Flemons P et al (2015) Accelerating the digitization of biodiversity research specimens through online public participation. *Bioscience* 65(4):383–396. doi:10.1093/biosci/biv005
- Encyclopedia of Life (EOL) (2016) <http://www.eol.org>. Accessed 20 Nov 2016
- Evans MEK, Merow C, Record S et al (2016) Towards process-based range modeling of many species. *Trends Ecol Evol* 31(11):860–871. doi:10.1016/j.tree.2016.08.005
- Farnsworth EJ, Chu M, Kress WJ et al (2013) Next-generation field guides. *Bioscience* 63(11):891–899. doi:10.1525/bio.2013.63.11.8
- Federhen S (2012) The NCBI Taxonomy database. *Nucleic Acids Res* 40:D136–D143
- Franz N (2010) Biological taxonomy and ontology development: scope and limitations. *Biodivers Inform* 7:45–66
- Global Biodiversity Information Facility (GBIF) (2016) Global Biodiversity Information Facility: free and open access to biodiversity data. <http://www.gbif.org>. Accessed 22 Nov 2016
- GBIF Science Committee (2016) GBIF science review 2016. <http://www.gbif.org/resource/82873>. Accessed 14 Nov 2016
- Global Genome Biodiversity Network (GGBN) (eds) (2011+, continuously updated) The GGBN Data Portal. GGBN Secretariat, NMNH, Washington, DC. Compiled by GGBN Technical Management, BGBM, Berlin, Germany. <http://data.ggbn.org>. Accessed 22 Nov 2016
- Gillison A, Asner G, Fernandes E et al (2016) Biodiversity and agriculture in dynamic landscapes: integrating ground and remotely-sensed baseline surveys. *J Environ Manag* 177:9–19
- GitHub Inc. (2016) GitHub. <https://github.com>. Accessed 22 Nov 2016
- Global Names Index (GNI) (2016) <http://gni.globalnames.org>. Accessed 20 Nov 2016
- Global Names Recognition and Discovery (GNRD) (2016) <http://gnrd.globalnames.org/>. Accessed 20 Nov 2016
- Gliozzo A, Biran O, Patwardhan S et al (2013) Semantic technologies in IBM Watson™. In: Proceedings of the fourth workshop on teaching natural language processing, Aug 4–9 2013. Sofia, Bulgaria, pp 85–92. <http://www.aclweb.org/anthology/W13-3413>. Accessed 29 Nov 2016
- Godfray H, Beddington J, Crute I et al (2010) Food security: the challenge of feeding 9 billion people. *Science* 327:812–818
- Gratton P, Marta S, Bocksberger G et al (2016) A world of sequences: can we use georeferenced nucleotide databases for a robust automated phylogeography? *J Biogeogr*. doi:10.1111/jbi.12786
- Gray CL, Hill SLL, Newbold T et al (2016) Local biodiversity is higher inside than outside terrestrial protected areas worldwide. *Nat Commun* 7:12306. doi:10.1038/ncomms12306
- Gries C, Gilbert E, Franz N (2014) Symbiota – a virtual platform for creating voucher-based biodiversity information communities. *Biodiv Data J* 2:e1114
- Gwinn NE, Rinaldo C (2009) The Biodiversity Heritage Library: sharing biodiversity literature with the world. *IFLA J* 35:25–34
- Hebert PDN, Cywinska A, Ball SL et al (2003) Biological identifications through DNA barcodes. *Proc Biol Sci* 270(1512):313–321. doi:10.1098/rspb.2002.2218

- Heidorn PBH (2011) Biodiversity informatics. *Bull Am Soc Inf Sci Technol* 37:38–44
- Hill AW, Guralnick R, Flemons P et al (2009) Location, location, location: utilizing pipelines and services to more effectively georeference the world's biodiversity data. *BMC Bioinf* 10(Suppl 14):S3. doi:[10.1186/1471-2105-10-S14-S3](https://doi.org/10.1186/1471-2105-10-S14-S3)
- Hinchliff CE, Smith SA, Allman JF et al (2015) Synthesis of phylogeny and taxonomy into a comprehensive tree of life. *Proc Natl Acad Sci U S A* 112(41):12764–12769. doi:[10.1073/pnas.1423041112](https://doi.org/10.1073/pnas.1423041112)
- Hoborn D, Apostolico A, Arnaud E et al (2013) Global biodiversity informatics outlook: delivering biodiversity knowledge in the information age. GBIF Secretariat 41 p. <http://www.gbif.org/resource/80859>. Accessed 29 Nov 2016
- Holt RD (2016) Geographical variation in the availability of natural history field guides: personal reflections, causes, and consequences. *Am Nat* 188S:S90–S95
- Hortal J, De Bello F, Alexandre J et al (2015) Seven shortfalls that beset large-scale knowledge of biodiversity. *Annu Rev Ecol Evol Syst* 46:523–549
- The International Plant Names Index (IPNI) (2012) <http://www.ipni.org>. Accessed 20 Nov 2016
- Integrated Taxonomic Information System (ITIS) (2016) <http://www.itis.gov>. Accessed 11 Nov 2016
- Jensen LJ, Bork P (2010) Ontologies in quantitative biology: a basis for comparison, integration, and discovery. *PLoS Biol* 8(5):e1000374
- Jetz W, McPherson JM, Guralnick RP (2012) Integrating biodiversity distribution knowledge: toward a global map of life. *TREE* 27:151–159
- Joly A, Goëau H, Glotin H et al (2015) LifeCLEF 2015: multimedia life species identification challenges. In: Mothe J, Savoy J, Kamps J et al (eds) *Experimental IR meets multilinguality, multimodality, and interaction: 6th international conference of the CLEF Association, CLEF'15, Toulouse, France, September 8–11, 2015, Proceedings*. Springer International Publishing, Cham, pp 462–483. doi:[10.1007/978-3-319-24027-5\\_46](https://doi.org/10.1007/978-3-319-24027-5_46)
- Kaschner K, Kesner-Reyes K, Garilao C (2016) AquaMaps: predicted range maps for aquatic species. [www.aquamaps.org](http://www.aquamaps.org), Version 08/2016
- Kelling S, Hochachka WM, Fink D et al (2009) Data-intensive science: a new paradigm for biodiversity studies. *Bioscience* 59(7):613–620. doi:[10.1525/bio.2009.59.7.12](https://doi.org/10.1525/bio.2009.59.7.12)
- King AMQ (ed) (2011) *Virus taxonomy: classification and nomenclature of viruses: ninth report of the International Committee on Taxonomy of Viruses*. Elsevier, Amsterdam
- Kumar N, Belhumeur PN, Biswas A (2012) Leafsnap: a computer vision system for automatic plant species identification. In: Fitzgibbon A, Lazebnik S, Perona P et al (eds) *Computer Vision – ECCV 2012: 12th European conference on computer vision, Florence, Italy, October 2012. Proceedings Part II*. Springer, Heidelberg, pp 502–516
- Lapage SP, Sneath PHA, Lessel EF et al (eds) (1992) *International code of nomenclature of bacteria*. ASM Press, Washington, DC
- Lapp H, Morris RA, Catapano T et al (2011) Organizing our knowledge of biodiversity. *Bull Assoc Inf Sci Technol* 37:38–42
- Leary PR, Remsen DP, Norton CN et al (2007) uBioRSS: tracking taxonomic literature using RSS. *Bioinformatics* 23:1434–1436
- Lis JA, Lis B, Ziaja DJ (2016) In BOLD we trust? A commentary on the reliability of specimen identification for DNA barcoding: a case study on burrower bugs (Hemiptera: Heteroptera: Cydnidae). *Zootaxa* 4114(1):83–86. doi:[10.11646/zootaxa.4114.1.6](https://doi.org/10.11646/zootaxa.4114.1.6)
- Lucidcentral (2016) Lucid. <http://www.lucidcentral.com>. Accessed 21 Nov 2016
- Mabee PM, Ashburner M, Cronk Q et al (2007) Phenotype ontologies: the bridge between genomics and evolution. *Trends Ecol Evol* 22(7):345–350. doi:[10.1016/j.tree.2007.03.013](https://doi.org/10.1016/j.tree.2007.03.013)
- Malaverri JG, Vilar B, Medeiros CB (2009) A tool based on web services to query biodiversity information. In: *Proceeding of the 5th international conference on web information systems and technologies (WEBIST)*, Lisbon Portugal, 23–26 March 2009, pp 305–310
- Malone J, Holloway E, Adamusiak T et al (2010) Modeling sample variables with an Experimental Factor Ontology. *Bioinformatics* 26(8):1112–1118

- Manda P, Balhoff JP, Lapp H (2015) Using the phenoscape knowledgebase to relate genetic perturbations to phenotypic evolution. *Genesis* 53:561–571
- McNeill J, Barrie FR, Buck WR et al (2012) International Code of Nomenclature for algae, fungi, and plants. *Regnum Veg* 154(1):208
- Midford PE, Decechi TA, Balhoff JP et al (2013) The vertebrate taxonomy ontology: a framework for reasoning across model organism and species phenotypes. *J Biomed Semant* 4:34
- Miller SE, Hausmann A, Hallwachs W et al (2016) Advancing taxonomy and bioinventories with DNA barcodes. *Philos Trans R Soc B Biol Sci* 371(1702):20150339
- Map of Life (MOL) (2016) Map of Life: putting biodiversity on the map. <https://mol.org>. Accessed 22 Nov 2016
- Mora C, Tittensor DP, Adl S et al (2011) How many species are there on Earth and in the ocean? *PLoS Biol* 9(8):e1001127. doi:10.1371/journal.pbio.1001127
- Morris RA, Barve V, Carausu M et al (2013) Discovery and publishing of primary biodiversity data associated with multimedia resources: The Audubon Core strategies and approaches. *Biodivers Inform* 8(2):185–197. doi:10.17161/bi.v8i2.4117
- Mukherjee S, Stamatis D, Bertsch J et al (2016) Genomes OnLine Database (GOLD) v.6: data updates and feature enhancements. *Nucleic Acids Res*. doi:10.1093/nar/gkw992
- Nakhleh L (2013) Computational approaches to species phylogeny inference and gene tree reconciliation. *TREE* 28:719–728
- National Center for Biotechnology Information: Taxonomy Database (NCBI) (2016) <https://www.ncbi.nlm.nih.gov/taxonomy>. Accessed 20 Nov 2016
- Noy NF, Shah NH, Whetzel PL et al (2009) BioPortal: ontologies and integrated data resources at the click of a mouse. *Nucleic Acids Res*, 37 (Web Server issue):W170–W173. doi:10.1093/nar/gkp440
- OBO Technical WG (2016) The OBO Foundry: phenotypic quality. <http://obofoundry.org/ontology/pato.html>. Accessed 22 Nov 2016
- Open Tree of Life (2016) Open Tree of Life. <https://tree.opentreeoflife.org/>. Accessed 22 Nov 2016
- Page RDM (2008) Biodiversity informatics: the challenge of linking data and the role of shared identifiers. *Brief Bioinform* 9(5):345–354. doi:10.1093/bib/bbn022
- Page LM, MacFadden BJ, Fortes JA et al (2015) Digitization of biodiversity collections reveals biggest data on biodiversity. *Bioscience* 65:841–842. doi:10.1093/biosci/biv104
- Parr CS, Guralnick R, Cellinese N et al (2012) Evolutionary informatics: unifying knowledge about the diversity of life. *TREE* 27:94–103
- Parr CS, Wilson N, Leary P et al (2014) The Encyclopedia of Life v2: providing global access to knowledge about life on Earth. *Biodiv Data J* 2:e1079
- Parr CS, Wilson N, Schulz KS, Leary P, Rice J, Hammock J, Corrigan B (2016) TraitBank: practical semantics for organism attribute data in Special Issue on Semantics for Biodiversity. *Semantic Web* 7(6):577–588. doi:10.3233/SW-150190
- Patterson DJ (2014) Helping protists to find their place in a Big Data world. *Acta Protozool* 53:115–128
- Patterson DJ, Cooper J, Kirk PM et al (2010) Names are key to the big new biology. *TREE* 25:686–691
- Patterson D, Mozzherin D, Shorthouse DP et al (2016) Challenges with using names to link digital biodiversity information. *Biodivers Data J* 4(4):e8080. doi:10.3897/BDJ.4.e8080
- Pereira HM, Ferrier S, Walters M et al (2013) Essential biodiversity variables. *Science* 339(6117):277–278
- Peterson AT, Papeš M, Soberón J (2015) Mechanistic and correlative models of ecological niches. *Eur J Ecol* 1(2):28–38. doi:10.1515/eje-2015-0014
- Phenoscape (2016a) Phenoscape. <http://phenoscape.org>. Accessed 22 Nov 2016
- Phenoscape (2016b) Phenoscape wiki. [http://phenoscape.org/wiki/EQ\\_for\\_character\\_matrices](http://phenoscape.org/wiki/EQ_for_character_matrices). Accessed 29 Nov 2016

- Poelchau M, Childers C, Moore G et al (2015) The i5k Workspace@NAL – enabling genomic data access, visualization and curation of arthropod genomes. *Nucleic Acids Res* 43(Database issue):D714–D719. doi:[10.1093/nar/gku983](https://doi.org/10.1093/nar/gku983)
- Poelen JH, Simons JD, Mungall CJ (2014) Global biotic interactions: an open infrastructure to share and analyze species-interaction datasets. *Ecol Inform* 24:148–159
- Poisot T, Gravel D, Leroux S et al (2015) Synthetic datasets and community tools for the rapid testing of ecological hypotheses. *Ecography (Cop)* 38:001–007. doi:[10.1111/ecog.01941](https://doi.org/10.1111/ecog.01941)
- ProtectedPlanet (2016) Protected planet: discover our thematic areas. <https://protectedplanet.net>. Accessed 22 Nov 2016
- Purves D, Scharlemann JPW, Harfoot M et al (2013) Time to model all life on Earth. *Nature* 493(7432):295–297. doi:[10.1038/493295a](https://doi.org/10.1038/493295a)
- Pyle RL, Michel E (2008) ZooBank: developing a nomenclatural tool for unifying 250 years of biological information. *Zootaxa* 1950:39–50
- Ratnasingham S, Hebert PDN (2007) BOLD: the barcode of life data system. *Mol Ecol Notes* 7:355–364
- Read WJ, Demetriou G, Nenadic G et al (2016) The BioHub knowledge base: ontology and repository for sustainable biosourcing. *J Biomed Semant* 7:30
- Rees T (compiler) (2016) The interim register of marine and nonmarine genera. <http://www.irmng.org>. Accessed 21 Nov 2016
- Remsen D (2016) The use and limits of scientific names in biological informatics. *ZooKeys* 550:207–223. doi:[10.3897/zookeys.550.9546](https://doi.org/10.3897/zookeys.550.9546)
- Ride WDL, Cogger HJ, Dupuis C et al (eds) (1999) International code of zoological nomenclature, 4th edn. International Trust for Zoological Nomenclature, London
- Rios NE, Bart HL (2010) GEOLocate (Version 3.22) [Computer software]. Tulane University Museum of Natural History, Belle Chasse, LA
- RO Project (2016) oborel/obo-relations. <https://github.com/oborel/obo-relations/>. Accessed 22 Nov 2016
- Roskov Y, Abucay L, Orrell T et al (2016) Species 2000 & ITIS catalogue of life, 2016 annual checklist. Digital resource at [www.catalogueoflife.org/annual-checklist/2016](http://www.catalogueoflife.org/annual-checklist/2016). Species 2000: Naturalis, Leiden, the Netherlands ISSN 2405-884X
- Ruggiero MA, Gordon DP, Orrell TM et al (2015) A higher level classification of all living organisms. *PLoS One* 10(4):e0119248. doi:[10.1371/journal.pone.0119248](https://doi.org/10.1371/journal.pone.0119248)
- Sanderson MJ, Donoghue MJ, Piel WH et al (1994) TreeBASE: a prototype database of phylogenetic analyses and an interactive tool for browsing the phylogeny of life. *Am J Bot* 81:183
- Santschi L, Hanner RH, Ratnasingham S et al (2013) Barcoding life's matrix: translating biodiversity genomics into high school settings to enhance life science education. *PLoS Biol* 11(1):1–8 doi:[10.1371/journal.pbio.1001471](https://doi.org/10.1371/journal.pbio.1001471)
- Schmitz OJ, Hambäck PA, Beckerman AP (2000) Trophic cascades in terrestrial systems: a review of the effects of carnivore removals on plants. *Am Nat* 155:144–153
- Schwartz MD, Betancourt JL, Weltzin JF (2012) From Caprio's lilacs to the USA National Phenology Network. *Front Ecol Environ* 10(6):324–327. doi:[10.1890/110281](https://doi.org/10.1890/110281)
- Seltmann KC, Péntzes Z, Yoder MJ et al (2013) Utilizing descriptive statements from the Biodiversity Heritage Library to expand the Hymenoptera Anatomy Ontology. *PLoS One* 8:e55674
- Shen Y-Y, Chen X, Murphy RW (2013) Assessing DNA barcoding as a tool for species identification and data quality control. *PLoS One* 8(2):1–5. doi:[10.1371/journal.pone.0057125](https://doi.org/10.1371/journal.pone.0057125)
- Simons JD, Yuan M, Carollo C et al (2013) Building a fisheries trophic interaction database for management and modeling research in the Gulf of Mexico Large Marine Ecosystem. *Bull Mar Sci* 89:135–160
- Slashdot Media (2016) sourceforge. <https://sourceforge.net/>. Accessed 22 Nov 2016
- Smith B, Ashburner M, Rosse C et al (2007) The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat Biotechnol* 25:1251–1255

- Smith VS, Rycroft SD, Harman KT et al (2009) Scratchpads: a data-publishing framework to build, share and manage information on the diversity of life. *BMC Bioinf* 10(Suppl 14):S6. <http://www.biomedcentral.com/1471-2105/10/S14/S6>
- Smithsonian Institution (2016a) eMammal. <https://emammal.si.edu/>. Accessed 22 Nov 2016
- Smithsonian Institution (2016b) Global genome initiative. <https://ggi.si.edu>. Accessed 22 Nov 2016
- Soldatova LN, King RD (2006) An ontology of scientific experiments. *J R Soc Interface* 3:795–803
- Solomon DJ, Laakso M, Björk BC (2013) A longitudinal comparison of citation rates and growth among open access journals. *J Inf Secur* 7:642–650
- Specify Software Project (2016) Specify. <http://specifyx.specifysoftware.org>. Accessed 21 Nov 2016
- Staats M, Arulandhu AJ, Gravendeel B et al (2016) Advances in DNA metabarcoding for food and wildlife forensic species identification. *Anal Bioanal Chem* 408(17):4615–4630. doi:10.1007/s00216-016-9595-8
- Stucky BJ, Deck J, Conlin T et al (2014) The BiSciCol Triplifier: bringing biodiversity data to the Semantic Web. *BMC Bioinf* 15:257
- Suttle KB, Thomsen MA, Power ME (2007) Species interactions reverse grassland responses to changing climate. *Science* 315:640–642
- Tarnecki JH, Wallace AA, Simons JD et al (2016) Progression of a Gulf of Mexico food web supporting Atlantis ecosystem model development. *Fish Res* 179:237–250. doi:10.1016/j.fishres.2016.02.023
- Taxonomic Databases Working Group (TDWG) (2016) Biodiversity information standards: TDWG. <http://www.tdwg.org>. Accessed 22 Nov 2016
- The GRIN-Global Project (2016) The GRIN-Global Project. <http://www.grin-global.org>. Accessed 3 Dec 2016
- The Phyloinformatics Research Foundation, Inc. (2016) TreeBASE: a database of phylogenetic knowledge. <https://treebase.org/>. Accessed 22 Nov 2016
- Theobald EJ, Ettinger AK, Burgess HK et al (2015) Global change and local solutions: Tapping the unrealized potential of citizen science for biodiversity research. *Biol Conserv* 181:236–244
- Thessen AE (2016) Adoption of machine learning techniques in ecology and earth science. *One Ecosyst* 1:e8621
- Thessen AE, Cui H, Mozzherin D (2012) Applications of natural language processing in biodiversity science. *Adv Bioinforma* 2012:1–17. doi:10.1155/2012/391574
- Thessen AE, Bunker DE, Buttigieg PL et al (2016) Emerging semantics to link phenotype and environment. *PeerJ* 3:e1470
- Tylianakis JM, Didham RK, Bascompte J et al (2008) Global change and species interactions in terrestrial ecosystems. *Ecol Lett* 11:1351–1363
- University of Oxford (2016) BRAHMS database. <http://herbaria.plants.ox.ac.uk/bol/>. Accessed 21 Nov 2016
- United States Department of Agriculture (USDA) (2016) National agricultural library thesaurus and glossary. <http://agclass.nal.usda.gov>. Accessed 22 Nov 2016
- United States Geological Survey (USGS) (2016) Biodiversity Information Serving Our Nation (BISON) – Explore & download U.S. species occurrence data & maps. <https://bison.usgs.gov/#home>. Accessed 22 Nov 2016
- Walls RL, Deck J, Guralnick R et al (2014) Semantics in support of biodiversity knowledge discovery: an introduction to the Biological Collections Ontology and related ontologies. *PLoS One* 9:e89606
- Walter DE, Winterton S (2007) Keys and the crisis in taxonomy: extinction or reinvention? *Annu Rev Entomol* 52(1):193–208. doi:10.1146/annurev.ento.51.110104.151054
- White JW, Hunt LA, Boote KJ et al (2013) Integrated description of agricultural field experiments and production: The ICASA Version 2.0 data standards. *Comput Electron Agric* 96:1–12

- Widrow B, Hartenstein R, Hecht-Nielson R (2005) Eulogy: Karl Steinbuch 1917-2005. IEEE Computational Intelligence Society Newsl 5
- Wieczorek J, Bloom D, Guralnick R et al (2012) Darwin Core: an evolving community-developed biodiversity data standard. PLoS One 7(1):e29715. doi:[10.1371/journal.pone.0029715](https://doi.org/10.1371/journal.pone.0029715)
- Wilson EO (ed) (1988) Biodiversity. National Academies Press, Washington, DC
- Wilson EO (1999) The diversity of life. WW Norton & Company, New York
- Wong EHK, Hanner RH (2008) DNA barcoding detects market substitution in North American seafood. Food Res Int 41:828–837
- WoRMS Editorial Board (2016) World Register of Marine Species. <http://www.marinespecies.org> at VLIZ. Accessed 21 Nov 2016 doi:[10.14284/170](https://doi.org/10.14284/170)
- Xiang Z, Mungall CJ, Ruttenberg A et al (2011) Ontobee: a linked data server and browser for ontology terms. In: Proceedings of international conference on biomedical ontology (ICBO), University at Buffalo, 28–30 July 2011, pp 279–281
- Yoder MJ, Mikó I, Seltmann KC et al (2010) A gross anatomy ontology for Hymenoptera. PLoS One 5:e15991

# Chapter 18

## Lessons from Bioinvasion of Lake Champlain, U.S.A.

Timothy B. Mihuc and Friedrich Recknagel

**Abstract** Freshwater lakes provide ideal habitat for invasive species, such as the zebra mussel, which can weaken lake ecological integrity by altering food web structure and dynamics. This case study utilized 23 years Lake Champlain data to examine relationships among water quality, invasive species, native mysids (*Mysis diluviana*) and the zooplankton community. Canonical correspondence analysis (CCA) was employed to ordinate and qualitatively assess long-term patterns across the datasets, and the hybrid evolutionary algorithm (HEA) revealed quantitative relationships and thresholds. Results from both methods are complementary and suggest that: (1) zebra mussels directly affect rotifer densities by preying on slow moving rotifers, and (2) zebra mussels indirectly affect cladocerans, copepods and mysids by both preying on rotifers and grazing on phytoplankton. The direct and indirect effects of zebra mussels on the zooplankton community as well as on mysids adversely affect the ecological integrity of Lake Champlain. Data ordination by CCA and inferential modelling by HEA proved useful for elucidating long-term food web patterns in the complex Lake Champlain ecosystem.

### 18.1 Introduction

Bioinvasion is an anthropogenic global problem that can alter food web structures and dynamics, drive trophic cascading, and irreversibly affect biodiversity and ecological integrity (Higgins and Vander Zanden 2010). Over time invasive species can also cumulatively change physical and chemical habitat conditions making ecosystems less resilient to stresses (Strayer et al. 2006). Biological invasion interacts in complex ways with transport, land use and global change (Crowl et al. 2008). Growing connectedness and globalization increasingly expose aquatic ecosystems to invasive species. Temperate freshwater lakes are particularly vulnerable

---

T.B. Mihuc (✉)

State University of New York at Plattsburgh, Plattsburgh, NY, USA

e-mail: [mihuctb@plattsburgh.edu](mailto:mihuctb@plattsburgh.edu)

F. Recknagel

University of Adelaide, Adelaide, SA, Australia

e-mail: [friedrich.recknagel@adelaide.edu.au](mailto:friedrich.recknagel@adelaide.edu.au)



to key invasive species including the zebra mussel (*Dreissena polymorpha*), alewife (*Alosa pseudoharengus*) and spiny water flea (*Bythotrephes longimanus*). Invasive zebra mussels have degraded aquatic habitats in Europe and North America for 25 years with estimated total economic costs for electric generation and water treatment of \$267 million in the U.S. between 1989 and 2004 (Connelly et al. 2007).

Zebra mussels are found throughout the Great Lakes, Lake St. Clair and the Mississippi river watershed. Lake Champlain is one of the largest lakes in the United States and has been referred to as the “Sixth Great Lake”. It has been invaded by zebra mussels and alewife since 1992 (Mihuc et al. 2012).

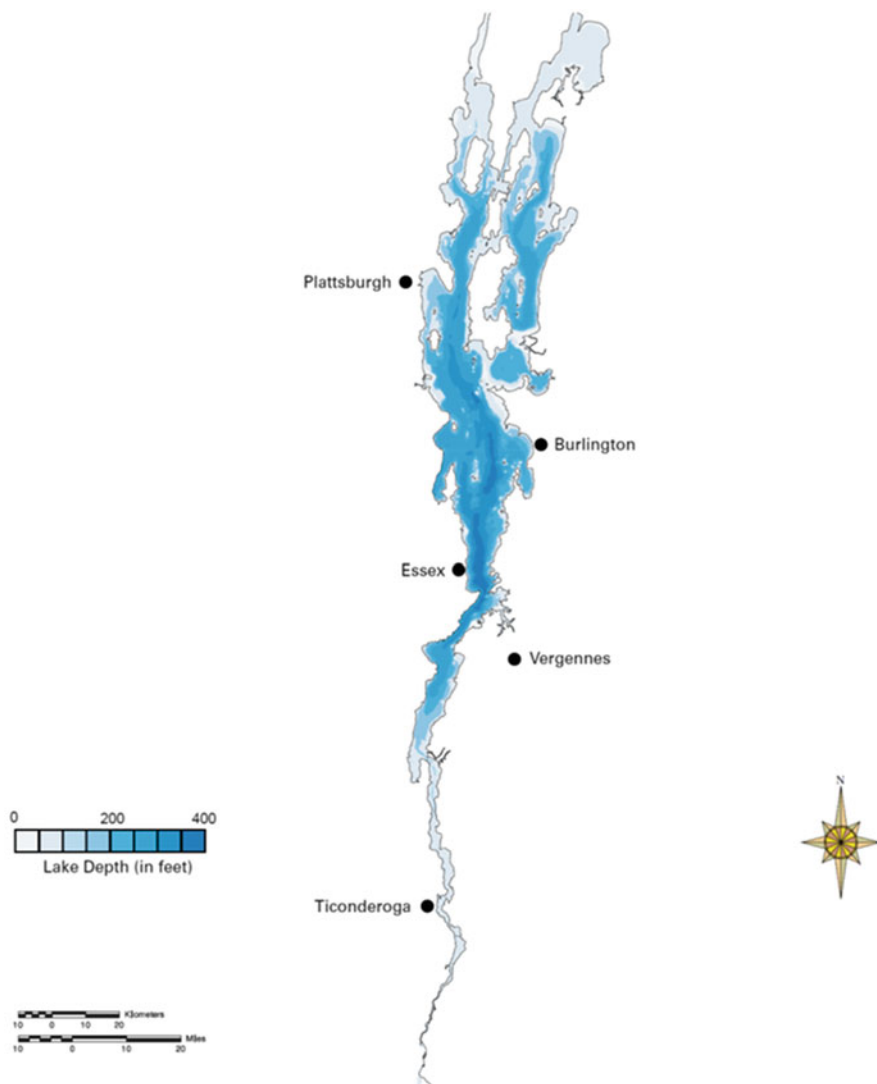
Synergies among invasive species, aquatic communities and habitats are highly complex and dynamic, but poorly understood (Strayer et al. 2006). Various approaches have been applied to forecast invasion and assess impacts of non-native species in freshwater lakes. Process-based models have been built using the modelling software Ecopath where the inclusion of *Dreissena* mussels in food web models enabled comparison of carbon flow and biomass before and after their invasion (Stewart and Sprules 2011). Individual based models have been developed to examine alterations of invasive mussel life cycles rising temperatures (Griebeler and Seitz 2007). Survival analysis and maximum likelihood techniques have been applied to estimate probabilities of lake invasion by invasive mussels (Leung et al. 2004).

This study applies CCA and HEA models to examine direct, indirect, and cascading effects between invasive zebra mussels and alewives, and the zooplankton community and water quality using data collected by the Lake Champlain Long-Term Monitoring program (LTM) from 1992 to 2015. General trends in water quality and zooplankton community dynamics have previously been discussed in Smeltzer et al. (2012) and Mihuc et al. (2012).

## 18.2 Case Study: Lake Champlain

The dimictic Lake Champlain (Fig. 18.1) is located between New York and Vermont (US) and Quebec (Canada). It has an estimated volume of 25.8 km<sup>3</sup>, a surface area of 1120 km<sup>2</sup>, a mean depth of 22 m and a drainage basin of 21,326 km<sup>2</sup>. Although primarily used as recreational lake, it also serves as a source of drinking water. However, runoff from agricultural and urban land uses in the drainage basin have resulted in eutrophication of the lake. The lake is divided into several distinct lake segments that differ in eutrophication levels, with the most Northern and Southern parts being meso-eutrophic, and central parts being meso-oligotrophic. Other major concern include the impact of the zebra mussel (*Dreissena polymorpha*) and alewife (*Alosa pseudoharengus*) on Lake Champlain, as well as recent arrivals of the spiny water flea, *Bythotrephes longimanus*.





**Fig. 18.1** Bathymetric map of Lake Champlain (1 m = 3.28 feet). From Lake Champlain Basin Atlas at: <http://atlas.lcbp.org/>

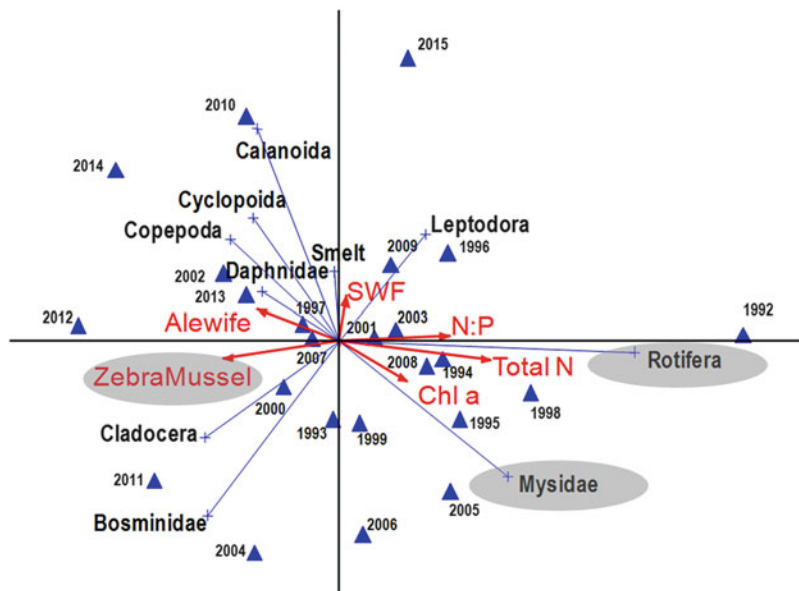
Lake water quality and zooplankton data were obtained from the Lake Champlain Long-term Monitoring program, and alewife data were acquired from the Vermont Dept. of Environmental Conservation. Data from the deep main lake from 1992 to 2015 have been averaged for July–August and are listed in Table 18.1.

**Table 18.1** Summary of annual water quality and biological data of Lake Champlain averaged for July–August from 1992 to 2015

Water quality variables	Mean/Min/Max	Biological variables	Mean/Min/Max
Secchi Depth (m)	5.2/3.3/6.8	Zebra Mussels (Ind m <sup>-3</sup> )	93300.7/0/ 350000
Total Nitrogen [Total N] (mg L <sup>-1</sup> )	407.1/325/516	Mysids (Ind m <sup>-3</sup> )	297.3/64/973.6
Total Phosphorus (mg L <sup>-1</sup> )	11.35/7.6/16.1	Cyclopoida (Ind m <sup>-3</sup> )	1615.5/127.4/ 3620.5
Chlorophyll-a [Chl a] (µg L <sup>-1</sup> )	3.8/1.7/6.1	Calanoida (Ind m <sup>-3</sup> )	883.9/18.4/ 2151.8
TN/TP [N:P]	36.97/23.3/51.7	Bosminidae (Ind m <sup>-3</sup> )	1875.1/145.5/ 6782.9
		Daphnia (Ind m <sup>-3</sup> )	955.3/36.8/ 3169.5
		Copepoda (Ind m <sup>-3</sup> )	2663.7/300.5/ 8427.2
		Leptodora (Ind m <sup>-3</sup> )	4.6/0.1/21.85
		Rotifers (Ind m <sup>-3</sup> )	3988.5/154.8/ 27582.4
		Spiny water flea SWF (Ind m <sup>-3</sup> )	1.3/0/28.4
		Alewife (CPUE)	49982.4/0/ 352136.8
		Smelt (CPUE)	259.6/59.9/746.3

### 18.3 Data Ordination by CCA

CCA has been applied to ordinate and display complex long-term monitoring data of Lake Champlain in two-dimensional space, and reveal qualitative relationships between the native zooplankton community, water quality, and invasive zebra mussels and alewives. Figure 18.2 displays CCA results for the biological and water quality data summarized in Table 18.1. The ordination results for water quality parameters and invasive species are plotted as red arrows that originate from the plot origin whereby the length of the arrows indicate the data variance of these variables. The location of plots for biological variables such as rotifers reflect their abundance. Locations for sample years are symbolized by blue triangles corresponding with shifts in the composition of the zooplankton community across the sample years. The ordination in Fig. 18.2 achieved a 51.2% correlation between water quality, invasive species and the zooplankton community. The results reflect long-term shifts in the zooplankton community such as the higher rotifer and mysid abundances shown in the right lower quadrant of the x–y axis for the early 1990s. Rotifer abundance was inversely related to zebra mussel abundance. Increases in alewife and zebra mussel densities in the left quadrants accompanied increases in Total N, N:P and Chl a in the right quadrants.



**Fig. 18.2** CCA analysis of the biological and water quality data summarized in Table 18.1. Results for water quality parameters and invasive species are plotted as red arrows; shaded taxa reflect strong negative relationships with zebra mussel abundance

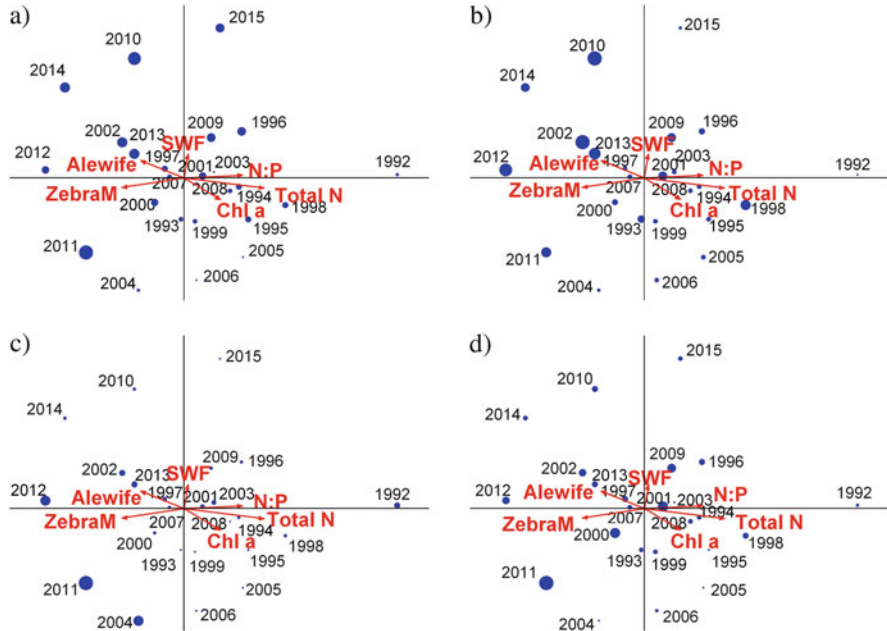
The copepod taxa cyclopoida and calanoida increased in abundance in response to the alewife invasion in the mid-2000s (Fig. 18.3a, b). Overall ordination results in Fig. 18.3 indicate long-term shifts not only in copepods but also in daphnic and bosminid crustaceans in response to increased zebra mussel and alewife abundance.

Results in Fig. 18.4a indicate a decline in mysid abundance in recent years compared to the late 1990s that coincided with steadily growing zebra mussel populations since the mid-1990s (Fig. 18.4c) and increasing alewife densities since the mid-2000s (Fig. 18.4b). These results correspond with findings by Ball et al. (2015).

CCA results suggest possible trophic interactions among zooplankton taxa, mysids, and zebra mussels in Lake Champlain that are further explored below by inferential modeling using HEA.

## 18.4 Inferential Modelling by HEA

HEA has been developed for inferential modelling of multivariate ecological data (Cao et al. 2014; Recknagel and Ostrovsky 2016). For more details about HEA please see Chap. 9. In this study, HEA is used to reveal and quantify direct and

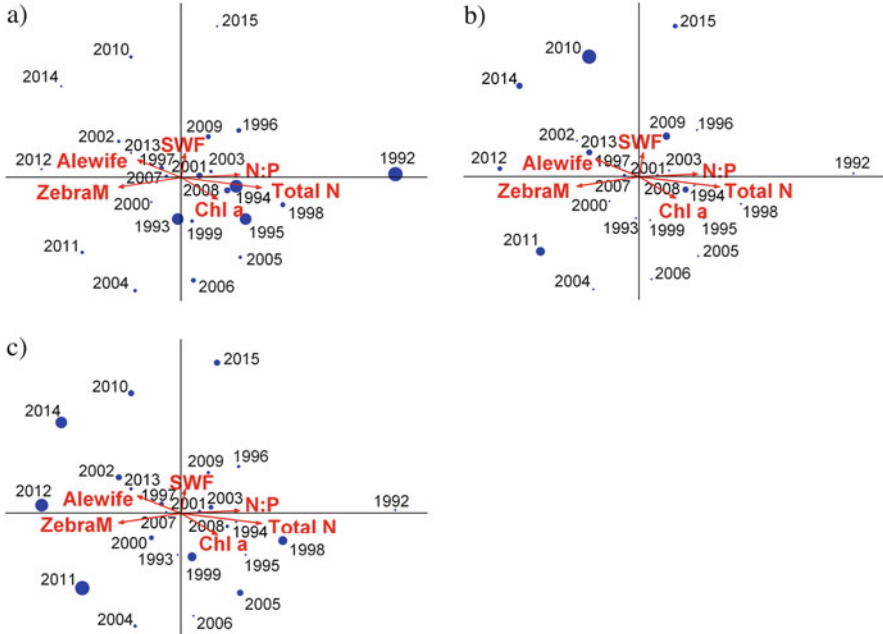


**Fig. 18.3** CCA related to: (a) Calanoids; (b) Cyclopoids; (c) Daphnids; (d) Bosminids. The size of the *blue circles* reflects increases in abundance; gradients for water quality and invasive species are illustrated by *red arrows*

indirect relationships among the zooplankton community, mysids and the invasive zebra mussel *Dreissena polymorpha* in Lake Champlain.

Rotifer population dynamics were related to grazing of chlorophyll-a, and predation by crustaceans and zebra mussels as illustrated in Fig. 18.5a. The IF-condition of the model in Fig. 18.5b indicates that the fastest decline in rotifer abundance is associated with increased numbers of zebra mussels above 395 individuals  $m^{-3}$ . The corresponding sensitivity function for the zebra mussels in Fig. 18.5c illustrates that rotifer density sharply decreases in response to increasing numbers of zebra mussels, most likely in response to the mussel's filtration of small bodied and slow moving rotifers. Sensitivity functions in Fig. 18.5c further reveal that there are no obvious predation effects on rotifers by growing numbers of calanoids and cyclopoids. Interestingly, rotifer and mysid population densities are positively associated.

The inferential model in Fig. 18.6b suggests that chlorophyll-a concentrations remained slightly higher when zebra mussel numbers were less than 110.781  $Ind\ m^{-3}$ . However, sensitivity results in Fig. 18.6c show a nearly neutral relationships between chlorophyll-a concentrations and zebra mussels. The sensitivity function for rotifers indicates that chlorophyll-a concentrations are slightly increasing with growing numbers of rotifers. This finding points at an indirect effect of zebra mussels since numbers of rotifers greater than 10,000  $Ind\ m^{-3}$  occurred only before



**Fig. 18.4** CCA related to: (a) Mysidae; (b) Alewife; and (c) Zebra mussels. The size of the blue circles reflects increases in abundance; gradients for water quality and invasive species are illustrated by red arrows

the arrival of zebra mussels. *Daphnia* seems to have the greatest grazing effects on chlorophyll-*a*, whereas *Bosmina* has no effect on chlorophyll-*a* concentrations.

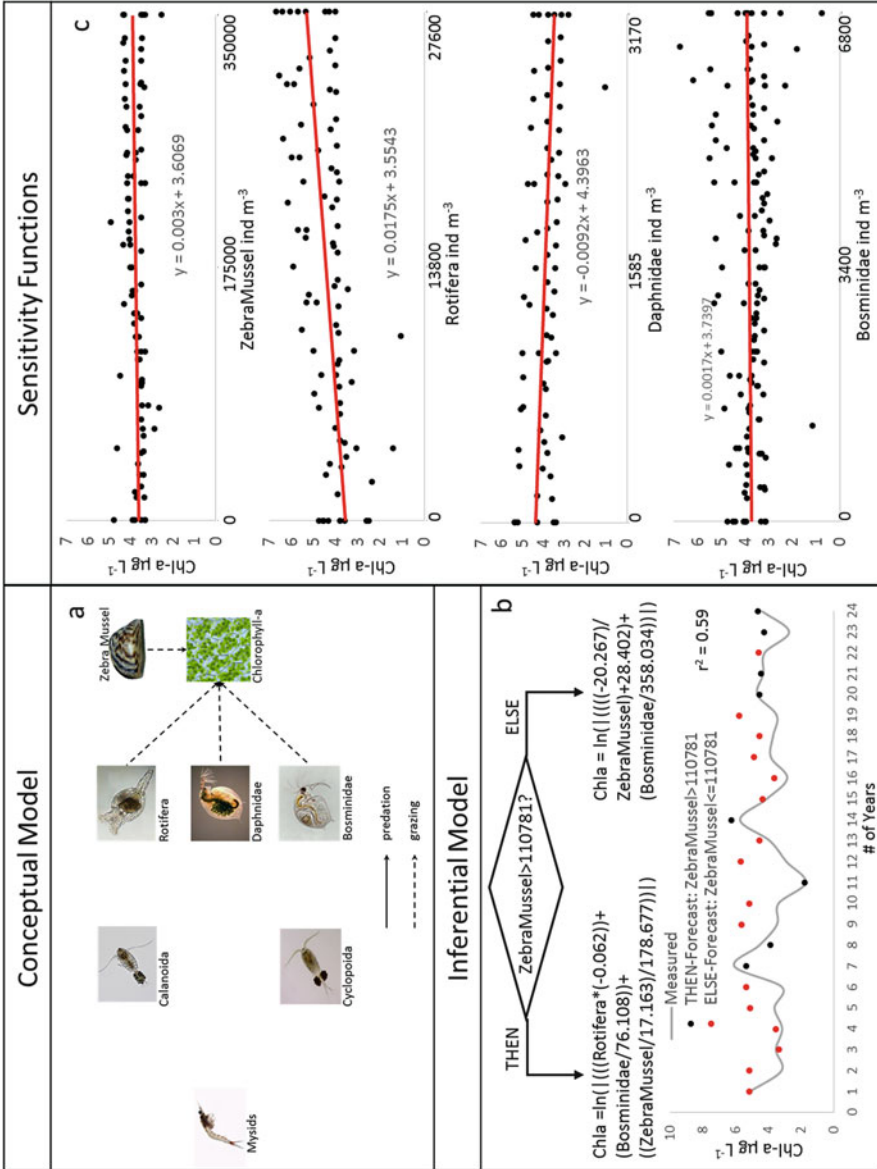
The model for daphnia presented in Fig. 18.7 indicates mutual relationships among daphnia, calanoids, mysids and cyclopoids (Fig. 18.7c). The IF-condition of the model in Fig. 18.7b also shows a range rather than a distinct threshold for mysids that would indicate the separation of high from low individual numbers of daphnia.

The bosmina model (Fig. 18.8b) identified a cyclopoid abundance of  $445 \text{ Ind m}^{-3}$  as a threshold below which high numbers of bosmina can be expected. However, sensitivity functions in Fig. 18.8c suggest a positive relationship with cyclopoida and neutral relationships with calanoids and mysids.

Results from HEA models (Figs. 18.5, 18.6, 18.7, and 18.8) illustrate direct and indirect effects of zebra mussels on zooplankton and mysids in Lake Champlain. Zebra mussels directly prey on small bodied rotifers (see Fig. 18.5b) causing secondary effects on mysids. The model in Fig. 18.6 also revealed direct effects of zebra mussels on chlorophyll-*a*, as well as the indirect effects on zooplankton community and mysids associated with zebra mussels feeding on rotifers.

Figure 18.9 demonstrates that small bodied cyclopoids increased in abundance following alewife invasion in 2005–06. These taxa most likely are able to avoid





**Fig. 18.6** Relationships of chlorophyll-a with zebra mussels, rotifers, daphnia and bosmina: (a) conceptual model; (b) inferential model; and (c) sensitivity functions

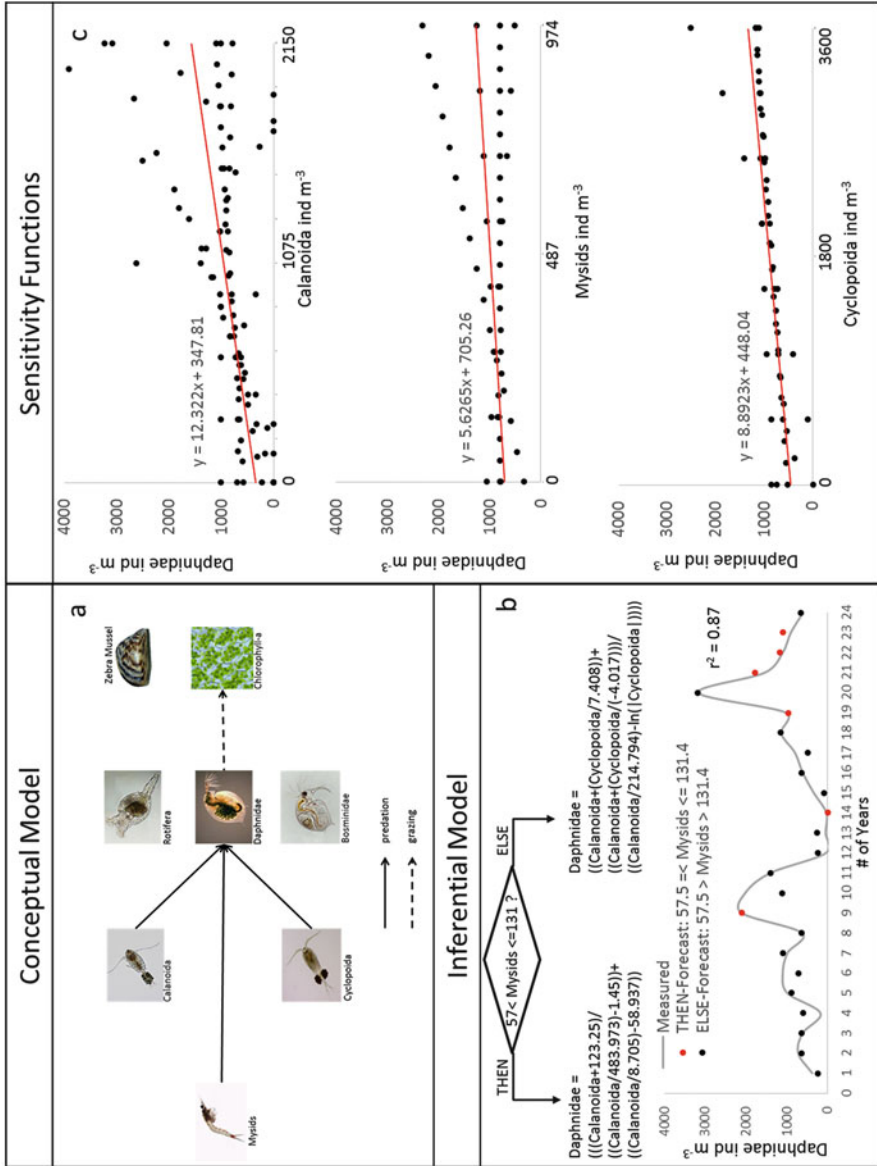
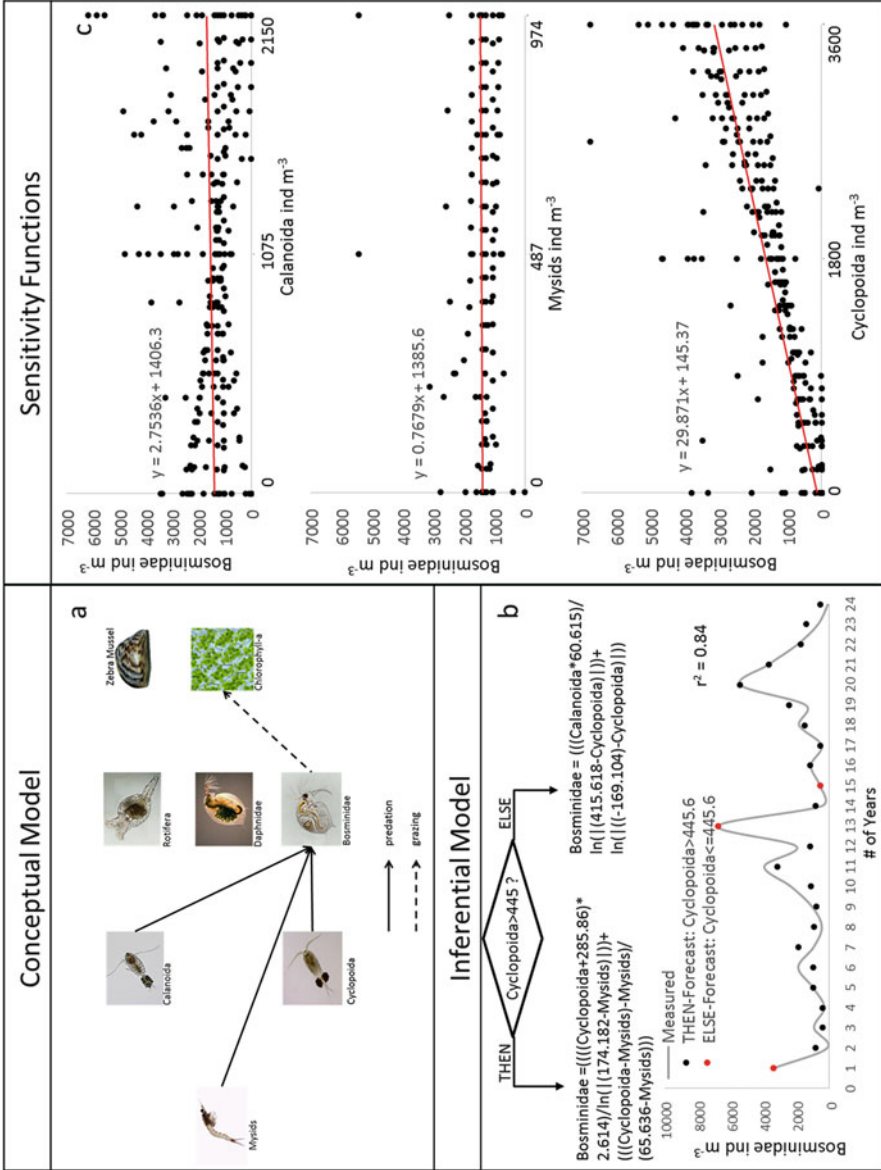
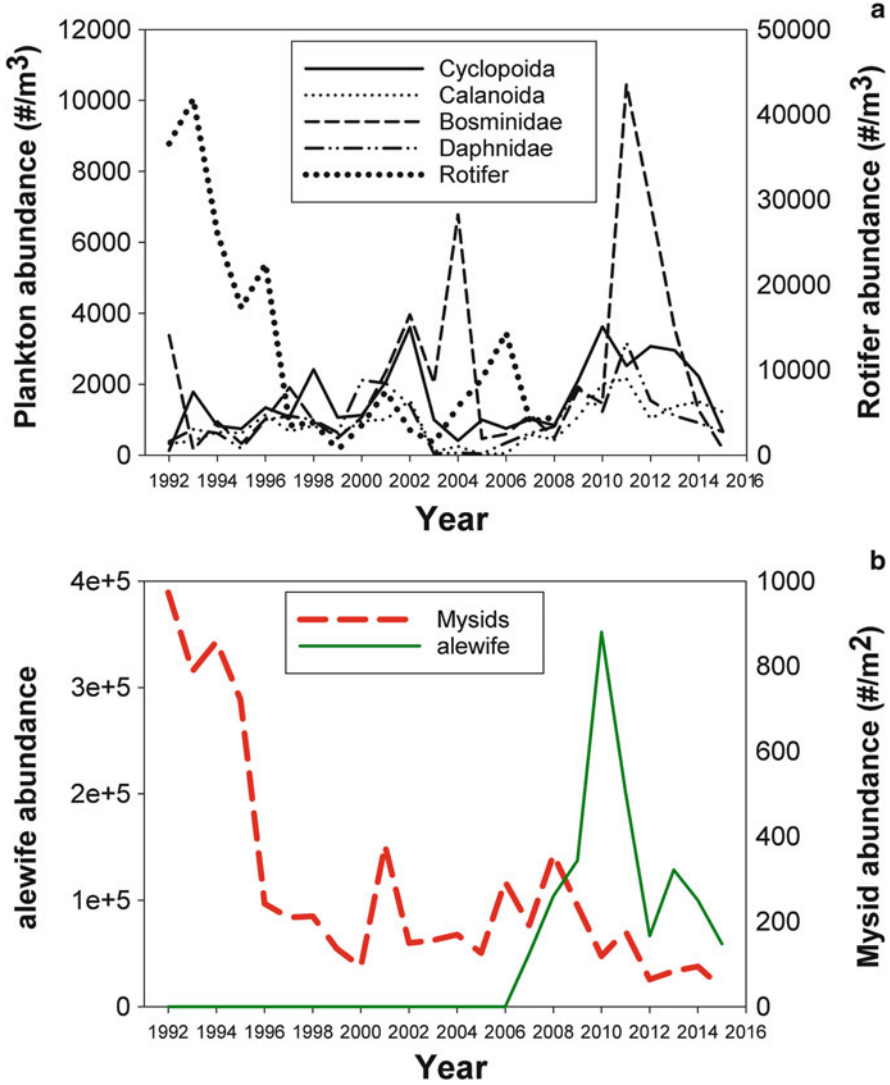


Fig. 18.7 Relationships of daphnia with copepods, calanoids and mysids: (a) conceptual model; (b) inferential model; and (c) sensitivity functions





**Fig. 18.8** Relationships of bosmina with copepods, calanoids and mysids: (a) conceptual model; (b) inferential model; (c) sensitivity functions



**Fig. 18.9** Long-term trends of native zooplankton (a) and mysids and alewife (b) in Lake Champlain from 1992 to 2015

predation by alewives, which selectively feed on larger bodied zooplankton. Small bodied bosmina and daphnia also seem to be less affected by alewives. The arrival of the spiny water flea *Bythotrephes longimanus* in 2014 may also have contributed to the recent declines in cladocerans and copepods in Lake Champlain.

## 18.5 Conclusions

This study demonstrated that analysis and modelling of long-term ecological data can assist in better understanding complex inter-relationships and possible future trends in lake ecosystems affected by invasive species that may become irreversible if disregarded.

Results of the study suggest that over time zebra mussels and alewives synergistically reshape trophic food webs of lakes by cascading positive and negative effects. Zebra mussels subdue directly rotifers causing indirect positive effects on crustaceans and copepods, and negative effects on mysids. Alewives entering a food web that has already been altered by zebra mussels, directly prey on mysids and copepods with further indirect positive effects on cladocerans. Since cladocerans are most efficient filter feeders, the question arises how abundant cladocerans influence the phytoplankton community and detritus concentrations in Lake Champlain. Whilst existing chlorophyll-a data displayed nearly steady-state conditions during the study period, this question can only be answered by future research including phytoplankton community data and modelling.

Findings from these studies may also improve public awareness of consequences of uncontrolled bioinvasion on natural lakes that over time will be worsened by catalyzing effects of global climate change (e.g. Bellard et al. 2016). These consequences not only cause ecological costs but moreover growing economic costs.

## References

- Ball SC, Mihuc TB, Myers LW, Stockwell JD (2015) Ten-fold decline in *Mysis diluviana* in Lake Champlain between 1975 and 2012. *J Great Lakes Res* 41:502–509
- Bellard C, Leroy B, Thuiller W et al (2016) Major drivers of invasion risks throughout the world. *Ecosphere* 7(3):1–14
- Cao H, Recknagel F, Orr PT (2014) Parameter optimisation algorithms for evolving rule models applied to freshwater ecosystem. *IEEE Trans Evol Comput* 18:793–806
- Connelly NA, O'Neil CR, Knuth BA, Brown TL (2007) Economic impacts of zebra mussels on drinking water treatment and electric power generation facilities. *Environ Manage* 40:105–112
- Crowl TA, Crist TO, Parmeter RR et al (2008) The spread of invasive species and infectious disease as drivers of ecosystem change. *Front Ecol Environ* 6(5):238–246
- Griebeler EM, Seitz A (2007) Effects of increasing temperatures on population dynamics of the zebra mussel *Dreissena polymorpha*: implications from an individual-based model. *Oecologia* 151:530–543
- Higgins SN, Vander Zanden MJ (2010) What a difference a species makes: a meta-analysis of dreissenid mussel impact on freshwater ecosystems. *Ecol Monogr* 80(2):179–196
- Leung B, Drake JM, Lodge DM (2004) Predicting invasions: propagule pressure and the gravity of allee effects. *Ecology* 85(6):1651–1660
- Mihuc TB, Dunlap F, Binggeli C et al (2012) Long-term patterns in Lake Champlain's zooplankton: 1992–2010. *J Great Lakes Res* 38:49–57
- Recknagel F, Ostrovsky I (2016) Inferential modelling of time series by evolutionary computation. In: Obrador B, Jones ID, Jennings E (eds) NETLAKE toolbox for the analysis of high-

- frequency data from lakes (Factsheet 11). Technical report. NETLAKE COST Action ES1201, pp 57–60. <http://eprints.dkit.ie/id/eprint/542>
- Smeltzer E, Shambaugh A, Stangel P (2012) Environmental change in Lake Champlain revealed by long-term monitoring. *J Great Lakes Res* 38:6–18
- Stewart TJ, Sprules G (2011) Carbon-based balanced trophic structure and flows in the offshore Lake Ontario food web before (1987–1991) and after (2001–2005) invasion-induced ecosystem change. *Ecol Model* 222:692–708
- Strayer DL, Eviner VT, Jeschke JM, Pace ML (2006) Understanding the long-term effects of species invasion. *Trends Ecol Evol* 21(11):645–651

# Chapter 19

## The Global Lake Ecological Observatory Network

Paul C. Hanson, Kathleen C. Weathers, Hilary A. Dugan,  
and Corinna Gries

**Abstract** This chapter explores a socio-technological (S-T) approach to information management within the Global Lake Ecological Observatory Network (GLEON). In S-T systems, information management, relevant organizational policies, and the supporting technologies are integral components of the network fabric. They derive from the needs of the community, articulated through representative governance, and they service the needs of the community by engaging data providers as partners in scientific endeavors. Through a brief history of GLEON, we recount the emergence of the S-T approach as part of GLEON's philosophy as a learning organization. It is clear that there is still much to be learned about streamlining data curation and publishing, especially from an international network of observatories with diverse data and sensor networks. Grass-roots networks such as GLEON often do not have the resources—human, financial, and infrastructure—required for persistent and highly efficient data curation and publishing. However, strategies that address directly the needs of the network community, such as providing credit to data providers, tracking the progress of projects that use the data, and sharing high-value synthesized data sets, quickly gain acceptance and garner commitment by the community. Today, S-T systems require ‘humans in the loop’ for data curation, which, in turn, results in constraints on scalability of these systems. One of the great challenges that lie ahead will be connecting GLEON S-T, which represents a diverse international community, with existing external data curation and archiving services.

---

P.C. Hanson (✉) • C. Gries

Center for Limnology, University of Wisconsin, Madison, WI, USA

e-mail: [pchanson@wisc.edu](mailto:pchanson@wisc.edu); [cgries@wisc.edu](mailto:cgries@wisc.edu)

K.C. Weathers

Cary Institute of Ecosystem Studies, Millbrook, NY, USA

e-mail: [weathersk@caryinstitute.org](mailto:weathersk@caryinstitute.org)

H.A. Dugan

Center for Limnology, University of Wisconsin, Madison, WI, USA

Cary Institute of Ecosystem Studies, Millbrook, NY, USA

e-mail: [hdugan@gmail.com](mailto:hdugan@gmail.com)

## 19.1 Introduction: A Brief History of GLEON

The Global Lake Ecological Observatory Network (GLEON; [www.gleon.org](http://www.gleon.org)) is an international, grassroots network of environmental and computer scientists, information technology experts, and, increasingly, citizens and artists, whose mission is to *conduct innovative science by sharing and interpreting high-resolution sensor data to understand, predict and communicate the role and response of lakes in a changing global environment* (Fig. 19.1; Weathers et al. 2013; Hanson et al. 2017). However, GLEON did not begin in 2005 with this mission. Rather, it began as a group of limnologists and information technology (IT) professionals who wished to build a persistent and scalable network of buoys in lakes around the world. GLEON's initial focus was on sensor observatories (data) and lakes (Fig. 19.2), but has since evolved into three networks: people, lakes and data (Weathers et al. 2013; Hanson et al. 2017). GLEON now has more than 500 members from more than 50 countries.

## 19.2 GLEON as Three Networks: People, Lakes and Data

GLEON is comprised of networks of people, lakes and data. The three networks are described below.

### 19.2.1 People

GLEON is largely a volunteer organization and relies on its membership to create products such as research projects, publications, code, models, applications, and educational/outreach materials. Members also elect and fill governance roles and create operational committees. In fact, the governance and operations structure of GLEON has played a critical role in its success as a network and provides the guidance for the development, evaluation, and implementation of science, technology, education and outreach initiatives. Understanding GLEON information management (IM) requires knowledge of the roles and responsibilities of its governance structure as well as an understanding of how the 'people network'—the engine of GLEON—operates (Fig. 19.3).

The future trajectory of scientific research has high uncertainty and, often, organizational structures and philosophical differences and similarities dictate the nature and methods of how science is accomplished (Uriarte et al. 2007; Eigenbrode et al. 2007). GLEON uses a socio-technological (S-T) systems approach to accomplish its goals. Here we define an S-T systems approach as 'a social system operating on a technical base' (Whitworth and Ahmad 2014). Our collaboration pays careful attention to the process and best practices of team





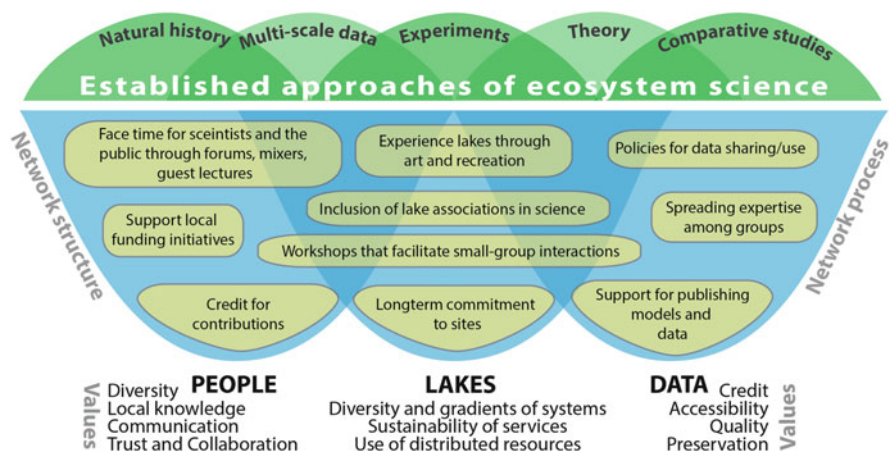


Fig. 19.2 Examples of GLEON buoys from around the world

science (e.g., Bennett et al. 2010; NRC 2015; Read et al. 2016), and we facilitate an organizational structure adapted to training and producing talented network scientists who can effectively navigate and shape the S-T interface and create innovative scientific products. One of the first activities of early GLEON members was to create Operating Principles and Procedures (OPP). This OPP not only set out the data sharing guidelines, but also created a governing structure, the Steering Committee.

An elected Steering Committee (SC), comprised of approximately 14 members from 10 different countries, provides leadership and vision for the organization. The Collaborative Climate Committee (CCC), also elected from the membership, helps guide the social environment of GLEON to the end of maximum engagement and empowerment of all members. The CCC is advisory to the GLEON SC and its Chair is a non-voting member of the SC. The GLEON Student Association (GSA)





**Fig. 19.3** GLEON is a network of people, lakes, and data, with a diversity of resources. Through network structure and process, GLEON is able to use its diversity of resources to address the five pillars of ecosystem science [adapted from Hanson et al. (2017)]

accounts for one-third of GLEON membership, plays a critical role in GLEON by organizing training programs and sessions within the annual All Hands’ meetings, co-running the Network Partnership Program, and helping to promote leadership opportunities for students throughout the network (Weathers et al. 2013). The GSA Chair is a non-voting member of the SC. Network-wide issues, such as best practices for the use of network data and priorities for the development of information management (IM) technologies, are relevant to each of these committees, and each committee, in turn, influences how important issues are addressed throughout GLEON operations.

### 19.2.2 Lakes and Lake Science

GLEON’s primary mission is to create and communicate knowledge about lake ecosystems. It operates principally through face-to-face and virtual meetings that are designed to initiate and accomplish research and synthesis projects. The team-forming models that GLEON has evolved toward fall into three categories: (1) initiatives that result directly from formal Working Groups; (2) ad hoc groups that form around ideas that do not fall neatly into the longer-term Working Group structure; and (3) partnerships—collaborations that have formed wholly or in part through professional networking at GLEON meetings.

GLEON has successfully demonstrated the value of this team-building approach through catalyzing multiple Working Groups organized under a diverse set of topics with contemporary relevance (Table 19.1). Via these working groups, data sets have been curated (e.g., Solomon et al. 2013) that utilize an unprecedented

**Table 19.1** GLEON Working Groups persist through multiple meetings and provide the infrastructure for discussions of science topics. Many GLEON product-oriented projects originate from Working Groups

Working group	Example project title
The theory group	Spring blitz
Reservoirs and lake management	State of the lakes survey
Citizen science	Lake observer mobile app
Lake physics and modeling	Parameter optimization techniques
Signal processing and fluorescence	DO and ChlA maximums across lakes
Lake metabolism	The age of carbon
Climate sentinels	Thermal responses to regional climate

collection of high-frequency sensor data from lakes worldwide. These data sets have been used to determine the relative importance of convection versus wind shear in lake mixing which controls CO<sub>2</sub> emissions from lakes (Read et al. 2012), predict ice-out dates (Pierson et al. 2011), examine short and longer-term effects of major storm events on lake ecosystem function (Jennings et al. 2012; Klug et al. 2012), and determine that lakes are warming at different rates around the globe (O'Reilly et al. 2015), all of which are major advances in knowledge that are direct or indirect results of GLEON network science and collaboration. Further, synthetic work stemming from GLEON provides insights to the future of harmful algal blooms in lakes (Brookes and Carey 2011), delivers best practices for studying lake metabolism (Staeher et al. 2010), and both demonstrates and articulates some of the future opportunities for microbial ecology using environmental sensors (Jones et al. 2008; Shade et al. 2009). Recently, large-scale data synthesis projects have led to data publications (Sharma et al. 2015) and the development of integrated databases (Soranno et al. 2015) with the potential to be extended internationally through GLEON collaborations. Collectively, these outcomes demonstrate how a network diverse in intellectual resources can capitalize on distributed data and information resources to address a diversity of science questions.

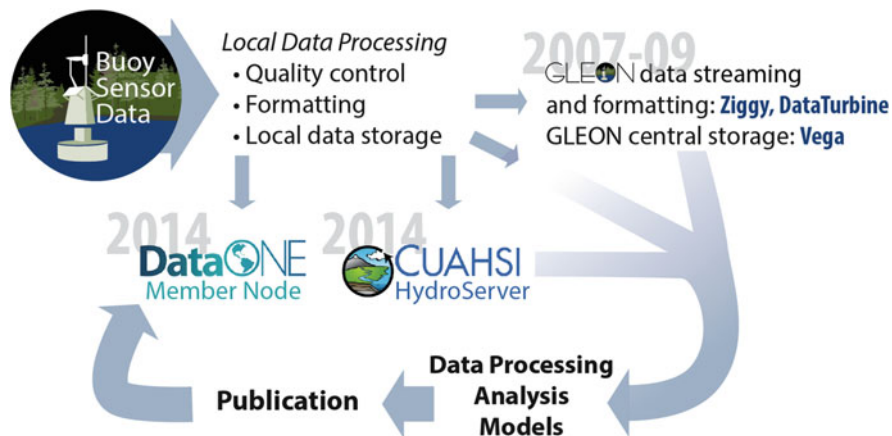
### 19.2.3 Data

All of these synthesis projects, and especially the ones on a larger scale (spatial, temporal, or data type) have advanced our experience and helped define successful, and unsuccessful, approaches and necessary skillsets in data management to support this type of research. The irony of these scientific and human network successes is that there is a major gap in our ability to discover, explore, and synthesize data on lake (and myriad other) ecosystems: *To date, data acquisition and harmonization is laborious and sometimes too difficult to accomplish, and it stands in the way of scientific advancement. The lack of an efficient way to collect and harmonize high frequency data is a significant bottleneck.*

### 19.3 GLEON’s Early History and Evolution

Following GLEON’s mission of ‘sharing and interpreting high resolution sensor data’, a vision for a cyberinfrastructure (CI) ecosystem was developed early on in GLEON’s history. Figure 19.4 illustrates that vision, and each component was realized with varying degrees of success. At the time, several CI research and development groups were faced with similar problems of streaming, harmonizing, storing, and making accessible large amounts of high frequency sensor data in near real time. Each sub group approached the same basic problem from their respective environmental, research and technical perspectives. However, it became clear that the goal of collecting and serving real time, high frequency sensor data was expensive, almost impossible to sustain (especially as a research project), and in need of extensive user input, not to mention being hampered by national and institutional data sharing (or rather data not sharing) policies. Hence, the initial GLEON CI vision was never accomplished. Here we recount each of the components of what should be a straightforward CI approach, its successes/failures, and lessons learned. We also discuss which of the components of our initial vision are still in use today and why.

Our initial vision was to have sensor data collected from buoys deployed around the world be pushed or pulled, near real time, into a central database and served out from there (again, near real time) (Fig. 19.4). As a grass-roots network, GLEON did not impose any specific requirements on the variables being collected or the hardware/software required to obtain and store data at the site-level. Each site funds its own research, including buoys, and its sampling infrastructure must service first the needs of the site-level funding source. As a consequence, GLEON is a network whose heterogeneity in data and local CI reflects the diversity of people, lakes, and cultures of the network. None-the-less, an end-to-end solution



**Fig. 19.4** An early vision of the GLEON cyber infrastructure ecosystem included fully integrated observational platforms (buoys), software for reformatting and formatting data for storage, and long-term storage in standardized repositories

for network-wide data streaming was, and is a terrific goal, but one that has not yet been accomplished due to the lack of major, distributed and sustained funding. The ideal progression, which we describe below, is to have data flow from sensors to the GLEON database. The GLEON database would then be open, and the data easily discovered, utilized in models and for visualizations, and harmonized with other data (e.g., watershed data and limnological data sampled by more traditional methods).

1. **From sensors to databases:** This is the realm of proprietary hardware and software of sensors and dataloggers, and rapidly changing telemetry technology. In remote areas, this remains problematic. For the purpose of formatting the datastream and submitting it to a database the software package ‘Ziggy’ was developed by GLEON members while DataTurbine (Open Source DataTurbine Initiative 2016) was available as an alternative.
2. **A data model:** The data model ‘Vega’, adapted from the Observations Data Model (ODM) developed by the Consortium of Universities for the Advancement of Hydrologic Science, Inc. (CUAHSI 2016) was developed, and data providers in GLEON agreed on a common vocabulary to describe measurement parameters and other metadata. Ziggy and DataTurbine were capable of formatting the data according to the Vega data model via specific configuration files for each data provider.
3. **Data discovery:** Originally a web interface was developed, which was later replaced by a desktop application, providing access and query capabilities into the harmonized data from >50 lakes in the GLEON network.
4. **Lake modeling:** The data were then available in a known standard format, and access and pre-processing could easily be automated before they were used in large-scale modeling.

The described components were developed, and several, mostly US lake research groups, contributed data for a short period of time at the beginning of GLEON (circa 2007–2009). However, supporting the development and management of the system became prohibitive, both because undergraduate and graduate students who helped in the development graduated, and because there were no specific resources to further develop and support the software. In addition, when these systems stopped being maintained, they were still in a developmental stage that rendered them too technical for most data contributors to use without major support from the developers. However, social and policy related issues of data sharing also contributed to its demise. Since then, commercial data logger software has improved dramatically, and several other groups were able to continue development [e.g., DataONE (2016), CUAHSI Water Data Center (2016), OPeNDAP (2016)] of more generally applicable solutions to data management. However, by now each GLEON member has developed a unique workflow and tools to manage the data locally complying with their respective institutional data sharing policy.

## 19.4 A World Café Approach

In 2011 the GLEON community re-assessed information management needs, opportunities and requirements. We took a *World Café* approach (The World Café Community Foundation 2016) to these issues of network-wide relevance, in which a large and somewhat intractable topic, such as IM for GLEON, is broken down into smaller, more tractable sub-issues, such as IM policy, IM tools, and IM training. Each tractable bit is a stop in the tour of cafés, and everyone at the meeting (up to 200 people) has the opportunity to contribute to each bit by spending 10–15 min at each café. After the tour, hosts from each café summarize the information, and a sub-committee is tasked with assembling the bits into a coherent message addressing the larger issue.

The World Café approach was used to justify and define the mission of the GLEON IT Task Force. An IT task force was formed in 2011 to respond to several information technology needs of GLEON, some of which were articulated and discussed at the World Café during the G13 meeting in Sunapee, New Hampshire (Table 19.2). The

**Table 19.2** The GLEON IT Task force was formed to make recommendations to the GLEON Steering Committee on changes to GLEON policy and process in support of technology use and development. The following are the task force’s five recommendations

- 
1. Team Data Sharing
    - (a) Sharing GLEON data, must be as open as possible, while respecting differences in international funding environments as they pertain to the data ownership and liability of the data-providing investigator.
    - (b) The GLEON data sharing policy must include a simple data access agreement.
    - (c) The GLEON data sharing policy must stipulate conditions for acceptable use (not-for-profit) and redistribution of data and metadata, and the citation verbiage.

---

  2. Team Attribution
    - (a) GLEON data should be linked to the organization(s) and individuals in the organization who played a key role in providing the data.
    - (b) Any errors discovered in the contributed data or improvements to the data must be reported back to the data provider.
    - (c) Any publication must be reported back to the providers of the data, who should be offered the opportunity to participate in any subsequent research.
    - (d) GLEON should investigate publishing GLEON data sets both to credit data providers and to provide a reference/citation for published data sets.

---

  3. Team Metadata
    - (a) We recommend GLEON adopt a metadata scheme describing different versions of the same data (such as raw, cleaned, gap-filled, QA/QC’d, modeled, etc.), similar to schemes employed by other organizations (e.g. NEON, CUAHHSI, and NetLake).

---

  4. Team Participation
    - (a) Points of contact within GLEON should be clearly identified so that sites have a GLEON resource person to contact with questions.
    - (b) Site data and metadata should be easily accessed on the GLEON website.

---

  5. Team Research Project Management
    - (a) Develop a project tracking system on the GLEON website that communicates the breadth of research conducted at GLEON, provides transparency and openness of the research process, and allows data providers to track and report on data use.
-

task force was divided into 5 teams according to topical areas. The primary objectives of the task force were to: (1) summarize the list of data provider and user needs, as begun at G13, and produce a set of requirements used to engage computer science (CS) colleagues in developing technology solutions; (2) make recommendations to the GLEON Steering Committee on changes to GLEON policy and process in support of technology use and development; (3) identify the most pressing needs for more reliable and usable sensor data for those sites that stream to Vega (Note: We mostly tabled this discussion point at G13. However, we have an opportunity to improve the current data streaming system, and our CS colleagues need our guidance); and (4) define the path forward for future oversight of GLEON IT.

By ‘technologies’, we mean technologies developed by/for GLEON in support of science (e.g., distributed computing, management of synthesized data sets), the human and data network (e.g., documenting participation in GLEON activities, providing credit for participation, linking participation to resources such as data, supporting data streaming and sensor network development), and dissemination of information about the organization (e.g., reporting on activities to GLEON membership, as well as outside entities). These are broad and complex categories of technologies. The Task Force addressed technology at two levels. At a high level, the task force sought to ensure that technology use and development is consistent with the organization’s goals and the members’ needs. At an operational level, the Task Force intended to make specific and targeted recommendations on technology, policy, and process that will improve GLEON, with the understanding that resources are limited and initiatives need to be prioritized.

Technology cannot be separated from policy and process, and so a concurrent goal was to synchronize GLEON’s policies and operations with technologies so that the three components can work most effectively in service of GLEON’s needs. We note that GLEON, as with technology, continually evolves and that it is important to build the flexibility for change into the S-T system; the resources in support of GLEON must continually adapt to the changing needs of GLEON.

Based on the recommendations of the IT Task Force, GLEON made two particularly notable advancements. The first was that the project management team implemented *GLEON’s Project Tracker* on the website, which now is used successfully to communicate research progress and allows data providers to track the use of their data, one of the main requirements identified by the IT task force. The second was that GLEON obtained a small planning grant awarded by the US NSF to *test information management approaches developed by other groups* in a series of workshops in 2012 and 2013. Specifically, a DataONE member node and a CUAHSI HydroServer were installed locally for testing purposes. Several GLEON data products are now available through the DataONE federation and the CUAHSI data center now hosts high frequency time series data from some US GLEON sites.

Meanwhile, a very successful, albeit non-automated, approach to large-scale research in GLEON is requesting data from the community that fulfill certain criteria relevant for the study, i.e., certain lake characteristics or certain parameters measured. Data providers may choose to participate and sometimes are required to provide significant data manipulations to format their data according to the project’s

standard, at which point data providers are frequently included in the author list of the resulting paper. Several of the resulting data products have been published in well-known repositories, such as DataONE via the Long Term Ecological Research (LTER) or GLEON data repositories (e.g., Sharma et al. 2015). This model was initiated by the early synthesis working groups at the National Center for Ecological Analysis and Synthesis (NCEAS 2016) and is very successful well beyond GLEON.

## 19.5 Lessons Learned

A decade of GLEON experience has rendered several lessons about information management in service of people, data, and lakes. These lessons include:

1. Invite more than data. Data can, and should be used, as an invitation to engage new collaborators and perspectives in the research enterprise. Rather than asking for data alone, we encourage GLEON scientists to ask for collaborators, with the recognition that there are many ways to contribute to productivity. This has been a crucial part of GLEON's ethos and operating principles.
2. Grassroots networks often generate highly heterogeneous data and manage similar data in equally heterogeneous and site-specific workflows. The collection of hardware and local IT infrastructure is thus also dizzyingly heterogeneous across sites. In addition, each site moves forward in its technological evolution and capacity at its own pace. While this diversity has inherent inefficiencies, it provides scientific advantages because the network has measurements relevant to almost any aquatic ecology topic, and therefore, scientific response can be rapid and adaptive. The diversity and asynchronous advancement also means that every site has the potential to be a leader in a particular area because of their unique design.
3. Harmonizing heterogeneous data into a convenient data product for re-use puts a large burden on the data providers. Once a data site (by site, we mean a functional unit that has responsibility and authority to collect, store, and distribute data) has developed an acceptable IM workflow, adding the step of providing the data in a community approved standard format is currently not providing enough incentive for the data provider to justify the extra effort.
4. Traditional funding streams for data repositories largely follow national borders and can not provide a comprehensive solution for an international network such as GLEON. This places international grass-roots networks in the untenable position of not having the internal resources to provide full IM services while at the same time not being able to engage, e.g., federal agencies, in the support of research activities from foreign countries.
5. Accordingly, international grassroots networks will have to organize as networks of independent repositories that agree on standard interfaces for data exchange which may or may not involve standard data formats. However,

extensive investments are necessary to develop and maintain needed infrastructure even if this system of systems consists only of a registry for resources (data and tools) and search capabilities without providing central IM services and data storage. Examples are the DataONE Federation (DataONE 2016) and the Group on Earth Observation (GEO 2016). Both projects have received national funding to develop metadata clearing houses that also serve international partners.

6. The S-T model used by GLEON has, in fact, provided an ideal platform for accelerating technology developments in the area of models and analytical code widely used in limnology. In contrast to data management, these models and code are considered research products, and represent contributions to advancing science. Extensive training in the use of these technologies and direct input of users to improve them has contributed to their great success. Further, these technologies have been released as open source which will dramatically increase the speed and efficiency with which model resources enter collaborative space, thereby catalyzing network science.
7. In true S-T spirit, GLEON has found that engaging with citizen scientists to further co-develop and utilize network science products enhances both the technology development as well as the use of those technologies, i.e., the developers and users are part of the same community (e.g., the LakeObserver app, [www.lakeobserver.org](http://www.lakeobserver.org), Cary Institute of Ecosystem Studies 2016).
8. The curation of resources (models and data) for publication, discovery, and delivery should be driven by research questions, the need for increased transparency and repeatability in science and the preservation of valuable, sometimes even irreplaceable information. However, the magnitude of the data discovery problem is still immense and may only be overcome by prioritization according to current science, education, and outreach needs, and by providing value to all participants in metrics (data as publications and data usage statistics).
9. As a community, we need to fully validate data publishing as a scientific contribution and completely integrate it into the scientific process, improve documentation and exposure of data manipulation procedures to enable reuse of data and tools by others, and embrace the practice of publishing models as open resources for community development and reuse.
10. We must enable the community of scientists, students and citizens focused on aquatic systems to develop networked team science strategies and infrastructure that can be transferred to other science domains and conversely encourage the community to re-use developments in other science domains.
11. It is essential to track progress at the project scale so that all who are involved, including data providers, project personnel, and the network community, have information on the state of the project. This provides the community with knowledge about which projects are underway and in many cases identifies opportunities for anyone in the community to contribute to those projects. Project tracking also provides information for reporting progress to funding providers.



12. At its inception, GLEON founders and shapers did not immediately recognize the importance of the ‘social’ part of IM; we now know that GLEON functions as an S-T system. In some ways, the story of GLEON is an empirical story—through trial and error and gathering feedback, a community of scientists has evolved from a product-centric group—with technologies and manuscripts the initial *raison d’être*, to a people-centric community, with additional foci of collaboration and team building compelled by a common need for discovery and innovation (Weathers et al. 2013; Hanson et al. 2017).

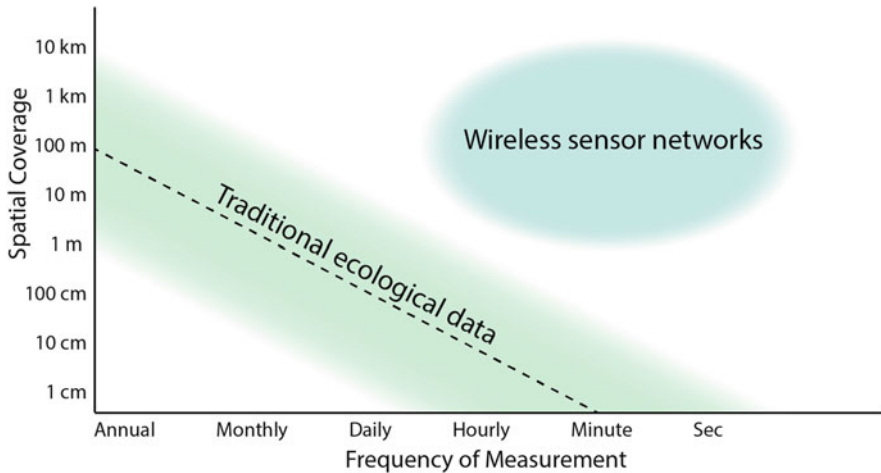
## 19.6 A Vision for the Future of GLEON Information Management

The vision for the future of GLEON information management focuses on three primary activities: (1) advancing socio-technical systems; (2) advancing team science; and (3) growing capacity at lake observatories.

### 19.6.1 *Socio-technological Systems: The Need and the Vision*

Innovative research and scientific discoveries are increasingly being accomplished by collaborative teams (Cheruvilil et al. 2014; Guimerà et al. 2005; Wüchty et al. 2007)—i.e., by diverse networks of people who are able to discover, access, manage, and synthesize ‘big data’ from observations that span the globe and at frequencies that range from milliseconds to months (LaDeau et al. 2017). Data at large spatial and temporal scales and integrated from different disciplines are needed to interpret, forecast, and manage ecosystems (e.g., lakes and reservoirs) and ecosystem functions and services, such as ecosystem energy exchanges and clean and plentiful drinking water, especially in a rapidly changing environment (Fig. 19.5). In recognition of these needs, networks such as GLEON are forming to support new science and new approaches to the conduct of science. These new teams need to include and fully integrate domain scientists as well as data managers, technologists, programmers, and computer scientists to rise up to the challenges.

Network science is increasingly being shown to lead to innovative research (NRC 2015). Along with the era of ‘Big Data’, there is growing recognition of the need for interdisciplinary science and more extensive collaboration that crosses institutional, geographic and political, as well as technology and domain science boundaries (Wüchty et al. 2007; Cheruvilil et al. 2014; Uriarte et al. 2007). Confronting complex ecosystem problems, such as how food, water, and energy systems are responding to global change, using big data requires a great diversity of skills, especially, a strong suite of social skills necessary to harness communities’

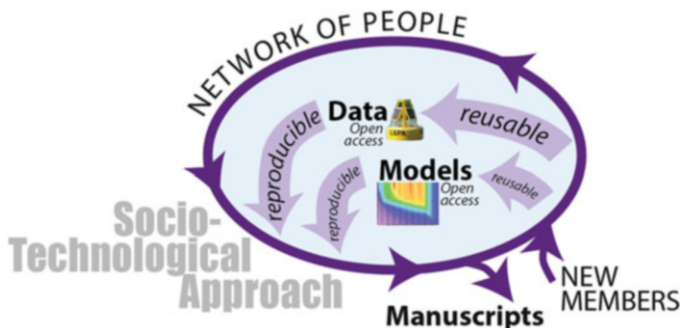


**Fig. 19.5** Data representing new space-time scales for science were part of the initial vision for GLEON (Porter et al. 2005). These data remain an important resource for network-level research

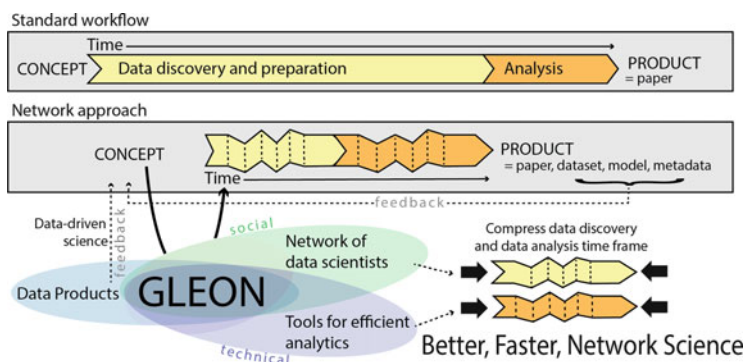
resources, inspire creativity, and communicate across domain boundaries (Porter et al. 2011; Cheruvelil et al. 2014). These skills, as well as the training necessary to develop them, are difficult to learn from any one institution or research group but can be a key attribute of the cumulative expertise and knowledge of a cohesive, functional research network (Weathers et al. 2013).

The current overarching goal of IM within GLEON is to grow a socio-technical (S-T) system (here defined as ‘a social system operating on a technical base’, Whitworth and Ahmad 2014) that captures the processes of assembling technology and human resources for a given science problem, and enables their reuse so that science problems can be solved more efficiently through collaboration (Fig. 19.6). Rather than develop new technology components, per se, we envision a ‘systems redesign’. In our experience, progress on systems-level CI has been impeded by the usually unstated assumption that *humans in the loop are barriers to be removed*. In contrast, we postulate that attention to S-T features are needed to overcome these obstacles, accelerate scientific inquiry, and to catalyze new networks of people (Fig. 19.7). In an S-T system, technology is adapted to human needs and humans are well trained to effectively use available technology and the human networks that result. Furthermore, community vetted guidelines or policies governing the system, community trust, platforms for developing collaborations, and human interactions underpin all aspects of S-T systems and are thus vitally important to system function.

Even with well-functioning data, model, and script archives in place today, innovative ideas for data syntheses, including data discovery and data access are ignited by personal interaction and communication (Fecher et al. 2015; Tenopir et al. 2011; Hogan and Weathers 2003)—through the *human network*. A series of human network events is necessary before data harmonization can start, including



**Fig. 19.6** In an S-T system, people are an integral part of the information management system. Teams both produce and consume data and are involved in the creation of technologies and models used in the iterative process of science



**Fig. 19.7** GLEON’s goal for a better, faster, network science in which the time from concept to product is shortened through tight integration of network resources, including GLEON’s human resources

building trust and establishing collaborative groups, which in highly functioning network teams must be followed by negotiating the rules for exchange and use of resources, and acknowledgement of those who contributed (e.g., Cheruvilil et al. 2014).

In order to proceed toward data preparation and analysis, it is critical to understand data structures and vocabularies, QA/QC and aggregation of raw data plus reformatting to meet specific requirements of the analysis or modeling process. Although there are many common patterns in these data manipulation processes across observatories and collaborative groups, they are usually handled in a one-off approach, with research teams implementing them over and over again in slightly different ways, depending on script writing prowess. We have found that having experts negotiate the interaction between data consumers and providers improves the overall experience and that these experts are more adept at writing reusable data

manipulation scripts, potentially expediting the research process. Hence, we assert that future IM should have a people-network-guiding triumvirate: Data Science, Governance, and Community Development Teams to guide the research process of Eco-Science teams. Through this human component of S-T, we will attain better, more creative products, publications, and perspectives.

Future work in GLEON will go far beyond establishing and making available technology to support the data lifecycle (data collection, curation, storage, discovery, access, integration, analysis, and publication). It has been recognized that successful data curation as well as data preparation for synthesis research requires a specialized skillset (Hernandez et al. 2012). However, this skillset can efficiently be applied to data curation and basic data manipulation of a wide range of environmental science disciplines. Hence, we hypothesize that a data curation and preparation center can accelerate scientific inquiry by shifting the ratio of time spent on data discovery and preparation vs. data analysis, which is currently cited to be as much as 80% vs. 20%, respectively, for synthesis research (Lohr 2014). Clearly these are estimates from ‘one-off’ research projects and efficiency can be achieved through advanced training/specialization of the workforce (data professionals), tool reuse, and understanding and implementation of advanced semantics and standards during data and tool curation.

### ***19.6.2 Advancing Team Science***

The network we seek to develop will evolve through time and be multilayered, connecting nodes of different types (e.g., models, data). A foundation layer will be the people network (Weathers et al. 2013; Read et al. 2016). Understood as a social network, GLEON will facilitate research interactions among scientists and citizens, focused on aquatic ecosystems. From one perspective, then, the activities of GLEON can be seen as a contribution to the facilitative mode of S-T integration (Fisher et al. 2015). As the research community focused on aquatic ecosystems is large and distributed, network facilitation must coordinate and integrate these interactions if it is to be efficient and effective, and this requires respecting the differences manifest across the community. These include conceptual, methodological, and professional differences between disciplines (O’Rourke and Crowley 2013), institutional boundaries that separate universities and other research units, and cultural and legal boundaries that mark relevant transitions between countries.

### ***19.6.3 Growing Capacity at Lake Observatories***

GLEON plans to continue to work closely with lake and reservoir observatories as well as citizen scientist groups that lead lake associations and management groups. Our goal is to provide training options to assist in making their data more accessible

and discoverable, which will also help in including these groups and their resources in collaborative efforts. Moreover, teaching observatories how to use community tools for data QA/QC, derived data products, and data discovery and access as part of the scientific process, i.e., ‘teach the teachers’ is an important aspiration. This has, for the last decade since the inception of GLEON, been identified as a critical, lacking set of resources by GLEON site members (i.e., observatories). Without these tools, many buoy data languish on the hard drives of observatory facilities. Finally, all of the data, models, workflows, and metadata need to be available to the community, enabling observatories to highlight the contributions they have made to broader science, education, and outreach efforts, as well as the policies regarding the use of resources in network science.

## 19.7 Conclusions

GLEON, as a socio-technical (S-T) system formed from an international community of scientists and a network of lake observatories, has made tremendous progress in data exchange among its members over the first 10 years of its existence. While an S-T system is a different and surprising outcome from our initial vision of building the cyberinfrastructure to accept, handle, and serve high frequency data streaming from lakes around the world, it has proven effective in supporting more than 100 publications and data products. GLEON has supported S-T development through inter personal trust building, developing mechanisms of attributing credit, and providing the platforms—both social and technical—for interdisciplinary collaboration.

Data management, however, remains primarily a manual exercise and happens at each observatory. In support of GLEON network science, data have also been collected from multiple sites and curated manually to answer specific research questions. These data products have proven to be high value and have supported multiple publications through re-use. Because data management has manual components, and because workflows in support of science tend to be ad hoc, GLEON scientists reinvent the process of data gathering, data cleaning, and data harmonization. Although each research question requires slightly different approaches to these steps, some generalizations are possible. Our experience suggests that the skill set of a data specialist could make the research process more efficient and accelerate scientific inquiry through standardization and tool reuse.

While we have made important strides within GLEON over the past decade, it is clear that a better system of systems for data discovery and access is needed in an approach that continues to honor local skills, policies, and requirements. The current approach of querying the community of data producers for data is successful but will always miss data that are not offered due to lack of time, resources, or interest at the local observatory. Collaborations with organizations outside of GLEON that have long-term sustainability and the human and technological

resources to tackle big data management problems may be needed to realize the goals of easily discoverable and usable data.

## References

- Bennett LM, Gadlin H, Levine-Finley S (2010) Collaboration and team science: a field guide. National Institutes of Health Publication No. 10-7660, National Institutes of Health, Bethesda. [https://ccrod.cancer.gov/confluence/download/attachments/47284665/TeamScience\\_FieldGuide.pdf?version=2&modificationDate=1285330231523&api=v2](https://ccrod.cancer.gov/confluence/download/attachments/47284665/TeamScience_FieldGuide.pdf?version=2&modificationDate=1285330231523&api=v2)
- Brookes JD, Carey CC (2011) Resilience to blooms. *Science* 334(6052):46–47
- Cary Institute of Ecosystem Studies (2016) Lake Observer: a mobile app for recording lake and water observations. <https://www.lakeobserver.org>. Accessed 18 Aug 2016
- Cheruvilil KS, Soranno PA, Weathers KC et al (2014) Creating and maintaining high-performing collaborative research teams: the importance of diversity and interpersonal skills. *Front Ecol Environ* 12:31–38
- Consortium of Universities for the Advancement of Hydrologic Science, Inc. (CUAHSI) (2016) ODM Databases. <http://his.cuahsi.org/odmdatabases.html>. Accessed 18 Aug 2016
- Consortium of Universities for the Advancement of Hydrologic Science, Inc. Water Data Center (CUAHSI Water Data Center) (2016) The Water Data Center. <https://www.cuahsi.org/wdc>. Accessed 18 Aug 2016
- DataONE (2016) DataONE: Data Observation Network for Earth. <https://www.dataone.org/>. Accessed 18 Aug 2016
- Eigenbrode SD, O'Rourke MR, Wulforst JD et al (2007) Employing philosophical dialogue in collaborative science. *BioScience* 57:55–64
- Fecher B, Friesike S, Hebing M (2015) What drives academic data sharing? *PloS One* 10(2): e0118053. doi:10.1371/journal.pone.0118053
- Fisher E, O'Rourke M, Evans R et al (2015) Mapping the integrative field: taking stock of socio-technical collaborations. *J Respons Innov* 2(1):39–61
- GEO (2016) GEO: Group on Earth Observations. <https://www.earthobservations.org/index.php>. Accessed 18 Aug 2016
- Guimerà R, Uzzi B, Spiro J et al (2005) Team assembly mechanisms determine collaboration network structure and team performance. *Science* 308:697–702
- Hanson PC, Weathers KC, Kratz TK (2017) Networked lake science: how the Global Lake Ecological Observatory Network (GLEON) works to understand, predict and communicate lake ecosystem responses to global change. *Inland Waters* 4:543–554
- Hernandez RR, Mayernik MS, Murphy-mariscal ML et al (2012) Advanced technologies and data management practices in environmental science: lessons from academia. *BioScience* 62(12): 1067–1076
- Hogan K, Weathers KC (2003) Psychological and ecological perspectives on the development of systems thinking. In: Berkowitz AR, Nilon CH, Hollweg KS (eds) *Understanding urban ecosystems: a new frontier for science and education*. Springer, New York, pp 233–260
- Jennings E, Jones S, Arvola L et al (2012) Episodic events in lakes: an analysis of drivers, effects, and responses using high frequency data. *Freshw Biol* 57:589–601
- Jones SE, Chiu CY, Kratz TK et al (2008) Typhoons initiate predictable change in aquatic bacterial communities. *Limnol Oceanogr* 53:1319–1326
- Klug JL, Richardson DC, Ewing HA et al (2012) Ecosystem effects of a tropical cyclone on a network of lakes in NE North America. *Environ Sci Technol* 46:11693–11701
- LaDeau SL, Han BA, Rosi EJ, Weathers KC (2017) The next decade of big data in ecosystem science. *Ecosystems* 20(2767):274–283

- Lohr S (2014) For big-data scientists, 'janitor work' is key hurdle to insights. *New York Times*, 17 Aug 2014. [http://www.nytimes.com/2014/08/18/technology/for-big-data-scientists-hurdle-to-insights-is-janitor-work.html?\\_r=0](http://www.nytimes.com/2014/08/18/technology/for-big-data-scientists-hurdle-to-insights-is-janitor-work.html?_r=0)
- National Research Council (NRC) (2015) Enhancing the effectiveness of team science. In: Committee on the Science of Team Science, Cooke NJ, Hilton ML (eds) Board on behavioral, cognitive, and sensory sciences, division of behavioral and social sciences and education. The National Academies Press, Washington, DC
- NCEAS (2016) NCEAS: National Center for Ecological Analysis and Synthesis. <https://www.nceas.ucsb.edu>. Accessed 18 Aug 2016
- O'Reilly CM, Sharma S, Gray DK et al (2015) Rapid and highly variable warming of lake surface waters around the globe. *Geophys Res Lett* 42:10773–10781
- O'Rourke M, Crowley SJ (2013) Philosophical intervention and cross-disciplinary science: the story of the Toolbox Project. *Synthese* 190:1937–1954
- Open Source DataTurbine Initiative (2016) DataTurbine. <http://dataturbine.org>. Accessed 18 Aug 2016
- OPeNDAP (2016) OPeNDAP. <http://opendap.org>. Accessed 18 Aug 2016
- Pierson DC, Weyhenmeyer GA, Arvola L et al (2011) An automated method to monitor lake ice phenology. *Limnol Oceanogr Methods* 9:74–83
- Porter J, Arzberger P, Braun H-W, Bryant P, Gage S, Hansen T et al (2005) Wireless sensor networks for ecology. *BioScience* 55(7):561. doi:10.1641/0006-3568(2005)055[0561:WSNFE]2.0.CO;2
- Porter JH, Hanson PC, Lin CC (2011) Staying afloat in the sensor data deluge. *TREE* 1484:1–9
- Read JS, Hamilton DP, Desai AR et al (2012) Lake size dependency of wind shear and convection as controls on gas exchange. *Geophys Res Lett* 39:L09405. doi:10.1029/2012GL051886
- Read EK, O'Rourke M, Hong GS et al (2016) Building the team for team science. *Ecosphere* 7(3):e01291. doi:10.1002/ecs2.1291
- Shade A, Carey CC, Kara E et al (2009) Can the black box be cracked? The augmentation of microbial ecology by high-resolution, automated sensing instruments. *ISME J* 3:881–888
- Sharma S, Gray D, Read J et al (2015) A global database of lake surface temperatures collected by in situ and satellite methods from 1985–2009. *Sci Data* 2:150008. doi:10.1038/sdata.2015.8
- Solomon CT, Bruesewitz DA, Richardson DC et al (2013) Ecosystem respiration: drivers of daily variability and background respiration in lakes around the globe. *Limnol Oceanogr* 58:849–866
- Soranno PA, Bissell EG, Cheruvilil KS et al (2015) Building a multi-scaled geospatial temporal ecology database from disparate data sources: Fostering open science through data reuse. *GigaScience* 4:28. doi:10.1186/s13742-015-0067-4
- Staehr PA, Bade D, Van de Bogert MC et al (2010) Lake metabolism and the diel oxygen technique: state of the science. *Limnol Oceanogr Methods* 8:628–644
- Tenopir C, Allard S, Douglass K et al (2011) Data sharing by scientists: practices and perceptions. *PLoS One* 6(6):e21101. doi:10.1371/journal.pone.0021101
- The World Café Community Foundation (2016) The World Café. <http://www.theworldcafe.com>. Accessed 18 Aug 2016
- Uriarte M, Ewing HA, Eviner VT, Weathers KC (2007) Scientific culture, diversity and society: suggestions for the development and adoption of a broader value system in science. *BioScience* 57:71–78
- Weathers KC, Hanson PC, Arzberger P et al (2013) The Global Lake Ecological Observatory Network (GLEON): the evolution of grassroots network science. *Bulletin of Limnol Oceanogr* 22:71–73
- Whitworth B, Ahmad A (2014) Socio-technical system design. In: Soegaard M, Dam RF (eds) The encyclopedia of human-computer interaction, 2nd edn. The Interaction Design Foundation, Aarhus, Denmark. [https://www.interaction-design.org/encyclopedia/socio-technical\\_system\\_design.html](https://www.interaction-design.org/encyclopedia/socio-technical_system_design.html)
- Wüchty S, Jones BF, Uzzi B (2007) The increasing dominance of teams in production of knowledge. *Science* 316:1036–1039

# Chapter 20

## Long-Term Ecological Research in the Nakdong River: Application of Ecological Informatics to Harmful Algal Blooms

Dong-Gyun Hong, Kwang-Seuk Jeong, Dong-Kyun Kim, and Gea-Jae Joo

**Abstract** In recent decades, the importance of long-term ecological research (LTER) has been highlighted because of the growing interest in global environmental changes. Specifically, LTER data allows one to track the history of target ecosystems (e.g., trends of particular ecological entities) and enables one to understand the causal relationships of ecosystem functioning. One ecological problem is harmful algal blooms (HABs) in freshwater environments. It is generally perceived that global warming and local eutrophication are responsible for serious and frequent HAB events, and various efforts have been made to explain and forecast HABs. LTER data for HABs typically consist of various forcing functions and variables; thus, the selection of appropriate data-analysis methods for a HAB database is necessary. This chapter presents a series of studies related to the prediction and elucidation of two HABs, such as summer cyanobacteria (e.g., *Microcystis aeruginosa*) and winter diatom (e.g., *Stephanodiscus hantzschii*) that occur in the regulated Nakdong River, South Korea. First, HABs, water quality, and zooplankton patterns were analyzed using self-organizing maps (SOMs). Those major factors that have a close relationship to HABs, i.e., water temperature, pH, and rainfall, were selected. We created a predictive model and control scenario for HABs using a variety of methods (evolutionary computation, artificial neural network) in the real world based on confirmed information. We also suggest potential further studies of the Nakdong River.

---

D.-G. Hong • G.-J. Joo  
Department of Biological Sciences, Pusan National University, Busan 46241,  
Republic of Korea  
e-mail: [hong0728@hanmail.net](mailto:hong0728@hanmail.net); [gjoo@pusan.ac.kr](mailto:gjjoo@pusan.ac.kr)

K.-S. Jeong (✉)  
School of Public Health, Dongju Colledge, Busan 49318, Republic of Korea  
e-mail: [kknd.ecoinfo@gmail.com](mailto:kknd.ecoinfo@gmail.com)

D.-K. Kim  
Department of Physical and Environmental Sciences, University of Toronto, Toronto, ON,  
Canada, M1C1A4  
e-mail: [dkkim1004@gmail.com](mailto:dkkim1004@gmail.com)



This chapter focuses on: (1) properties of the limnological dataset of the Nakdong River derived from Korean Long-Term Ecological Research (KLTER), (2) analysis and time-series modelling of KLTER dataset by means of machine learning techniques, and (3) benefits of applied ecological informatics for KLTER dataset.

## 20.1 Introduction

Streams and rivers are regarded as major water resources. To utilize water more efficiently from natural freshwater systems, humans have modified streams and rivers, including the construction of hydraulic features such as weirs and dams. Stream modification, which includes artificial structures, is also employed to provide for issues affecting water resource management resulting from climate change (e.g., increased flooding and drought).

One emerging issue in water resource management is that of harmful algal blooms (HABs). It is well known that HABs typically occur in eutrophic systems where the water body is stagnant. An increase in water depth and retention time and a decrease in water velocity result from river modifications, and these morphological changes function as environmental stressors, which may induce phytoplankton succession leading to HABs (principally cyanobacteria). Excessive HABs produce odor-causing components and increase water toxicity (Mihaljević and Stević 2011).

Prediction of HABs and identification of causing factors are the first actions to ameliorate water-quality deterioration associated with ecosystem degradation in the Nakdong River (Kim et al. 2007a) (Fig. 20.1). Due to the complexity of the relationship among ecosystem components, understanding the causal relationship between HABs should be accomplished using the appropriate methodologies. Ecological informatics, based on machine learning techniques, is known to be a suitable method for understanding with HAB problems (Jeong et al. 2001). Furthermore, long-term datasets including HABs provide a formidable amount of information for a machine-learning algorithm because the history of environmental problems is recorded in the dataset. The utilization of LTER data for HAB prediction will contribute to an understanding of HABs in river systems.

In this chapter, we summarize the application of ecological informatics techniques to HAB prediction, using LTER data collected in a regulated river system in East Asia (the Nakdong River, South Korea). Seasonal variation of HABs in the Nakdong River is illustrated in Fig. 20.2. The Nakdong River case studies of HAB predictions are discussed, including the identification of the relationship between HABs and environmental variables. Next, we show how to develop appropriate strategies for water quality/quantity management using modeling approaches. We also discuss future directions for the use of LTER data in ecological informatics.

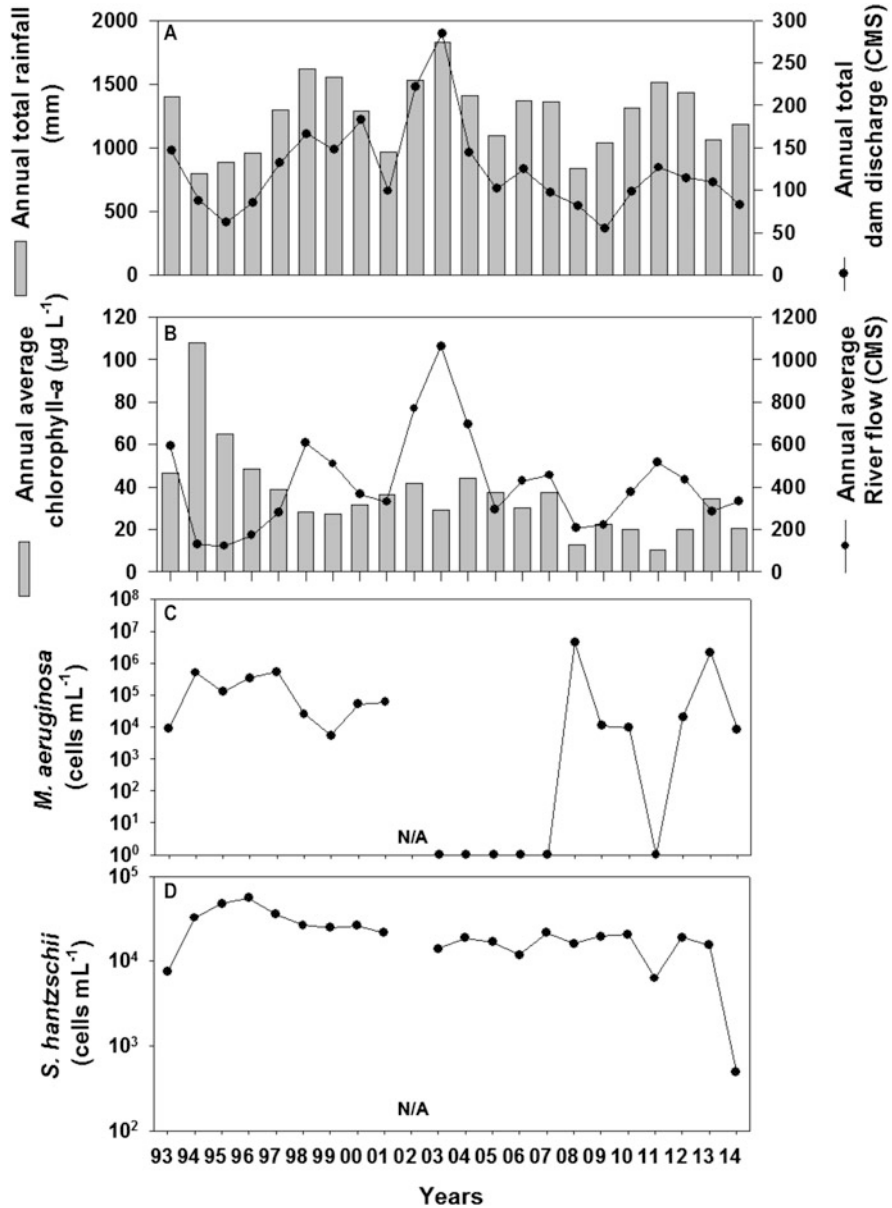


Fig. 20.1 Annual trends of the major environmental variables at Site 9 [(a) rainfall (bar) and total dam discharge (line), (b) river flow rate (line) and chlorophyll-a (bar), (c) dominant cyanobacteria average cell density, (d) dominant diatom average cell density)

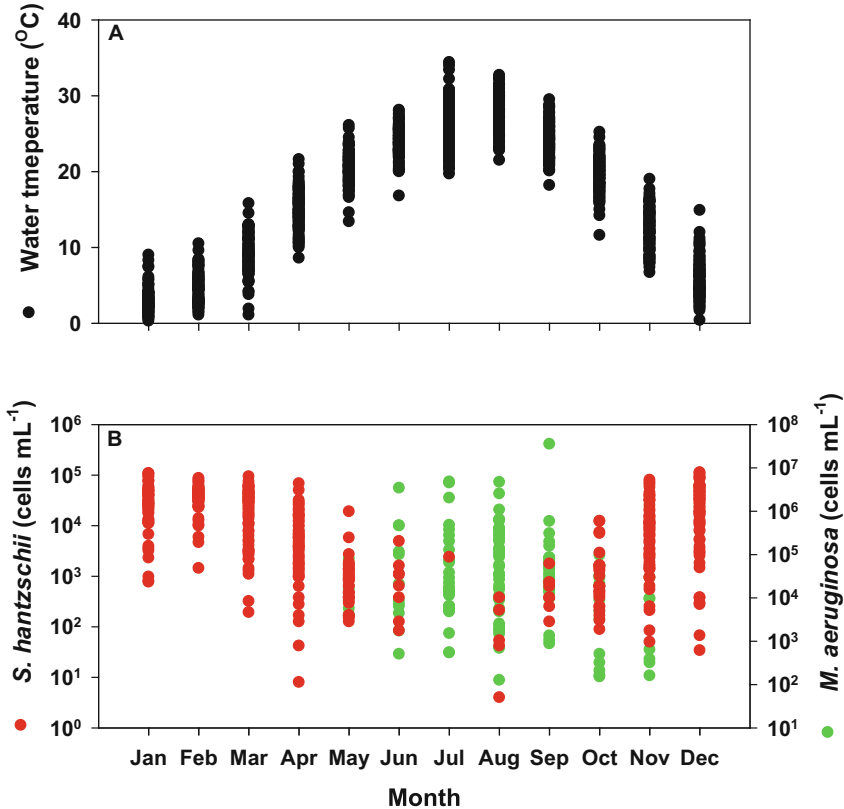
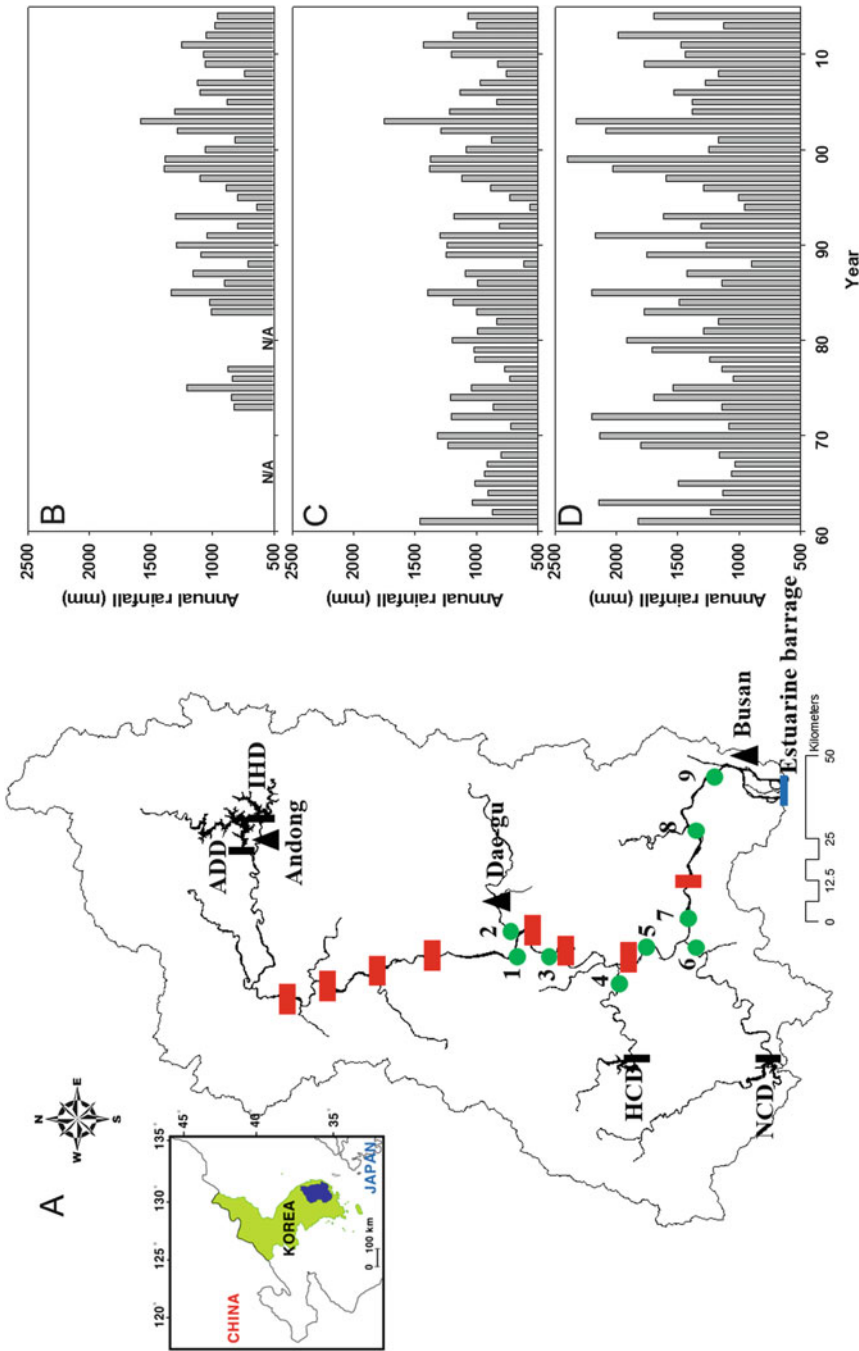


Fig. 20.2 Monthly water temperature (a), Observed cell densities of *M. aeruginosa* (green) and *S. hantzschii* (red) (b) at Site 9 (1993~2014)

## 20.2 Characteristics of the Nakdong River

The Nakdong River is located in the northern temperate region (35~37°N, 127~129°E), with four distinct seasons (cold winter in January, 30-year average  $-1.7^{\circ}\text{C}$ ; hot summer in August, average  $24.8^{\circ}\text{C}$ ). The annual average rainfall is 1202 mm, with more than 60% occurring in summertime due to seasonal monsoons (referred to as Jangma in Korean; mainly late June to mid-July) and several typhoons (mostly during July to September). In contrast, 5% of annual rainfall occurs in wintertime, which implies a seasonal drought following the rainy season. In South Korea, summer concentrated rainfall is a major factor affecting annual water quality. Due to the imbalance in annual rainfall, South Korea is recognized as the most water-stressed country among the Organization for Economic Co-operation and Development (OECD) countries (Marchal et al. 2012).

Spatial heterogeneity in rainfall distribution is also present throughout the Nakdong River basin (Fig. 20.3). The upper reaches of the river experience a



**Fig. 20.3** Locations of dams/weirs and KLTER monitoring sites (a) (filled rectangle, dams: ADD, Andong; IHD, Im-ha; HCD, Hapcheon; NCD, Namgang; filled circle, study sites: 1 Waegwan; 2 Gum-ho River; 3 Go-yeong; 4 Hwang River; 5 Juk-po; 6, Nam River; 7 Nam-ji; 8 Ha-nam; 9 Mulgeum, filled square, new weirs since 2011, filled triangle, weather station, rainfall in the Nakdong River basin). Annual precipitations observed from the weather stations at Andong (b), Daegu (c), and Busan (d)

frequent shortage in rainfall (annual average, 985.9 mm), while excessive rainfall occurs in the lower reaches of the river (annual average, 1539.9 mm; Korean Meteorological Administration 2008). The reduced rainfall of the upper reaches frequently increases problems in water security (Fig. 20.3).

Except for the upper reaches and tributaries, the riverbed slope for the main channel is low (ca. 17:10,000). This pattern is more clearly observed in the lower 160 km of the river where the transition to the estuary is very gradual (1/10,000), and hence a long retention time is observed. It has been reported that water velocity in the main channel increases only during the summer rainy season (Ha et al. 2002). Due to low water velocity from autumn to the following spring, sedimentation increases, resulting in a predominance of sand in the riverbed. The dominance of sand in the river bed appears about 330 km upriver from the estuarine barrage (Jeong et al. 2010a). Considering the river length (ca. 525 km), sand deposition occurs quickly in the main channel.

To mitigate meteorological-hydrological variation, four large multipurpose dams and one estuarine barrage are in operation (Fig. 20.3). They were built to ameliorate water security problems caused by the spatiotemporal heterogeneity of rainfall. The Nakdong River has become a “regulated river system.” Furthermore, the estuarine barrage has divided the estuarine area into freshwater and saline zones, resulting in the loss of a gradual transition (e.g., brackish zone). Levee construction intensively modified the river’s main channel, and it is difficult to find a natural or near-natural riparian zone except in the mountainous reaches upstream. Several industrial cities and two metropolitan cities are situated in the river’s mid and lower reaches, and the gradual increase in demand for water resources along with pollutant-loading is believed to accelerate eutrophication (Jeong et al. 2010a). On the other hand, the estuarine barrage has resulted in an accumulation of pollutants in the lower reach. Consequently, cyanobacteria blooms have been observed since 1991. An exceptional summer drought over 3 consecutive years (1994–1996) triggered an explosive increase in cyanobacteria density during the summer, with following diatom blooms in winter.

Korean national policy has exacerbated water quality issues in the Nakdong River. Since the Korean War in the mid-twentieth century, Korean national policies have focused on economic development, which facilitated channelization and wetland loss due to reclamation in the 1960s and 1970s. In particular, wetland loss implies a decrease in nutrient filtering opportunities, and the resultant excessive nutrient loading is responsible for eutrophication of the river system.

In 2011, an extensive 2-year river modification program concluded, including levee strengthening, dredging, and weir construction. The Nakdong River now has eight large weirs in the main channel, at 20-km intervals. As a consequence, fragmentation of the river channel and its associated habitats has occurred, cyanobacteria blooms have increased, and lentic organisms such as *Pectinatella magnifica* have been reported in the main channel (Jo et al. 2014; Seo et al. 2012).

### 20.2.1 Nakdong River's Limnological Characteristics

To understand HAB dynamics in the Nakdong River more effectively, the Limnology Laboratory of the Pusan National University initiated monitoring in 1993, in conjunction with an ongoing plankton LTER program at nine monitoring sites (six main channel sites and three tributary sites) at weekly or biweekly intervals for the last 20 years. The primary main channel site (Site 9; more frequently monitoring on a weekly basis) is 27 km away from the estuarine barrage and the remaining main channel sites (Sites 1, 3, 5, 7, and 8 on a biweekly basis) are located in the upper river reaches from the primary site at 20-km intervals (Fig. 20.3). Three major tributary sites are positioned near the confluences between the main channel and the tributaries (Sites 2, 4, and 6 on a biweekly basis). The tributaries for Sites 4 and 6 have multipurpose dams in their upstream reaches. We monitored the physico-chemical characteristics and the phyto- and zooplankton abundances for all study sites. The objectives of the LTER program were as follows: using the accumulated dataset, (1) examine the long-term water quality changes and HAB dynamics, (2) investigate the interactions between grazers (mainly zooplankton) and prey (mainly phytoplankton), (3) understand the corresponding behavior of plankton to hydrological variations through ecological modeling, and (4) explore appropriate management strategies for controlling HABs.

The limnological parameters at Site 9 (weekly monitoring) showed a strong seasonality in accordance with temperature and rainfall changes in the Nakdong River (Ha et al. 1998). Average concentrations of total nitrogen (TN) and total phosphorus (TP) were  $3.6 \pm 1.3 \text{ mg L}^{-1}$  and  $139.5 \pm 132.8 \text{ } \mu\text{g L}^{-1}$  ( $n = 1119$ , respectively). They tended to increase in dry springs (late April to May) and immediately decreased when concentrated summer rainfall occurred (Kim et al. 2011). The sites closest to the estuarine barrage, i.e., Hanam (Site 8) and Mulgeum (Site 9), showed higher nutrient concentrations than other sites.

The Geum-ho River is the most upstream among our monitoring sites. This river runs through the large city, Daegu, which derives a large amount of the point-source nutrient loading to the Nakdong River. Nutrient concentrations at the other tributaries (Nam River and Hwang River) were lower than at the Guem-ho River.

In the early 1990s, nutrient concentrations were very high. Temporal trends show a decrease in nutrient concentration over the past two decades due to significant water quality improvements (e.g., sewage treatment plant upgrade, reduction of total emission volume of water pollutants and nonpoint pollution source, and effective livestock wastewater management) (Son 2013a, b). In drought years, this nutrient concentration was connected to phytoplankton blooms.

Zooplankton abundance was higher in spring and fall and lower in summer and winter. The seasonal pattern of rotifers was similar to that for total zooplankton. This reflects the fact that rotifers (*Brachionus calyciflorus*, *B. rubens*, *Keratella cochlearis*, and *Polyarthra*) strongly dominated the zooplankton community in all locations. TN and TP increases in spring were related to increased time of water residence, which played the most important role in the abundance of large

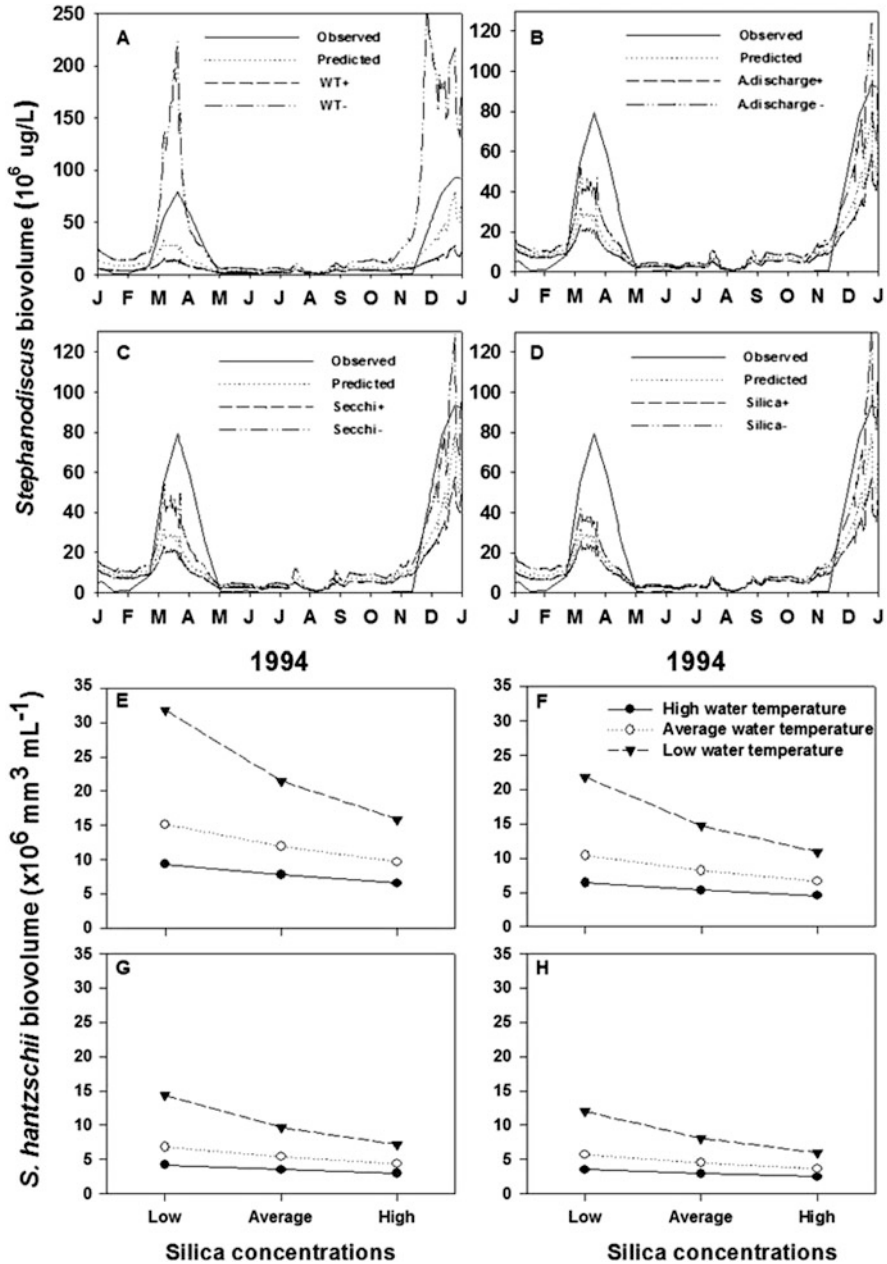
zooplankton (e.g., *Daphnia*), resulting in excessive grazing impacts. The relationship between zooplankton abundance and residence time is stronger in the lower river basin (Mulgeum) than in the mid-river basin (Kim et al. 2000, 2012a; Kim and Joo 2000).

In the Nakdong River, two HAB events consistently occur, i.e., summer cyanobacteria blooms mainly dominated by *Microcystis aeruginosa* and winter diatom blooms (Kim et al. 2011). A severe cyanobacteria bloom occurred in 1994–1996 when summer rainfall was relatively scarce (ca. 490.5 mm) compared with the 30-year average summer rainfall (ca. 660.2 mm). In contrast, it seemed that sufficient rainfall in 1998–1999 suppressed bloom formation. Jeong et al. (2011a) and Kim et al. (2011) emphasized that there was a significant negative relationship between rainfall and cyanobacteria blooms in the river system. Furthermore, Jeong et al. (2011a) reported that summer concentrated rainfall in year (t) affected the increase of dam discharge the following spring [i.e., spring in year (t + 1)]. The remaining phytoplankton bloom events are characterized by winter diatom species (*Stephanodiscus hantzschii*) (Ha et al. 1998, 2002). During winter, grazers typically disappear due to low water temperatures (4–6 °C) (Kim et al. 2000); thus, it is expected that a phytoplankton increase would not be limited by grazer populations. A series of LTER studies have revealed that low water temperature, increased water retention time, and a decrease in silica allowed the species to drastically increase during winter (Jeong et al. 2008; Kim et al. 2007a). Kim et al. (2008) found, based on growth experiments, that *S. hantzschii* isolated from the Nakdong River favored relatively lower water temperatures (4–8 °C). Kim et al. (2007a) stressed that a winter drought might fuel an abrupt increase in diatom abundance (Fig. 20.4). Kilham (1971) reported that *S. hantzschii* did not respond sensitively to low silica concentrations, which are frequently observed in the Nakdong River between November and the following February.

## 20.3 Ecological Informatics and the Nakdong River

### 20.3.1 *Applicability of Machine Learning to River Modeling (~2000s)*

The primary objective of the early modeling studies of the Nakdong River was to explore the ecological information residing in the dataset originating from multi-site studies (more than 500 sites). In that period, the most popular analytical method for large ecological datasets (species abundance with various metadata) used multivariate statistics such as canonical correlation analysis (CCA). Unfortunately, statistics-based approaches had limitations in explaining the ecological data complexity. Alternatively, applications of machine learning (ML) algorithms such as artificial neural networks (ANN), fuzzy logic (FL), or evolutionary computation (EC) to those ecological datasets were successful in the ordination of data or the prediction of interested species abundances (Park et al. 2003, 2004, 2006).



**Fig. 20.4** Diatom-forecasting model results by Kim et al. (2007a). Predictive uncertainty based on the perturbation ( $\pm$ S.D.) of water temperature (a), Andong dam discharge (b), Secchi depth (c), silica concentration (d). Sensitivity results of the diatom based on simultaneously perturbed inputs including silica, water temperature, dam discharge, and Secchi depth; low and high dam discharge and storage (e and f, respectively), low and high dam discharge and storage under the condition of high Secchi depth (g and h, respectively)



An early application of ML to river time-series data can be found in Jeong et al. (2001). After Recknagel (1997) applied an ANN to the prediction of cyanobacteria blooms in Lake Kasumigaura (Japan), they reported the successful application of a recurrent neural network (RNN) to time-series prediction of phytoplankton biomass (chlorophyll-*a*) using a 5-year weekly limnological dataset. Further application of the RNN to river phytoplankton blooms can be found in Jeong et al. (2006a), in which they developed a 7-day-ahead predictive model for *M. aeruginosa* and *S. hantzschii*. From those results, they reported that an internal loop generating additional input data to external input (i.e., the raw environmental data used in training) in RNN training was helpful for time-series prediction, and they determined that river phytoplankton blooms were largely affected by upstream dam flow–river flow controls and nutrient conditions.

Besides ANN modeling, a series of EC modeling attempts for the prediction of phytoplankton blooms were implemented for the Nakdong River data. EC is an adaptive method that mimics biological processes of evolution, natural selection, and genetic variation. Jeong et al. (2003) utilized a genetic programming (GP) algorithm for the development of an *M. aeruginosa* predictive equation model. In their study, they emphasized that the GP algorithm might assure the search for an optimal ecological model through a global search and that the predictability of the time-series cyanobacteria species was sufficient. Furthermore, by using a time-delayed input in the model training process, they successfully achieved a short-term future prediction. The other related study applied a hybrid evolutionary algorithm (HEA) based on rule discovery to the Nakdong River data for forecasting chlorophyll-*a* and for elucidating complex nonlinear relationships between input and output variables (Cao et al. 2006; Joo et al. 2003; Kim et al. 2007a).

### **20.3.2 Ecological Elucidation of HAB Dynamics in the Nakdong River**

After the evaluation of machine learning applicability, KLTER data for the Nakdong River was used to identify the relationship between HABs and the environment through sensitivity and scenario analyses. The basic assumptions for this process were described in Jeong et al. (2007); as the upstream dam discharge increased, smaller magnitude HABs were observed. This pattern also appeared in other regions such as Australia (Maheshwari et al. 1995; Maier et al. 2001, 2004; Walker and Thoms 1993). The negative relationship between river flow and dam discharge and HABs was simulated in a series of ecological modeling studies focused on phytoplankton dynamics. Sensitivity analysis revealed that increased dam discharge rates maintained a lower level of chlorophyll-*a* in the river (Jeong et al. 2010b). From a species point of view, sensitivity analysis indicated that increases in dam discharge and river flow hindered the development of HABs by *M. aeruginosa* (Jeong et al. 2006a; Kim et al. 2007a). Furthermore, when a simple scenario (water temperature, +3 °C; upriver dam discharge rate,  $-10 \text{ m}^3 \text{ s}^{-1}$ ) was

applied to the model developed by Jeong et al. (2006a), the magnitude of the summer HABs was extended and several HABs occurrences were predicted (Joo et al. 2008).

There is also another example of ecological explanation based on machine learning applications. Kim et al. (2007a) generated empirical nonlinear equations using GP, and elucidated complex mechanisms of winter diatom blooms that recurrently occurred in the Nakdong River. Through different sensitivity analyses, they articulated that the winter blooms could be mainly driven by *S. hantzschii* that outcompeted the other phytoplankton at the low level of silica concentration and water temperature in the Nakdong River. Specifically, they speculated that a decrease in water temperature might cause stress to and threaten the survival of other phytoplankton species. Intensive competition for nutrients, such as  $\text{SiO}_2$ , between the diatoms in autumn (mainly from September to November) caused when  $\text{Si:P} < 10$ , often resulted in a dominance of *S. hantzschii* in lake environments (Kilham et al. 1986; Kolmakov et al. 2002). Furthermore, a decrease in water temperature affected the survival rate of grazers (e.g., zooplankton). Kolmakov et al. (2002) attempted a multiple sensitivity analysis and concluded that *S. hantzschii* proliferation in the river resulted from a combination of three environmental factors (low water temperature, Si:P ratio and photosynthetically active radiation).

As shown in Jeong et al. (2006b), the selection of appropriate input variables was important to assure the representativeness of the developed models. One simple method to facilitate the selection could come from an EC application (Kim et al. 2007a). EC modeling allows one to search for appropriate input variables in the model development process. If multiple candidate equation models are available, it is possible to compare the input variable selectivity between the models. Frequently adopted input variables can be regarded as more important in determining target variables. Jeong et al. (2003) and Kim et al. (2007a) reported that water temperature, dissolved oxygen (DO), pH, and Secchi transparency were frequently selected among variables and these variables were helpful for selecting the best out of multiple candidate models. Furthermore, the fewer number of input variables makes the model structure understood more easily when the model performance is reasonably acceptable. Thus, it is desirable to adopt several variables that can be easily monitored and that greatly increase the model performance. In this context, Kim et al. (2007a) considered variable selectivity thoroughly, and Kim et al. (2012b) showed different sensitivity of input variables according to short- (1-week) and long-term (1-year) forecasting.

In addition to relating the input environments–output pattern, one attempt revealed that machine learning offers the potential to implement autoregressive data processing. Jeong et al. (2008) developed a simple neural network model, namely the temporal autoregressive RNN. They expected that a delayed input of external data (i.e.,  $n$ -week previous cell density) and internal data loop by a time-delayed recurrent neural network might allow the network model to predict future cell density (i.e., cell density of  $n$ -weeks ahead) using previous cell density. Jeong et al. (2008) reported that the timing of bloom formation and accuracy of cell density prediction were acceptable, and the seasonality of input data did not

affect predictability. From this evidence, it is believed that ML (machine learning) algorithms are flexible enough for the purposes of ecological modeling.

### **20.3.3 *Model Applications: Scenarios for Smart Flow***

In a regulated river system, humans control the river flow using hydraulic structures such as dams and weirs in order to manage water resource demands. This artificial control is typically more intensive in East Asia because summer concentrated rainfall is practically the only source of water. If a dry summer persists, the hydraulic structures regulate the rate of water flow more strictly to maintain minimum water levels. Cyanobacteria density has explosively increased when a stable water environment persisted in the river (Joo et al. 2003; Kim et al. 2007b), and Jeong et al. (2007) warned of an increased probability of cyanobacteria proliferation when two or more consecutive summers were dry.

Several studies implemented in the Nakdong River considered the potential applicability of “smart flow” for the reduction of HABs. Cyanobacteria density is frequently diminished when a sudden discharge of water occurs (Maier et al. 2001; Webster et al. 2000). Based on this information, summer rainfall and controlled dam discharge can decrease summer cyanobacteria blooms and winter diatom blooms until spring of the next year (i.e., summer to the next spring; Jeong et al. 2011a). In order to control blooms more efficiently, we added an estuarine barrage for regulation. Hong et al. (2014) hypothesized a simultaneous operation of “upriver dams discharge increased and the estuarine barrage discharge decreased” can make flushing + dilution effect for reducing phytoplankton density (Fig. 20.5).

## **20.4 Future Research**

In our studies, we have demonstrated a variety of data sources, from plankton to mammals, available from the Nakdong River basin. These data can be used for an interdisciplinary approach. We introduce a variety of research results and possible future research projects.

### **20.4.1 *Scale Up: From Sites to Basin***

The Four Major Rivers Restoration Project was completed in 2011 by the Korean government, which resulted in enormous changes to the river ecosystem. These changes are comparable to those for the estuarine barrage constructed in 1987, which resulted in the clear separation of the pre-existing brackish area into freshwater and saline zones. Both weir construction (eight weirs normal pool level; max: 47.0 m, average: 24.3 m) and intensive dredging of the main channel increased

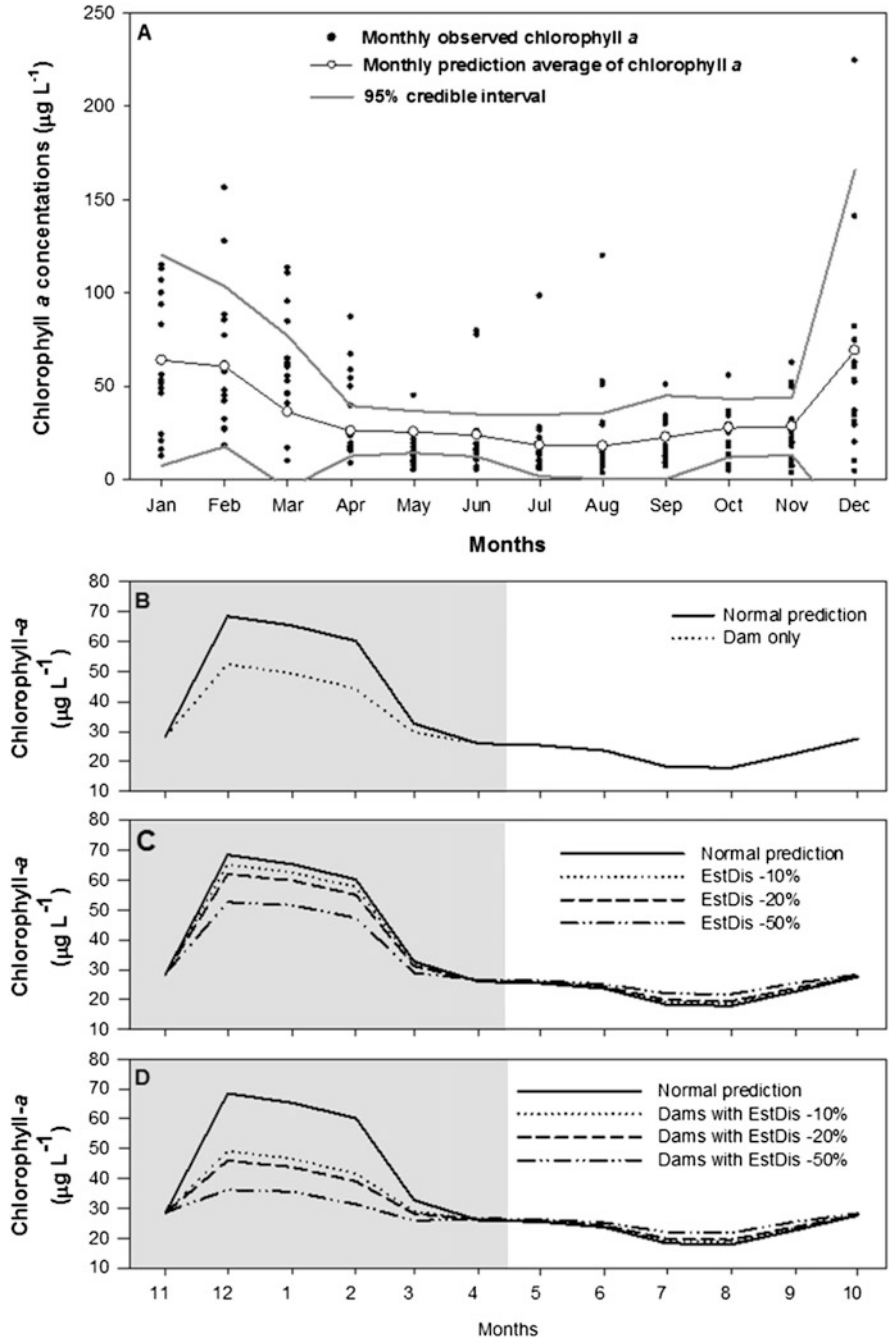
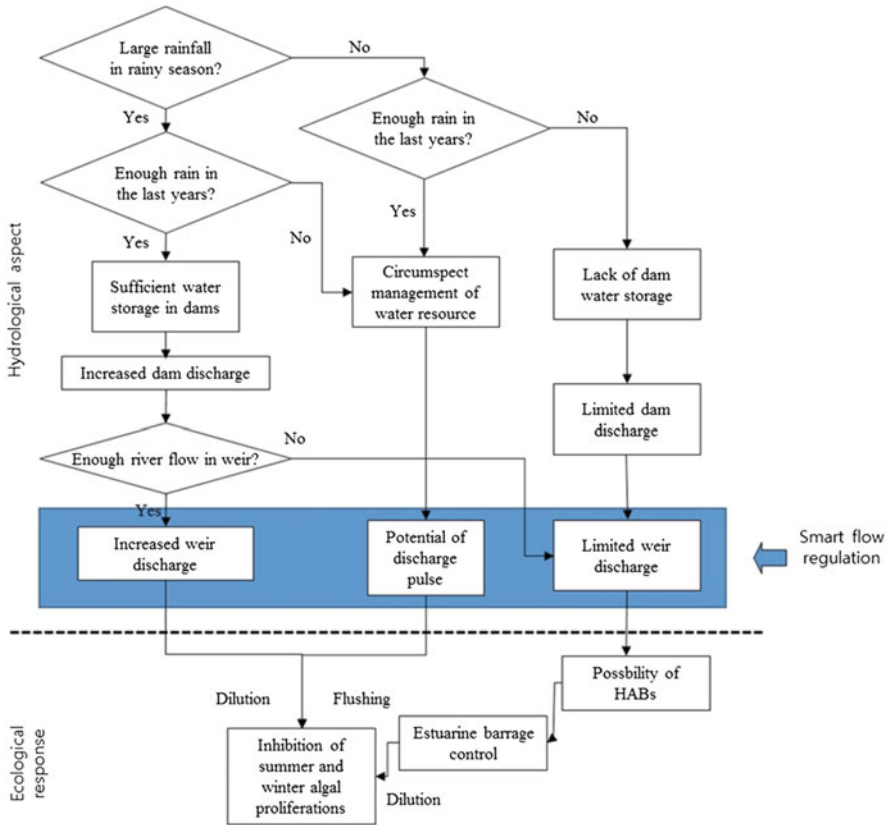


Fig. 20.5 Chlorophyll-a prediction model using HEA and scenario analysis results [(a) The thin line with blank circles is the monthly averaged predicted chlorophyll a concentration, and the black dots are the observed chlorophyll-a concentrations for every month. The thick solid lines



**Fig. 20.6** Diagram depicting the river regulation mechanism with respect to climate changes and phytoplankton population dynamics [modified from Jeong et al. (2007)]. According to rainfall in rainy season, dam and weir storage and discharge, HAB control strategies can be adopted to regulated river using dilution and flushing by artificial structures

water retention time and have caused drastic changes in the phytoplankton assemblage. Since the completion of the river project, summer HABs have been observed across the entire river channel (The blooms occurred mainly in the lower part of the river before the project). In this regard, appropriate remediation strategies should be urgently established to prevent the HAB expansion. For example, given the highly intensive summer rainfall and multiple dams’ hydrocapacity in the Nakdong River basin, Jeong et al. (2007) proposed effective flow regulations (e.g., strong pulse discharge) in order to flush out the HABs in the river (Fig. 20.6). They emphasized

**Fig. 20.5** (continued) indicate 95% of the credible interval for the prediction. Chlorophyll-a concentration responses based on different scenarios of hydrological control. (b) Sole control of upriver dams; (c) Sole control of the estuarine barrage. (d) Simultaneous application of both dams and the estuarine barrage (Hong et al. 2014)]

that river flow control by means of upriver dam discharge regulation was a crucial factor for mitigation of HAB occurrence in those years with sufficient or moderate annual rainfall. However, excessive proliferation of algae in dry years was not considered. Further development of ecological models may elucidate the relationship between HABs and water environments with regard to river modifications. HABs are also influenced by nutrient concentrations, which are largely related to basin nutrient loading. If information about basin loading and in-water nutrients is available, sensitivity and scenario analyses may help in the development of appropriate nutrient control strategies, with an aim of reducing HABs in conjunction with flow regulation.

#### **20.4.2 Ecological Informatics and KLTER**

KLTER data accumulated from the Nakdong River is a valuable source of information to assess the impact of climate variations on biological entities in a regulated river system. A recent study has shown that HABs in the Nakdong River were closely related to the Indian Ocean Dipole (IOD) (Jeong and Joo 2016). The researchers emphasized that as the IOD became positive, moisture convection was directed toward western Africa, leading to severe droughts throughout East Asia, including the Korean Peninsula. The droughts are considered as a causal factor of cyanobacteria increases in South Korea. In this regard, we anticipate that well-developed ecological models will be able to guide water-resources management in response to climate changes. Subsequently, the relationship identified in the previous modeling stage will allow us to discover feasible adaptations to climate change (Jones et al. 2012).

The Nakdong River estuary is an important habitat for migratory birds (swans, mallards, wild geese, etc.) in the East Asian-Australasian flyway. Currently, more than 10 years' KLTER data for migratory birds are available, and one migratory species, the little tern (*Sternula albifrons*), has been intensively studied using a number of ecological informatics methods. Hybrid evolutionary algorithm (HEA) was used to find the most appropriate conditions for nest site selection for the bird species (Jeong et al. 2011b), which successfully elucidated the site selection hypotheses postulated by other research groups. In addition, a continuous wavelet transformation (CWT) has revealed that this species significantly responded to the Korean monsoon rainfall, and the arrival of individual birds at the estuary the following year was negatively affected by the previous year's monsoon rainfall and onset dates (Jang et al. 2014).

A fish community investigation (seasonal monitoring) in the Nakdong River has been conducted for more than 10 years at three sites in the river's main channel. Specifically, the distribution and dispersal of exotic species [largemouth bass (*Micropterus salmoides*) and bluegill (*Lepomis macrochirus*)] have been intensively monitored (Jo et al. 2011). Researchers successfully identified a gradual increase of exotic species' relative abundance and their continuous impact on native

prey species. Long-term monitoring for a mammalian species (nutria, *Myocastor coypus*) has been initiated to discover relationships between species dispersion and their surrounding environments. Using 3 years of monitoring data, Hong et al. (2015) determined that a persistently cold winter (total number of days below  $-4^{\circ}\text{C}$ ) was significantly correlated with nutria occurrence. Furthermore, an extensive comparison between monitoring data and the literature showed that alien species' habitats are extending north in response to increases in temperature. Thus, the accumulation of exotic species data will allow identification of the relationship between species distribution and global warming.

Since the completion of the Four Major Rivers Restoration Project, *Pectinatella magnifica*, a bryozoan, has been widely observed throughout the Nakdong River basin. There are limited studies of this species. Particularly in 2014, it was the most frequent environmental news topic in Korea. The causes of bryozoan events are presumably due to increased retention time along with higher water temperatures (Choi et al. 2015; Jo et al. 2014); however, further studies are required to reveal the causes.

## 20.5 Conclusions

The Nakdong River is a regulated river, and a number of artificial structures (dams, weirs, and estuarine barrage) regulate flow in this river. Inter-annual variation of precipitation in the Nakdong River basin has significantly influenced dam operation, thereby affecting the frequency of HAB occurrence as well as the hydraulic residence time of water. KLTER helps us understand complex patterns associated with flow regulation in Korean river ecosystem. KLTER data allow quantitative assessments of ecosystem health and integrity with respect to a large, complex, and regulated river ecosystem in South Korea.

**Acknowledgements** This work was supported by National Research Foundation of Korea (NRF) Grants funded by the Korea government (Grant No. NRF-2014M3C8A4030721).

## References

- Cao H, Recknagel F, Joo G-J, Kim D-K (2006) Discovery of predictive rule sets for chlorophyll-a dynamics in the Nakdong River (Korea) by means of the hybrid evolutionary algorithm HEA. *Ecol Inf* 1:43–53
- Choi J-Y, Joo G-J, Kim S-K et al (2015) Importance of substrate material for sustaining the bryozoan *Pectinatella magnifica* following summer rainfall in lotic freshwater ecosystems, South Korea. *J Ecol Environ* 38:375
- Ha K, Kim H-W, Joo G-J (1998) The phytoplankton succession in the lower part of hypertrophic Nakdong River (Mulgum), South Korea. *Hydrobiologia* 369-370:217–227

- Ha K, Jang M-H, Joo G-J (2002) Spatial and temporal dynamics of phytoplankton communities along a regulated river system, the Nakdong River, Korea. *Hydrobiologia* 470:235–245
- Hong D-G, Jeong K-S, Kim D-K, Joo G-J (2014) Remedial strategy of algal proliferation in a regulated river system by integrated hydrological control: an evolutionary modelling framework. *Mar Freshw Res* 65:379–395
- Hong S, Do Y, Kim J et al (2015) Distribution, spread and habitat preferences of nutria (*Myocastor coypus*) invading the lower Nakdong River, South Korea. *Biol Invasions* 17: 1485–1496
- Jang J-D, Chun S-G, Kim K-C et al (2014) Long-term adaptations of a migratory bird (Little Tern *Sternula albifrons*) to quasi-natural flooding disturbance. *Ecol Inf* 29:166–173
- Jeong K-S, Joo G-J (2016) Effect of Indian Ocean Dipole signal on freshwater cyanobacterial dynamics. *Inland Waters* 6:414–422
- Jeong K-S, Joo G-J, Kim H-W et al (2001) Prediction and elucidation of phytoplankton dynamics in the Nakdong River (Korea) by means of a recurrent artificial neural network. *Ecol Model* 146:115–129
- Jeong K-S, Kim D-K, Whigham P et al (2003) Modelling *Microcystis aeruginosa* bloom dynamics in the Nakdong River by means of evolutionary computation and statistical approach. *Ecol Model* 161:67–78
- Jeong K-S, Recknagel F, Joo G-J (2006a) Prediction and elucidation of population dynamics of a blue-green algae (*Microcystis aeruginosa*) and diatom (*Stephanodiscus hantzschii*) in the Nakdong River-Reservoir System (South Korea) by artificial neural networks. In: Recknagel F (ed) *Ecological informatics: scope, techniques and applications*. Springer, Berlin, pp 255–273
- Jeong K-S, Kim D-K, Joo G-J (2006b) River phytoplankton prediction model by artificial neural network: model performance and selection of input variables to predict time-series phytoplankton proliferations in a regulated river system. *Ecol Inf* 1:235–245
- Jeong K-S, Kim D-K, Joo G-J (2007) Delayed influence of dam storage and discharge on the determination of seasonal proliferations of *Microcystis aeruginosa* and *Stephanodiscus hantzschii* in a regulated river system of the lower Nakdong River (South Korea). *Water Res* 41: 1269–1279
- Jeong K-S, Kim D-K, Jung J-M et al (2008) Non-linear autoregressive modelling by temporal recurrent neural networks for the prediction of freshwater phytoplankton dynamics. *Ecol Model* 211:292–300
- Jeong K-S, Hong D-G, Byeon M-S et al (2010a) Stream modification patterns in a river basin: field survey and self-organizing map (SOM) application. *Ecol Inf* 5:293–303
- Jeong K-S, Kim D-K, Shin H-S et al (2010b) Flow regulation for water quality (chlorophyll *a*) improvement. *Int J Environ Res* 4:713–724
- Jeong K-S, Kim D-K, Shin H-S et al (2011a) Impact of summer rainfall on the seasonal water quality variation (chlorophyll *a*) in the regulated Nakdong River. *KSCE J Civil Eng* 15: 983–994
- Jeong K-S, Jang J-D, Kim D-K, Joo G-J (2011b) Waterfowls habitat modeling: simulation of nest site selection for the migratory Little Tern (*Sterna albifrons*) in the Nakdong estuary. *Ecol Model* 222:3149–3156
- Jo H, Jang M-H, Jeong K-S et al (2011) Long-term changes in fish community and the impact of exotic fish, between the Nakdong River and Upo Wetlands. *J Ecol Environ* 34:59–68
- Jo H, Joo G-J, Byeon M et al (2014) Distribution pattern of *Pectinatella magnifica* (Leidy, 1851), an invasive species, in the Geum River and the Nakdong River, South Korea. *J Ecol Environ* 37:217–223
- Jones JA, Creed IF, Hatcher KL et al (2012) Ecosystem processes and human influences regulate streamflow response to climate change at long-term ecological research sites. *BioScience* 62: 390–404
- Joo G-J, Jang M-H, Park S-B (2003) The application of an algal fence for the reduction of algal intake into the water intake facility. *Kor J Limnol* 36:467–472



- Joo G-J, Kim D-K, Yoon J-D, Jeong K-S (2008) Climate changes and freshwater ecosystems in South Korea. *J KSEE* 30:1–6
- Kilham P (1971) A hypothesis concerning silica and the freshwater planktonic diatoms. *Limnol Oceanogr* 16:10–18
- Kilham P, Kilham SS, Hecky RE (1986) Hypothesized resource relationships among African planktonic diatoms. *Limnol Oceanogr* 31:1169–1181
- Kim H-W, Joo G-J (2000) The longitudinal distribution and community dynamics of zooplankton in a regulated large river: a case study of the Nakdong River (Korea). *Hydrobiologia* 438: 171–184
- Kim H-W, Hwang S-J, Joo G-J (2000) Zooplankton grazing on bacteria and phytoplankton in a regulated large river (Nakdong River, Korea). *J Plankton Res* 22:1559–1577
- Kim D-K, Jeong K-S, Whigham PA, Joo G-J (2007a) Winter diatom blooms in a regulated river in South Korea: explanations based on evolutionary computation. *Freshw Biol* 52:2021–2041
- Kim D-K, Cao H, Jeong K-S et al (2007b) Predictive function and rules for population dynamics of *Micyocystis aeruginosa* in the regulated Nakdong River (South Korea), discovered by evolutionary algorithms. *Ecol Model* 203:147–156
- Kim M-C, La G-H, Kim H-W (2008) The effect of water temperature on proliferation of *Stephanodiscus* sp. in vitro from the Nakdong River, South Korea. *Kor. J Limnol* 41:26–33
- Kim D-K, Hong D-G, Kim H-W et al (2011) Longitudinal patterns in limnological characteristics based on long-term ecological research in the Nakdong River. *J Ecol Field Biol* 34:39–47
- Kim D-K, Jeong K-S, Chang K-H et al (2012a) Patterning zooplankton communities in accordance with annual climatic conditions in a regulated river system (Nakdong River, South Korea). *Int Rev Hydrobiol* 97:55–72
- Kim D, Jeong K, McKay R et al (2012b) Machine learning for predictive management: short and long term prediction of phytoplankton biomass using genetic algorithm based recurrent neural network. *Int J Environ Res* 6:95–108
- Kolmakov VI, Gaevskii NA, Ivanova EA et al (2002) Comparative analysis of ecophysiological characteristics of *Stephanodiscus hantzschii* Grun in the periods of its bloom in recreational water bodies. *Russ J Ecol* 33:97–103
- Korean Meteorological Administration (2008) Weather information. Seoul, South Korea. <http://www.kma.go.kr>. Accessed 6 Feb 2017
- Maheshwari BL, Walker KF, McMahon TA (1995) Effects of regulation on the flow regime of the River Murray, Australia. *Reg Rivers Res Manage* 10:15–38
- Maier HR, Burch MD, Bormans M (2001) Flow management strategies to control blooms of the cyanobacterium, *Anabaena circinalis*, in the River Murray at Morgan, South Australia. *Reg Rivers Res Manage* 17:637–650
- Maier HR, Kingston GB, Clark T et al (2004) Risk-based approach for assessing the effectiveness of flow management in controlling cyanobacterial blooms in rivers. *River Res Appl* 20: 459–471
- Marchal V, Dellink R, Van Vuuren D et al (2012) OECD environmental outlook to 2050. <https://www.oecd.org/env/cc/49082173.pdf>. Accessed 20 Feb 2017
- Mihaljević M, Stević F (2011) Cyanobacterial blooms in a temperate river-floodplain ecosystem: the importance of hydrological extremes. *Aquat Ecol* 45:335–349
- Park Y-S, Céréghino R, Compin A, Lek S (2003) Application of artificial neural networks for patterning and predicting aquatic insect species richness in running waters. *Ecol Model* 160: 165–280
- Park Y-S, Chon T-S, Kwak I-S, Lek S (2004) Hierarchical community classification and assessment of aquatic ecosystems using artificial neural networks. *Sci Total Environ* 327:105–122
- Park Y-S, Tison J, Lek S et al (2006) Application of a self-organizing map to select representative species in multivariate analysis: a case study determining diatom distribution patterns across France. *Ecol Inform* 1:247–257
- Recknagel F (1997) ANNA – artificial neural network model for predicting species abundance and succession of blue-green algae. *Hydrobiologia* 349:47–57

- Seo D, Kim M, Ahn JH (2012) Prediction of Chlorophyll-a changes due to weir constructions in the Nakdong River using EFDC-WASP modelling. *Environ Eng Res* 17:95–102
- Son H-J (2013a) Changes of dominant phytoplankton community in downstream of the Nakdong River: from 2002 to 2012. *J KSEE* 35:289–293
- Son H-J (2013b) Long-term variations of phytoplankton biomass and water quality in the downstream of Nakdong River. *J KSEE* 35:263–267
- Walker KF, Thoms MC (1993) Environmental effects of flow regulation on the lower river Murray, Australia. *Reg Rivers Res Manage* 8:103–119
- Webster IT, Sherman BS, Bormans M, Jones G (2000) Management strategies for cyanobacterial blooms in an impounded lowland river. *Reg Rivers Res Manage* 16:513–525

# Chapter 21

## From Ecological Informatics to the Generation of Ecological Knowledge: Long-Term Research in the English Lake District

S.C. Maberly, D. Ciar, J.A. Elliott, I.D. Jones, C.S. Reynolds,  
S.J. Thackeray, and I.J. Winfield

**Abstract** Lakes are highly connected systems that are affected by a hierarchy of stressors operating at different scales, making them particularly sensitive to anthropogenic perturbation. Traditionally, lakes are studied as a whole system ‘from physics to fish’ and long-term monitoring programmes were initiated on this basis, some starting over a century ago. This chapter describes the long-term monitoring programme on the Cumbrian lakes, UK, how it is operated and how its scientific value is increased by combining it with additional activities. Case-studies are presented on the advances long-term research has made to testing ecological theory and understanding teleconnexions and phenology. Automatic high-frequency measurements are an important complementary approach that has been made possible by technological revolutions in computing, and telecommunications. They provide a window into the true dynamic nature of lakes that cannot be achieved by manual sampling. The large volume of data produced can now be quality controlled and analysed by bespoke software that has been developed in recent years by a global network of lake and data scientists. Finally, lake models constructed using the insights from monitoring, as well as experiments, are powerful ways to identify knowledge gaps and allow forecasts to be made of future responses to environmental change or management intervention. As other approaches become incorporated into lake research, such as Earth Observation and citizen science, the scale of knowledge about the system will increase, improving our ability to provide robust scientific advice for the sustainable management of these fragile, but important ecosystems.

---

S.C. Maberly (✉) • D. Ciar • J.A. Elliott • I.D. Jones • C.S. Reynolds • S.J. Thackeray • I.J. Winfield

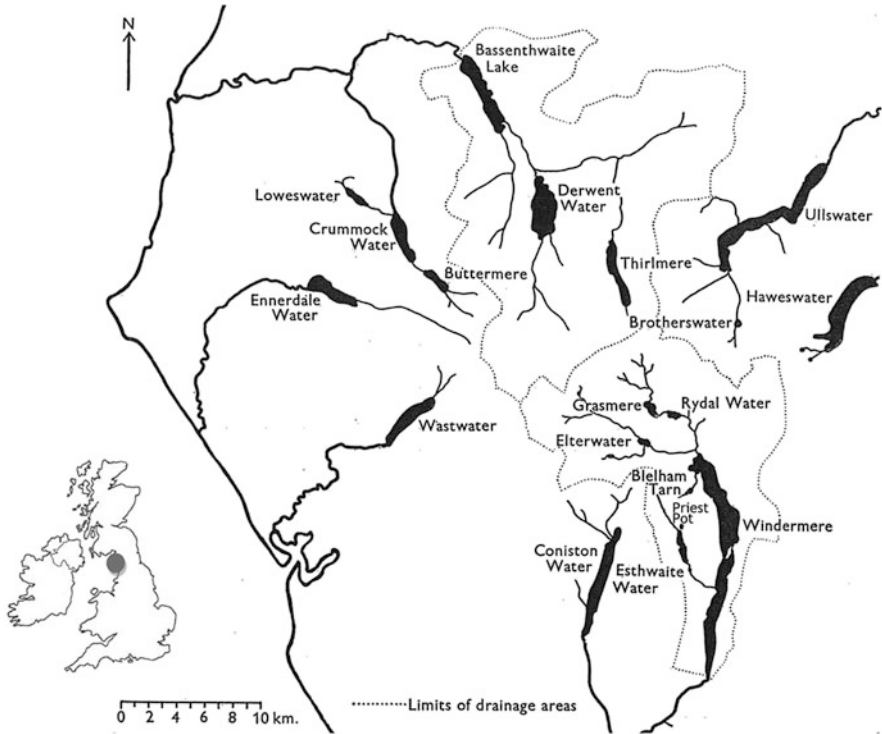
Centre for Ecology & Hydrology, Lancaster Environment Centre, Lancaster, UK  
e-mail: [scm@ceh.ac.uk](mailto:scm@ceh.ac.uk); [dciar86@ceh.ac.uk](mailto:dciar86@ceh.ac.uk); [alexe@ceh.ac.uk](mailto:alexe@ceh.ac.uk); [ianj@ceh.ac.uk](mailto:ianj@ceh.ac.uk); [csr@ceh.ac.uk](mailto:csr@ceh.ac.uk);  
[sjtr@ceh.ac.uk](mailto:sjtr@ceh.ac.uk); [ijw@ceh.ac.uk](mailto:ijw@ceh.ac.uk)

## 21.1 Introduction

Lakes are usually distinct features in the landscape whose clear boundaries belie the fact that they are highly-connected with their immediate landscape, airshed and, through biogeochemical cycles, with the planet. This was not fully recognised in the past when there was a tendency to focus on a lake as the system of interest, perhaps by analogy with approaches and concepts developed in the field of oceanography. However, modern studies recognise that lakes are influenced by stressors operating at a hierarchy of scales from local processes in the catchment, regional weather patterns, atmospheric deposition and invasion from regional species-pools, and global change (Maberly and Elliott 2012). The external stressors, and the complex internal interactions that they trigger, control the structure and function of lakes. Humans rely on lakes for a wide range of benefits including water supply, food production, hydropower generation, flood control, tourism and less tangible, but important, aesthetic and cultural fulfilment. They have also been used as a convenient means of waste disposal which brings into particularly sharp focus the tension between the requirement to derive goods and services from our environments and the need to use them sustainably.

The English Lake District, situated in Cumbria in north-west England (Fig. 21.1), is part of the Lake District National Park. The lakes were formed around 14,000 to 15,000 years ago after the retreat of glaciers at the end of the last ice age (Pearsall and Pennington 1947). The glaciers produced a pattern of lakes radiating from a central dome of high land with a current maximum elevation of 978 m. Over 200 lakes with an area of greater than 0.001 km<sup>2</sup> are present in the national park and 10 lakes have an area greater than 1 km<sup>2</sup> including England's largest natural lake, Windermere (Pickering 2001). The variety of lake elevation, catchment geology and soils, lake depth, morphometry, flushing rate and trophic state (Talling 1999) in such a small geographic area is unusual and provides a valuable opportunity to distinguish between the effects of local, regional and global stressors. Partly because of this, the lakes are among the best studied in the world in terms of intensity, and duration of research on all aspects of lake ecology.

Perhaps because of the apparent homogeneity of the open-water, there is a long tradition in freshwater research of undertaking 'ecosystem ecology' (Moss 2012) that studies different components of the system 'from physics to fish'. A holistic approach to limnology was also stimulated by the founding of specialist laboratories on the shores of lakes such as those on Plön in Germany in 1891 and on Lake Fure in Denmark in 1897, founded by Carl Wesenberg-Lund (Sand-Jensen 1997) and similar stations in Sweden, Hungary and elsewhere. In the United Kingdom, the Freshwater Biological Association (FBA), founded in 1929, established a laboratory on the shores of Windermere in the English Lake District (Talling 2008). In addition to experiments and development of new techniques and equipment, it was natural for the scientists to start to monitor regularly the physical, chemical and biological conditions in Windermere and nearby lakes.



**Fig. 21.1** Map of the English Lake District showing the major lakes and its location within Great Britain (*inset*). The watersheds for the main study lakes are shown by *dotted lines* [adapted with permission from Maberly and Elliott (2012)]

Following the establishment of appropriate infrastructure by the FBA in 1931, Windermere's fish populations immediately became a subject of research with Allen (1935) describing the diet and seasonal migrations of perch (*Perca fluviatilis*), its most numerous fish species. However, the subsequent outbreak of World War II in September 1939 shifted the nature of fish research at Windermere towards applied fisheries objectives given the constraints imposed on the UK's marine fisheries. A capture fishery for the lake's perch was quickly established (Worthington 1942) and developed (Worthington 1950), along with an associated culling programme of the lake's pike (*Esox lucius*) as described by Le Cren (2001). Together with contemporary work on Windermere's formerly commercially exploited Arctic charr (*Salvelinus alpinus*) (Kipling 1972), these fisheries activities evolved into a unique long-term monitoring programme for these three contrasting species which continues to the present (Craig et al. 2015). Its results have improved scientific understanding of the Windermere ecosystem and the 'overfishing problem' which challenges marine fisheries biologists around the world.

In 1945, John W.G. Lund began to study four lakes, (i.e. the North and South Basins of Windermere, Esthwaite Water and Blelham Tarn) in order to determine

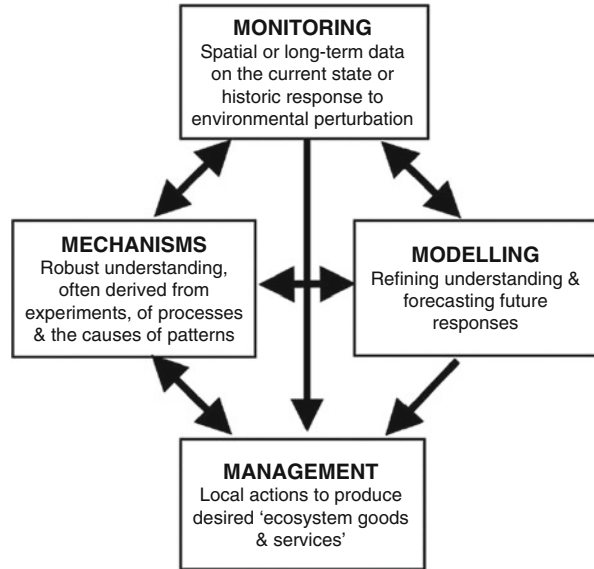
the causes of seasonal phytoplankton growth. This continued after his retirement in 1977 and became, unintentionally and serendipitously, a long-term monitoring programme. The research was transferred to the newly formed Institute of Freshwater Ecology in 1989 and this institute was later merged with other government funded institutes to form the Centre for Ecology & Hydrology (CEH). Today, CEH continues this long-term research, which comprises fortnightly studies on seven lake basins. The 65 years of research in this area was celebrated in a special issue of the journal *Freshwater Biology* in 2012 (Maberly and Elliott 2012).

Many key ecological concepts relating to food webs and energy flow, alternative stable states and ecological theory have been developed in lakes. However their study is also essential if we are to understand lakes so that we can manage them sustainably. Long-term monitoring produces an extremely valuable insight into how lakes have responded to past perturbations. Seasonal, inter-annual and decadal patterns of change can be discerned, and where many aspects of the system are studied in parallel, attribution of the causes of change can begin to be made.

Lakes are highly dynamic systems; the microbial populations that drive many of their functions have generation times that are much shorter than the typical traditional sampling interval. In addition, short-term events, such as a sudden storm, can have a large effect on the temperature structure of a lake with profound effects for a range of factors including underwater light climate (via the depth of the mixed layer), nutrient cycling and oxygen profiles. The development of appropriate sensors, computing hardware and software and communication technologies has allowed samples to be collected automatically at minute intervals, dramatically increasing our appreciation of higher-frequency dynamics.

However, no programme can monitor all the complex variables that affect lakes. Furthermore, as the size of the human population and the sophistication of society increases, new, previously unforeseen stressors and problems arise that need to be assessed and their effects understood. As a result, complementary approaches are needed to characterise more fully how lakes currently operate and may respond to future challenges. The conceptual model for how we undertake our research in the English Lake District was characterised by Maberly and Elliott (2012) as the 'four Ms' (Fig. 21.2): Monitoring, Mechanisms, Models and Management. These interact and reinforce each other. Experimentation can provide causal understanding of the mechanisms behind the patterns and processes observed in the field. They can range from laboratory experiments that allow the investigator a high level of control, but inevitably sacrifice ecological realism. Experimental approaches can be scaled up to address this issue, increasing realism at the expense of control and replication, to shore-based mesocosms (e.g. Liboriussen et al. 2005), in lake mesocosms (e.g. Lack and Lund 1974) and whole lake experiments (e.g. Schindler 1990). Secondly, the process-based understanding these monitoring and mechanistic studies produce can be encapsulated into process-based or statistical models. These help to refine understanding and identifying knowledge gaps and also allow lake responses to future change, including management intervention, to be forecast (e.g. Reynolds et al. 2001).

**Fig. 21.2** Interactions between different scientific approaches and links to ecosystem management use in the study of lakes in the English Lake District [adapted with permission from Maberly and Elliott (2012)]



This chapter describes how these distinct and complementary approaches have been applied in the English Lake District; a case study in holistic ecology that integrates science from different disciplines. This information can be communicated to decision-makers responsible for the sustainable management of an essential human resource in the face of growing and multiplying stressors.

## 21.2 Methods and Data

The early years of the FBA were characterised by wide-ranging and careful development of novel equipment and methods to study fresh waters. Once established, however, the methods used to collect and analyse the data in the long-term monitoring programme were fairly conservative: there was a tendency to trade-off the use of the most modern methods in favour of using a method that may have been in use for decades, in order to preserve the continuity of the long-term records. For example, Talling (1993) first introduced measurement of phytoplankton chlorophyll *a* as a routine method in 1964 using extraction in boiling methanol and spectrophotometric analysis. The same method has been used since then, although other solvents (e.g. ethanol), and methods of analysis (e.g. fluorescence or high performance liquid chromatography) are commonly used elsewhere. In another example, the method for the analysis of nitrate, was changed from using phenol disulphonic acid to using the cadmium/copper hydrazine reduction technique after 1971 (Heaney et al. 1988). A long overlap where both methods were used in parallel allowed differences between the two methods to be

thoroughly characterised. The chemical methods of water analysis that had been developed and tested were standardised and described in a widely used booklet (Mackereth et al. 1978).

The methods used to sample Windermere's fish populations have also been subjected to a very conservative approach, with those employed to monitor its Arctic charr, perch and pike populations being rooted in the passive sampling equipment and single-species approaches of the 1940s; each conducted annually over periods of approximately 6 weeks. Specifically, Arctic charr are monitored by essentially non-destructive gill netting during the late autumn on a spawning ground in the lake's north basin which produces information on relative abundance, sizes, ages and the timing of spawning (Winfield et al. 2008a). Somewhat similarly, perch are sampled using a bespoke design of trap set in inshore areas of both basins during the spring which produces information on relative abundance, sizes, ages and the timing of spawning (Paxton et al. 2004). Finally, pike are sampled using a single mesh gill net in inshore areas of both basins during the late autumn which produces information on relative abundance, sizes, ages, fecundities and diet (Winfield et al. 2008b, 2012).

In addition to the continuation of the above long-term sampling and data sets, since the early 1990s fish research at Windermere has also expanded to include new approaches. This has included the use of recreational angler catch-per-unit-effort data for the Arctic charr populations of both basins of the lake (Winfield et al. 2008a) and the systematic use of survey gill nets to sample the entire fish community rather than only selected species. The latter has enabled monitoring not only of certain native species but also of introduced and expanding species such as roach (*Rutilus rutilus*) (Winfield et al. 2008a, b). Activities have also included the pioneering use of hydroacoustics to gather information on fish abundance and distribution (Winfield et al. 2007; Jones et al. 2008; Hateley et al. 2013) and, most recently, the exploration of environmental DNA techniques as an alternative and non-destructive means of determining fish species presence and relative abundance (Hänfling et al. 2016). Notably, this development and application of hydroacoustics and environmental DNA (eDNA) approaches was facilitated by the availability of unparalleled long-term background data on the lake's fish community derived from the established netting techniques begun in the 1940s.

High-frequency monitoring in the past was severely limited by the available technology. For example, Frempong (1983) made high-frequency measurements of temperature over depth for periods of a few days per campaign between 1977 and 1978, recording data on multi-channel chart recorders, with obvious limitations for data collection and analysis. Over a decade later, Davison et al. (1994) were able to measure pH at high-frequency in a stream flowing into a lake district lake because of advances in the technology of sensors (pH electrodes with an inbuilt amplifier overcoming problems of measuring a high impedance signal) and dataloggers. A similar system was used by Maberly (1996) to collect 15-min temperature and pH measurements in Esthwaite Water that started at the end of 1992 and has continued until the present. The increasing availability of sensors, more capable dataloggers,



software and computing and telecommunications has revolutionised high-frequency measurements in lakes (see Sect. 21.4).

There has also been a revolution, brought about by computing power, in the way data are stored and analysed. Early in the monitoring programme, field data and subsequent laboratory analyses were stored in notebooks, data tabulated and analysed using mental arithmetic, slide-rules or more recently electronic calculators, and graphs were drawn by hand. Starting in the mid 1980s, data were transcribed to a relational database; initially R:Base, latterly Oracle, by AE Irish. Short 4-letter codes were used to represent lakes, variables etc. and data were entered twice independently and validated to improve data quality. The ability to store, collate and analyse data electronically has greatly increased their value and usability. The current Oracle database serves as a long-term storage system, collating the different types of data using standardised styles, ensuring correct data formatting for future use.

Initially, the high-frequency data, such as the 15-min pH-data, were downloaded manually from a data-logger to a field computer and then uploaded into a Microsoft Excel spreadsheet. Later, data were downloaded automatically into a spreadsheet by telemetry over the General Packet Radio Service (GPRS) data network directly to the data server. However, as the volume of data increased because of the greater number of sensors and increased frequency of collection, storing data in spreadsheets started to become impractical. The increasing size of recorded high-frequency data made manipulation and validation tasks unwieldy, and so the data repository was transferred to an Oracle-based relational database. The raw high-frequency data are now automatically ingested into the database, where a suite of quality control (QC) checks are triggered to validate the data, followed by aggregation processes to create hourly and daily summaries. These data are then made available for browsing, download and analysis through a web-based client. The automation and scripting of data ingestion and QC tasks creates a known processing chain for data provenance, from the initial observations to making the data available to browse and download, that was not possible when using spreadsheets for storage.

While the above methods provide acceptable access and means of analysis at a local level, the use of proprietary storage and access designs can be a barrier to the sharing, integration, and analysis of data which is necessary when working at regional and larger scales with data from a number of different sources (e.g. Woolway et al. 2016). To provide standardised data representation and wider access, both for observations and metadata, ongoing research is evaluating a move to a system that employs the Sensor Web Enablement specifications (Conover et al. 2010), and the Semantic Sensor Network Ontology (Compton et al. 2012). These standards, and the technologies that build upon them, provide a foundation for creating systems that can automatically access, integrate, and analyse data from multiple sources, and provide scientists with straightforward data access across the data repositories that implement these standards. This provides another step towards broader data discovery, access, and integration of the high-frequency data that was initiated by moving from spreadsheets to a relational database, and will continue by moving towards international standards.

## 21.3 Case Studies of the Application of Long-Term Research to Generating Ecological Knowledge

The purpose of long-term research is partly to detect changes in the structure and function of the system being studied and to attribute causes of change. More than that, along with other activities (Fig. 21.2 and see Sects. 21.4–21.6), it produces ecological knowledge of how lake ecosystems function. Three case-study examples are given below describing how novel, and globally-relevant, ecological knowledge has been generated from long-term research in the English Lake District.

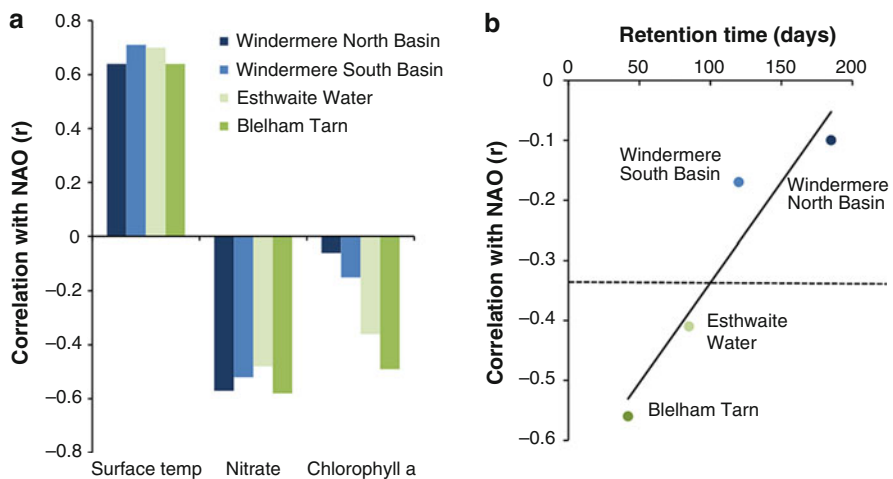
### 21.3.1 *Lake Teleconnexions*

To deliberately misquote the English metaphysical poet John Donne, ‘No lake is an island. . .’. This is obviously true for the input of energy, materials and propagules from the catchment and ‘air-shed’ but it is also, maybe more surprisingly, true of large scale processes seemingly operating at a distance. One of the first of these teleconnexions to be shown to affect lakes was the demonstration that the position of the North Wall of the Gulf Stream in the Western Atlantic affected the ecology of Windermere over 6000 km away (George and Taylor 1995). A Gulf Stream Index time-series, constructed from the position of the Gulf Stream, was strongly negatively correlated with interannual changes in the summer biomass of zooplankton in Windermere. The hypothesised mechanism behind the teleconnexion is that the Gulf Stream affects the strength of summer stratification which in turn controls the timing of edible phytoplankton which are in turn eaten by the zooplankton. This example demonstrates the sensitivity of lakes to physical forcing and also underlines the fact that bottom-up processes can have a profound effect on higher trophic levels.

In an example from Esthwaite Water, George (2002) showed that the summer phytoplankton biomass, as chlorophyll *a*, was also linked to the Gulf Stream Index. The suggested mechanism was the relationship between a negative Gulf Stream Index and strong summer winds causing a deeper early summer thermocline and greater entrainment of nutrients into the epilimnion from depth, which supports more phytoplankton growth.

The North Atlantic Oscillation (NAO) is a well-known weather pattern (Hurrell 1995) with derived records extending back to the 1860s based on sea-level air pressure at the Iceland Low and the Azores High. It controls winter weather in Europe by influencing the strength of winds blowing off the Atlantic. When the pressure difference is large between the Iceland Low and the Azores High (a positive NAO index), westerly winds from the Atlantic are strong, bringing mild, wet and windy weather in winter. A smaller pressure difference or a pressure reversal (a negative NAO index) produces cooler, drier and less windy weather in winter with more influence from air from the continent instead of the Atlantic.

George et al. (2004) analysed data from the winters (December to February) of 1961 to 1997. Four adjacent lakes in the English Lake District were studied: Windermere North Basin, Windermere South Basin, Esthwaite Water and Blelham Tarn, that had contrasting size, flushing time and productivity as well as long-term data. They showed that some features, such as winter water temperature, responded positively and coherently to a NAO index (George et al. 2004). The larger the NAO index, the warmer the water temperature (Fig. 21.3). In another example, the winter concentration of nitrate in the four lake basins was negatively correlated with the NAO index: higher concentrations were found in all four lakes when the NAO index was negative. Again the response was highly coherent (Fig. 21.3) and this was suggested to be caused by mild winter temperatures increasing the loss rate of nitrate from the catchment, via plant uptake and microbial denitrification, leading to higher concentrations in the lake. In contrast, some responses differed among the four lakes. So for example, winter chlorophyll *a* concentration was affected by the NAO index in the smaller, rapidly flushed lakes, but not in the larger less rapidly flushed lakes (Fig. 21.3). This non-coherent response appeared to be linked to the rainfall aspect of the NAO. High winter rainfall, associated with a positive NAO index, reduced winter concentrations of chlorophyll *a* in rapidly flushed lakes such as Blelham Tarn, but had little effect on less sensitive lakes with a longer retention time such as the North Basin of Windermere. Features such as the NAO can also affect higher trophic levels. Thus, for example, Elliott and colleagues (2000b) showed that the emergence of sea trout fry in a Lake District stream correlated with the NAO because of the influence of the NAO on water temperature.



**Fig. 21.3** The effect of the North Atlantic Oscillation on winter conditions in four contrasting lakes in the English Lake District [adapted with permission from George et al. (2004)]. (a) Correlation between winter water temperature, nitrate concentration and phytoplankton chlorophyll *a* and the North Atlantic Oscillation index. (b) The correlation of winter phytoplankton chlorophyll *a* and the North Atlantic Oscillation index as a function of mean lake retention time. The horizontal dashed line represents  $P = 0.05$

Where responses of lakes to a regional weather pattern are coherent, inter-annual variation in phenology and other responses, may also be coherent across large areas of the landscape or continent. However, some characteristics of a lake, especially those driven by their susceptibility to flushing, mean that lakes will vary in their sensitivity and potentially direction of response, to some features of the NAO and other weather patterns.

The NAO is one of a number of climate indices including the El Niño-Southern Oscillation (ENSO), the Arctic Oscillation (AO) and the Pacific Decadal Oscillation (PDO). Many of these indices are linked: thus the NAO and AO are closely related and the NAO and Gulf Stream position are also linked, albeit with a time lag so that the position of the Gulf Stream is correlated with the NAO 2 years earlier (George 2002). The position and strength of the jet stream at around 12 km in the troposphere above the Earth's surface also has a major effect on our weather and is potentially linked to regional climate indices such as the NAO or AO. Rossby Wave Breaking (Thorncroft et al. 1993) associated with the jet stream provides a potential mechanism that links processes in the atmosphere with conditions on the Earth's surface. Strong and Maberly (2011) showed that there was a strong correlation between the frequency of cyclonic and anticyclonic Rossby Wave Breaking and the surface temperature of lakes in the English Lake District. Rossby Wave Breaking explained between 54 and 69% of the interannual variation in lake temperature in the four seasons. The all-year-round effect of Rossby Wave Breaking contrasts with the NAO which is largely a winter phenomenon, although it can have longer-lasting effects in some lakes operating through a 'memory' within the food web (Straile 2000).

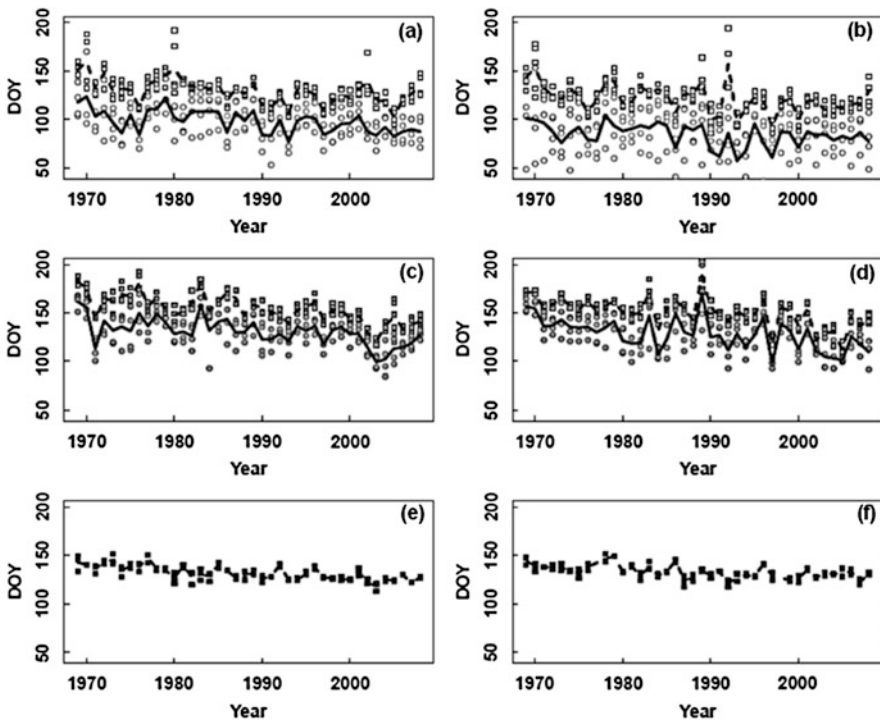
Long-term research in the English Lake District has led, therefore, to the appreciation that large scale processes such as the Gulf Stream, NAO and Rossby Waves control the structure and function of lakes and can act to produce coherent temporal responses for some properties over large spatial scales. Truly, 'no lake is an island'.

### **21.3.2 Phenology**

Shifts in the timing of the biological seasons provide some of the most compelling evidence that climate change is already having a discernible effect upon ecosystems (IPCC 2014). These changes are only apparent as a result of ongoing long-term monitoring schemes that continue to generate important biological data from ecologically-diverse systems across the world. Long-term research has revealed that species and populations vary greatly in the extent to which their seasonal activities have shifted, with possible consequences for ecological interactions and ecosystem functioning (Thackeray et al. 2010).

Generally, fresh waters have been under-represented in global assessments of phenological change and this has limited our ability to make projections and inferences regarding ecological impacts in these systems. Long-term research

from the English Lake District has made a major contribution to filling this important knowledge gap. The availability of multi-decadal time series data on taxa fulfilling various ecological roles, has allowed the detection of seasonal shifts in the timing of plankton population growth and fish spawning (Fig. 21.4), the identification of potential drivers of these changes, and assessments of possible ecosystem impacts. This research has shown that the seasonal growth period of the phytoplankton primary producers has changed over the long-term, in parallel with similar changes in terrestrial plants, but that these changes vary greatly among species (Thackeray et al. 2008; Meis et al. 2009; Feuchtmayr et al. 2012). This work challenged the assumption that water temperature is the sole influential driver of seasonal change, showing that varying nutrient availability over time can also have an important influence upon seasonality. Subsequent work has shown that seasonal shifts are not restricted to the phytoplankton but are also apparent at higher levels in the food web, for zooplankton grazers and fish (Thackeray et al. 2012, 2013;



**Fig. 21.4** Long-term changes in the seasonal timing of phytoplankton (a, b) and zooplankton (c, d) spring population growth, and perch spawning (e, f). Data are shown for the North (a, c, e) and South (b, d, f) basins of Windermere. *Points* show the original phenological event data and *lines* show average seasonal timings for distinct event classes. In plots a–d, *solid lines* show the mean seasonal timing of time-of-onset type metrics (*circles*), and *dashed lines* show the mean seasonal timing of time-of-peak/middle type metrics (*square symbols*). For the perch data, only peak/middle type metrics were calculated [reproduced with permission from Thackeray et al. (2013)]

Ohlberger et al. 2014) (Fig. 21.4). These studies have demonstrated that the extent to which seasonal shifts de-synchronise species interactions is highly variable; with stronger evidence of disrupted interactions affecting fish than affecting zooplankton grazers. The whole-system ethos underlying the monitoring of these sentinel systems has been central to our ability to detect physical and chemical drivers of change, as well as impacts felt at the ecosystem scale.

Data from long-term research on the Lakes in the English Lake District have also been used to explore, and refine, the way in which ecological information from lakes is processed to make inferences about phenological change. Often, the seasonal timing of biological events can be mathematically described in a variety of different ways e.g. the timing of peak abundance, or of populations exceeding a threshold. The choice of phenological metric can potentially affect estimates of changing seasonal timing, and also observed relationships with potential environmental drivers (Thackeray et al. 2012, 2013). Using our long-term data we have advocated to the wider research community that the choice of phenological metric, or indicator, must be made carefully in relation to research questions and objectives each time environmental data are to be analysed.

Phenological research in the English Lake District also illustrates the way in which expectations based upon theoretical considerations can be verified against “real world” observational data. Prior to the time series analyses cited above, Reynolds (1990, 1997) used prior knowledge on the dynamics of phytoplankton populations, accrued over years of careful experimentation and independent observation, to construct a theory of the impacts of shifting nutrient availability upon phytoplankton phenology. Later analyses then verified these expectations against observed ecological phenomena. Furthermore, using the independently-developed phytoplankton community model PROTECH (Sect. 21.5), it has been possible to re-create phenological responses observed in the field, in a virtual environment (Elliott et al. 2006). The results of long-term research in the English Lake District have also generated new expectations on the likely importance of phenological change in plankton communities (Thackeray et al. 2012), which are already being confronted with observational data from other systems (Atkinson et al. 2015).

### ***21.3.3 Testing Ecological Theory***

The duration and quality of the Windermere long-term data sets offer unique opportunities to test ecological theories at spatial and temporal scales which are otherwise unachievable. This is particularly the case for fish populations, for which typically large home ranges and great individual longevities pose particular challenges to evolutionary and population ecologists. Here, illustrative examples are given for tests of evolutionary ecology utilising the Windermere pike data and test of population biology using the Windermere perch data. In each case, the tests use data collected over many decades and in both basins of the lake and cover the lake’s major piscivore and planktivore, respectively.

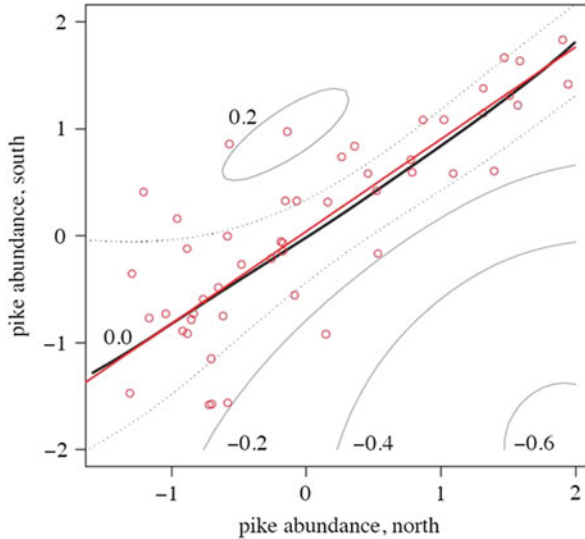
In evolutionary ecology, the ideal free distribution (IFD) theory originally developed by Fretwell and Lucas (1969) predicts that mobile animals living within an environment containing patches of contrasting quality will, if certain assumptions are met, distribute themselves such that mean individual fitness amongst those patches will be equalised. Originally proposed for birds, IFD quickly established itself as one of the most influential theories in evolutionary ecology and has been applied to a wide range of mobile taxa. However, most aquatic tests of IFD have been in the context of individuals foraging over relatively short time periods (e.g. Lampert et al. 2003; Shepherd and Litvak 2004). Furthermore, few of these or terrestrial studies have compared predicted with observed distributions and those that have done so have based their predictions on theory alone. To the best of our knowledge, large-scale studies using observed *a priori* knowledge of the relationship between fitness and population density were absent.

In the above context, Haugen et al. (2006) explored the movements of pike between the North and South basins of Windermere using 40 years of capture-mark-recapture tagging data. Fitness functions were derived, specific to pike population density and to basin, incorporating probabilities of survival and dispersal together with fecundity estimates. These descriptors were then used together with IFD theory to predict the changing distribution of pike between the two basins of the lake in response to changing conditions, with the intersection of the fitness surfaces for the two basins used to derive expected spatial distributions. In addition, these model-based predictions were then compared with observed multi-decade spatial distributions, which included an experimental manipulation of basin-specific pike population density carried out between 1956 and 1962.

Comparison of the spatial distributions predicted by IFD with those actually observed revealed a remarkably high degree of agreement (Fig. 21.5) and demonstrated that pike is ideal free distributed between Windermere's two basins. Over the study period as a whole, there was a net migration from the less productive North Basin to the more eutrophic South Basin. However, the experimental manipulation of pike population density in the late 1950s and early 1960s switched the net migration direction, further demonstrating that pike in Windermere choose their habitat in accordance with IFD theory. As remarked by Haugen et al. (2006), such a test of IFD theory had not been undertaken before on such a large field scale in aquatic or terrestrial systems and so this study has application beyond the understanding of pike movements within Windermere.

Fish have long been used as model species in studies of the dynamics of stage-structured populations and the plasticity of perch in this context makes this species particularly utilitarian (e.g. ten Brink et al. 2015). Recently developed theoretical models of stage-structured consumer–resource systems have shown that life-stage-specific biomass overcompensation (i.e. an increase in stage-specific biomass over that of pre-disturbance conditions) can arise in response to increased mortality rates if these release the surviving individuals from competition. If growth, maturation, and/or reproduction are food-dependent processes, as they have clearly been shown to be for perch in Windermere (Craig et al. 2015), then this indirect density-dependent effect may lead to higher growth rates, faster maturation, and/or



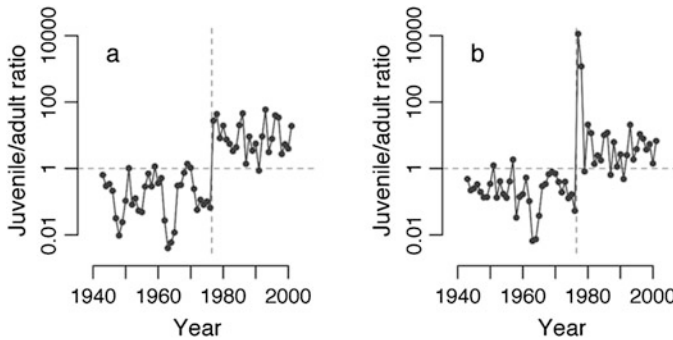


**Fig. 21.5** The isodar (i.e. the intersection line between fitness surfaces for the north and south basins, *thick black line*) and the corresponding 95% confidence boundary lines (*dotted lines*) predicted for Windermere pike by the ideal free distribution theory. *Numbers attached to grey lines* represent isocline values for the difference between basins of the estimated realised fitness values, where the zero isocline constitutes the predicted isodar. The *red open dots* represent annual pike abundance estimates over a 50 years period and the *red line* is the estimated linear isodar for these abundances. [Redrawn with permission from Haugen et al. (2006)]

increased adult fecundity. However, as for IFD theory, tests of such models are extremely rare in the field and have been largely confined to relatively short-term laboratory studies.

The long-term Windermere perch data were used by Ohlberger et al. (2011) to parameterise a stage-structured population model simulating the effects of increased adult mortality caused by a disease outbreak in 1976. Remarkably, this outbreak had a much higher prevalence among adult, compared to juvenile, individuals and was subsequently estimated to have killed 98% of adult perch in the lake (Bucke et al. 1979). The model predicted biomass overcompensation by juveniles in response to increased adult mortality caused by a shift in food-dependent growth and reproduction rates. The addition to the model of cannibalism between these life stages reinforced this compensatory response caused by the release of juveniles from intraspecific predation at high adult mortality rates. Model predictions were strongly supported by observations, revealing that the disease outbreak induced a strong decrease in adult biomass and a corresponding increase in juvenile biomass (Fig. 21.6). Age-specific adult fecundity and size-at-age were both higher after the disease outbreak, suggesting that the disease-induced mortality released adult perch from competition and so increased their somatic and reproductive growth. Higher juvenile survival after the disease outbreak caused by





**Fig. 21.6** Time series of the ratio between juvenile and adult perch biomass in the (a) North and (b) South Basins of Windermere, showing a transition of the biomass distribution from being dominated by adults (ratio  $<1$ ) to being dominated by juveniles (ratio  $>1$ ) following a major disease outbreak in 1976 (vertical dashed line). [Redrawn with permission from Ohlberger et al. (2011)]

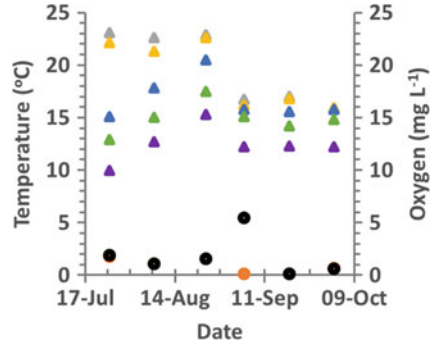
a release from cannibalism probably contributed to the observed biomass overcompensation. These findings have general implications for predicting population- and community-level responses to increased size-selective mortality caused by disease outbreaks or similarly size-specific impacts such as those arising from exploitation by fisheries or other forms of harvest.

## 21.4 Automatic High-Frequency Monitoring

Long-term monitoring of lakes provides an invaluable resource for understanding the ecosystem, but monthly, fortnightly or even weekly monitoring cannot capture time-scales at which meteorological drivers and ecosystem processes act. In the past this was a limitation which was difficult to overcome, but the recent advancements in sensor technology, computing power and communications has enabled a new branch of ecosystem monitoring to develop: automated, *in situ*, and high-frequency. An example is the deployment in 2006 of a monitoring buoy in Blelham Tarn, a small, but reasonably deep (surface area, 0.1 km<sup>2</sup>; maximum depth, 14.5 m), eutrophic lake in the English Lake District, which has since provided a suite of meteorological and limnological data every 4 min to supplement the traditional fortnightly monitoring which has been undertaken on this lake for decades. The sizable floating buoy, securely anchored to the lake bed, is sufficiently large to provide a stable platform for a meteorological station including instruments to measure solar radiation, air temperature, wind speed and relative humidity, with a chain of 12 temperature sensors hanging from 0.5 m below the surface of the lake to 12.0 m depth.

The benefit of these high-frequency data were highlighted by Jennings et al. (2012) in a study of episodic events taking place in different lakes across the world.

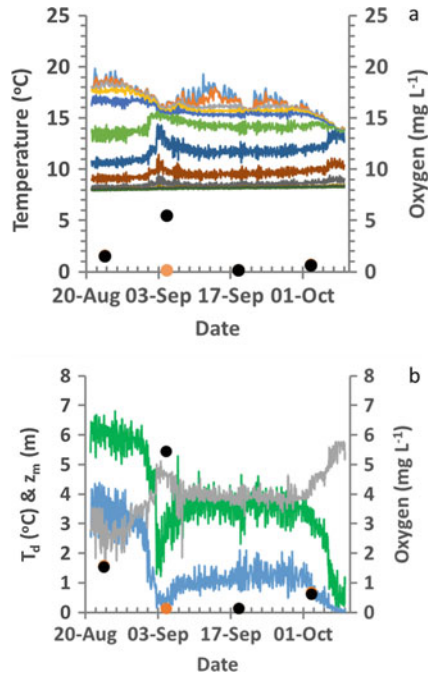
**Fig. 21.7** Oxygen concentration (circles) at 5 m (black) and 6 m (orange) and temperatures (triangles) at 0 m (grey), 2 m (yellow), 4 m (blue), 5 m (green) and 6 m (purple) manually collected every 2 weeks in Blelham Tarn in 2006. Note oxygen at 5 m and 6 m is virtually identical most of the time



For Blelham Tarn, the long-term research has shown oxygen depletion at depth to be a major problem, linked to thermal stratification restricting the supply of oxygen from the surface to depth (Foley et al. 2012). However, the precise meteorological and lake physical factors controlling oxygen depletion could not easily be elucidated using the long-term data. Analysis of the fortnightly collected data from 2006 showed that, by late summer, the deep water of the lake had become virtually anoxic, with oxygen concentrations below  $1 \text{ mg L}^{-1}$  at depths as shallow as 5 m. In early September, though, this fortnightly record showed substantial re-oxygenation at 5 m, though the lake clearly remained anoxic just a metre below. Analysis of the temperature profiles taken during the field sampling suggested a mixing event may have occurred bringing surface oxygen back down to 5 m, but little other information could be gleaned (Fig. 21.7). When did this mixing occur? How long did it last? What caused the mixing? Why did an apparently similar level of mixing not re-oxygenate that depth a month later?

Fortunately, the monitoring buoy that had been installed earlier in the summer, provided the means to answer these questions. Just replacing the fortnightly-resolution temperature series with the hourly-averaged temperature data collected from the buoy (Fig. 21.8) immediately indicated that a noticeable mixing event had, indeed, occurred in the lake. Its effects were felt down to 8 m in the water, but the event took place a couple of days before the field sampling. The high-frequency data also demonstrated that stratification re-asserted itself shortly after the fortnightly data had been collected.

To obtain more detailed answers to our ‘when’, ‘how’, ‘what’ and ‘why’ questions there is another obstacle to overcome. The advantages of high-frequency data collection are obvious, but a subtle disadvantage is that it is too easy to become overwhelmed by the sheer weight of the data. One data point every 4 min means that in 2 h of data collection there are more data to analyse than provided in a whole year of manual sampling every 2 weeks. Dealing with these data therefore demands the use of efficient and systematic analysing techniques. To this end, scientists from GLEON (The Global Lake Ecological Observatory Network), a grassroots organisation of limnologists working with high-frequency *in situ* data, and NETLAKE, an European Union Cost Action aimed at the European buoy-user community, have

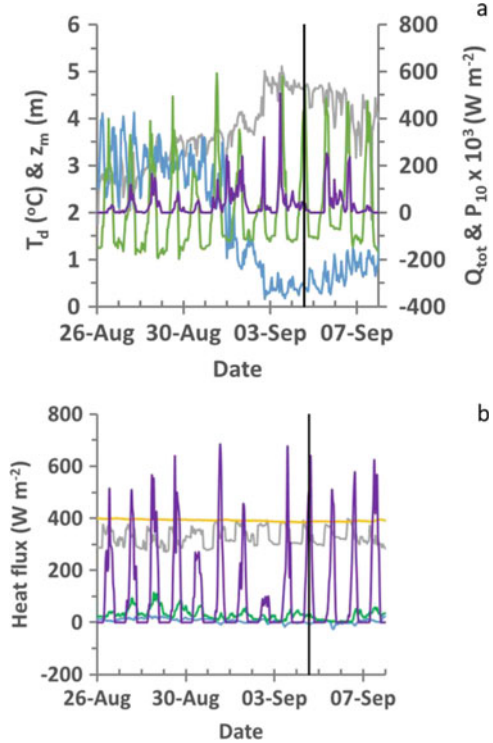


**Fig. 21.8** Measurements of oxygen concentration and temperature in Blenheim Tarn in 2006. (a) Manually collected oxygen concentrations (circles) at 5 m (black) and 6 m (orange) and temperatures at 0.5 (light blue), 1 (orange), 2 (light grey), 3 (gold), 4 (blue), 5 (green), 6 (dark blue), 7 (dark red), 8 (dark grey), 9 (mustard), 10 (indigo) and 12 m (dark green) measured from the buoy; (b) Oxygen concentration (circles) at 5 m (black) and 6 m (orange), temperature difference,  $T_d$ , between 4 m and 5 m (blue) and between 4 m and 6 m (green) and mixed depth,  $z_m$  (grey). Note in both panels that oxygen at 5 m and 6 m is virtually identical most of the time

collaborated to develop user-friendly software to facilitate analysis of data typically collected from this type of monitoring buoy. One such piece of software, Lake Analyzer (Read et al. 2011), was written specifically to calculate useful lake physics parameters from high-frequency temperature profiles and meteorological data (Global Lakes Ecological Observatory Network, Lake Analyzer. <http://www.gleon.org/research/projects/lake-analyzer>).

Applying Lake Analyzer to the data from Blenheim Tarn enabled the mixed depth, here defined as the first depth in the interpolated temperature profile for which the density gradient exceeded  $0.1 \text{ kg m}^{-3} \text{ m}^{-1}$ , to be readily calculated. Combining this information with the calculated temperature differences between 4 m and 5 m, which became oxygenated, and 4 m and 6 m, which stayed deoxygenated, allowed the mixing event to be understood (Fig. 21.9). While surface temperatures had been cooling and stratification weakening for several weeks, the substantive mixing took place on 1st and 2nd September; the mixed layer deepened during this time from 3 to 5 m, entraining the 5 m monitoring depth into the oxygenated epilimnion, the temperature difference between 4 m and 5 m falling

**Fig. 21.9** Temperature and heat fluxes in Blelham Tarn in 2006. **(a)** Temperature difference between 4 m and 5 m,  $T_d$  (blue), mixed depth,  $z_m$  (grey), total heat flux,  $Q_{tot}$  (green), and scaled turbulent wind energy flux,  $P_{10}$  (purple). Time of manual sampling marked with a black line; **(b)** Surface heat fluxes: long-wave out (yellow), long-wave in (grey), net solar radiation (purple), sensible heat flux (blue) and latent heat flux (green). Time of manual sampling marked with a black line. Note long-wave in and solar designated positive if a downwards flux (lake heating), while long-wave out, sensible heat and latent heat designated positive if an upward flux (lake cooling)



to a fraction of a degree in the process. While some effects of mixing were clearly felt at 6 m, nevertheless there remained a non-negligible temperature, and therefore density, gradient above that depth, providing sufficient resistance to prevent the penetration of oxygenated water. In fact, by the time of field sampling, around mid-morning on 4th September, the stratification was already beginning to recover, and within a couple of days the 5 m monitoring depth had become detraind from the epilimnion and the process of deoxygenation had resumed. It also became clear that when the site was monitored on 2nd October the 5 m depth was still just beneath the mixed layer and, though the stratification was severely weakened, the density gradient above 5 m was still larger than it had been during the early September mixing event. Thus, entrainment was beginning and re-oxygenation imminent, but had not yet occurred.

A further GLEON/NETLAKE software development was Lake Heat Flux Analyzer (Woolway et al. 2015), which facilitates the somewhat complicated calculations of atmospheric surface fluxes from the high-frequency data readily collected on *in situ* monitoring buoys (Global Lakes Ecological Observatory Network. HeatFluxAnalyzer Web. <http://heatfluxanalyzer.gleon.org/>). This software enabled the calculation of the various surface fluxes driving the cycles of heating and cooling in a lake, and the terms necessary for calculating the vertical turbulent wind energy flux (Wuest et al. 2000). Highlighting the fortnight around the mixing

event it can be seen (Fig. 21.9) that it was sustained wind mixing which initially drove the de-stratification on 1st September, but on 2nd September the wind was mostly quiescent except for a couple of hours. It was, however, clearly an unusually cool day with the day-time heating being substantially lower than for the rest of the period. This low heating was insufficient to provide resistance to just 2 h of very strong winds that afternoon which quickly deepened the mixed layer by a whole metre, re-oxygenating the 5 m layer in the water column. In fact, the wind provided even greater mixing energy the next day, but as the heating had returned to more typical values those winds were not sufficient to deepen the mixed layer further.

More information still can be obtained. The total heat flux is a combination of different heating and cooling processes: upward and downward long-wave radiation, evaporative cooling, the sensible heat flux driven by air-water temperature differences, and solar radiation. Investigation of these individual fluxes (Fig. 21.9) calculated using Lake Heat Flux Analyzer, showed that the unusually low heating on 2nd September was a result of depressed solar radiation: it was evidently a very cloudy day. Thus, a full pattern finally emerges. The lake was cooling towards the end of August, and strong winds on 1st September weakened the stratification further, allowing the combination of exceedingly overcast conditions and a couple of hours of high winds on 2nd September finally to open up the 5 m layer to oxygenation. Very strong winds the following day, accompanied, as they were, by a more typical level of diurnal heating for the season, proved insufficient to force the epilimnion any deeper. By 4th September, when the manual monitoring took place, the stratification was already recovering and the deoxygenation process beginning again, remaining unchecked until, coincidentally, just after the field sampling a month later.

Monitoring buoys, such as the one on Blelham Tarn, are now proliferating in lakes across the globe, adding substantially to data collected from traditional field sampling. New types of sensors are increasingly becoming available, affordable and sufficiently reliable and stable to be left automatically monitoring in a lake. The further development of analysis software, real-time telemetry providing the means for real-time web visualisation, processing and forecasting, and the willingness of scientists to collaborate and take part in multi-lake studies using high-frequency data (e.g. Solomon et al. 2013; Woolway et al. 2016) give reasons to be optimistic of the future potential for a profusion of scientific breakthroughs using these complementary techniques.

## 21.5 Modeling

### 21.5.1 *Benefits of Modeling*

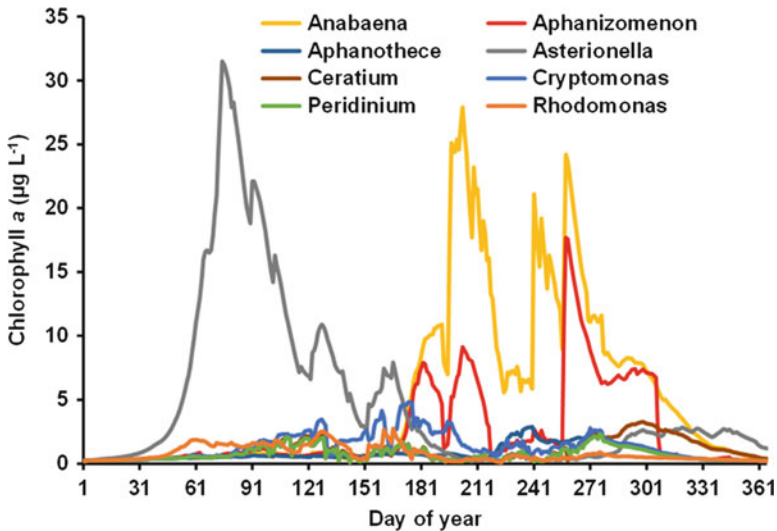
Long-term research greatly benefits from the development and application of models. Models allow researchers to explore concepts and test hypotheses related

to their observations. They can also facilitate the application of knowledge to other systems different from those being studied. Ultimately, they are a simplified expression of quantified understanding gained through the research. Here, we follow the example of such a model and illustrate its practical application within the English Lake District.

### 21.5.2 *The PROTECH model*

PROTECH [Phytoplankton RespOnses To Environmental CHange; (Elliott and Reynolds 2010; Reynolds et al. 2012)] is a process-based lake phytoplankton community model and is well established in its field (Trolle et al. 2012). It was developed from a combination of laboratory experiments (Reynolds 1989) and the Blelham Tarn field experiments (e.g. Reynolds et al. 1982, 1983). It predicts phytoplankton species growth in daily time steps at different depths and responds to changes in temperature, light and nutrient availability. A particular strength is its ability to model different phytoplankton species using morphological relationships from Reynolds (1989) (Fig. 21.10).

Of course, testing such a model was greatly facilitated by the availability of the long-term data and has resulted in PROTECH being applied to many of the lakes in the English Lake District with these data [e.g. Bassenthwaite Lake (Elliott et al. 2006), Blelham Tarn (Elliott et al. 2000a), Esthwaite Water (Elliott 2010) and Windermere (Elliott 2012)]. It has also been applied locally at sites with less



**Fig. 21.10** Example PROTECH simulation of eight different phytoplankton in Esthwaite Water in 2003 [After Elliott (2010)]

extensive datasets such as at Loweswater (Norton et al. 2012), Wastwater (Elliott and Thackeray 2004) and Ullswater (Bernhardt et al. 2008). These examples cover a range of lake types from oligotrophic to eutrophic and from shallow to deep (14.5–76 m). Such diversity and wide boundary ranges in typology provides a rich source of data for testing that is vital for model development.

### 21.5.3 Hypothesis Testing

After sufficient assessment, which builds up confidence in a model, the model can be a useful tool to test hypotheses or “what if” scenarios. A good example from the English Lake District is that of the vendace [*Coregonus albula* (L.)] in Bassenthwaite Lake. Vendace are a rare and protected fish species in the UK and, today, one of the few natural populations occurs in Bassenthwaite Lake. They are sensitive to warm water temperatures and to low oxygen concentrations and an analysis of observed data from 1990–99 suggested that the available habitat for the fish was sometimes constricted by increased water temperatures and decrease oxygen concentrations (George et al. 2006). George et al. (2006) went further and suggested that these factors restricting vendace habitat would be further enhanced by projected climate change.

In order to test this hypothesis, PROTECH was coupled to a lake oxygen model, LOX (Bell et al. 2006) and driven by 20 years of daily weather data (representing the last two decades of the twenty-first century) from a Regional Climate Model (RCM) to simulate vendace habitat volume in Bassenthwaite Lake (Elliott and Bell 2011). The results showed a forecast mean increase in water temperature of  $>2$  °C. In contrast, there was  $<10\%$  decrease in oxygen concentration under the future climate because, although the timing of phytoplankton growth was affected by climate change, total algal biomass and hence the amount of carbon reaching the hypolimnion was unaltered because nutrient loads to the lake were unchanged. Unfortunately for the vendace, the temperature increase alone was sufficient to see a marked decrease in available habitat volume: in all 20 future years, there were periods of  $> 7$  days where there was no suitable habitat available and in 16 years there were periods of  $> 20$  consecutive days when no habitat was available.

The example of Elliott and Bell (2011) shows the power of models in exploring hypotheses and asking “what if” questions through the combination of observed data recording the patterns of change and threats to the species; the application of the model allowed an independent assessment to be made of those threats. It also allows for expansion into the unknown, to assess quantitatively and to predict how future changes might impact upon species. Such combinations of knowledge remain powerful scientific tools and will inevitably help our understanding of complex environmental systems.

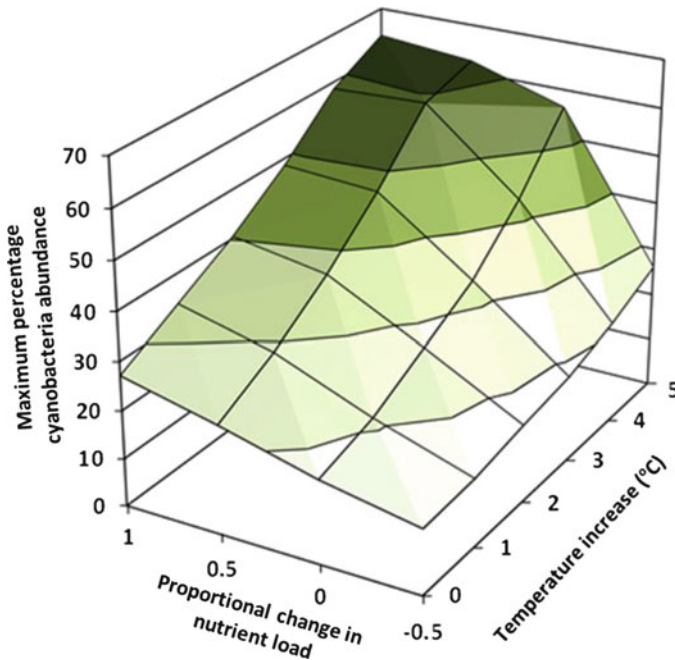


### 21.5.4 Models for Management

Finally, beyond the scientific benefits of such a tool, models can also be applied to help with management issues that affect lakes by forecasting the effects of different scenarios of change on particular aspects of lake condition. For example, in another application of PROTECH to Bassenthwaite Lake, the relative effects of changing nutrient load and water temperature upon cyanobacteria blooms were assessed (Elliott et al. 2006). The study showed how combined increases in nutrient load and water temperature both enhanced the size of cyanobacteria blooms (Fig. 21.11). However, it also demonstrated that while increased water temperature did increase bloom size, this change was greatly reduced if the nutrient supply to the lake was reduced. Therefore, lake managers could reduce the impact of a global climate pressure upon lakes though local management changes by striving to reduce nutrient loads.

## 21.6 Conclusions and Prospects

An ecosystem approach is needed in order to understand fully how a system functions and responds to past and future environmental perturbation because external pressures trigger a complex series of bottom-up and top-down interactions



**Fig. 21.11** The maximum annual percentage abundance of cyanobacteria in the simulated phytoplankton of Bassenthwaite Lake with changing nutrient load and water temperature [After Elliott et al. (2006)]



that can affect all aspects of a system (Maberly and Elliott 2012). In a long-term programme, the data that are produced should be able to detect and attribute the effects, not only known pressures, but also new types of perturbation that are often the result of unexpected consequences resulting from Man's activities. An excellent example of this is the routine measurements that led to the unexpected discovery of the Antarctic ozone hole (Farman et al. 1985).

Lakes are particularly sensitive sentinel systems because they integrate responses from the land and the atmosphere (Williamson et al. 2008). New technology is increasing the scale at which processes can be detected and understood. For example, the high-frequency technology allows the full dynamics of lakes to be captured. A second development that represents a huge opportunity for the study of lakes is modern Earth Observation platforms. These can produce valuable synoptic scientific information, albeit of the surface water, from a much larger range of lakes than can be measured by traditional means. The value of this approach is growing with the proliferation of systems with greater spatial resolution, allowing smaller lakes to be studied, and a greater temporal coverage, potentially allowing more frequent retrieval of data. This, in conjunction with global datasets on meteorology, geography, land-use and human populations and land-use practices, some of which also are produced by satellite, confer a powerful new opportunity to study and manage lakes. This approach is being used by the UK project *GloboLakes* ([www.Globolakes.ac.uk](http://www.Globolakes.ac.uk)). A third growing approach is the use of citizen scientists. These can increase greatly the amount of data collected after appropriate training. In the biodiversity field, this has been used very successfully to map changes in species distribution (Roy et al. 2015). In limnology they have also been successfully used to collect more data, more frequently, than can be afforded by more traditional means (Canfield et al. 2002; Lottig et al. 2014).

In the current climate of austerity, long-term research programmes are under increasing threat (Birkhead 2014). Long-term monitoring is invaluable for many reasons but often, wrongly, regarded as a 'Cinderella science' (Nisbet 2007). This threat is occurring internationally with, for example, threats of closure to the Canadian Experimental Lakes Area, which has produced high-quality research into how lakes respond to perturbation over 45 years, that were eventually resolved. These issues tend to be cyclical and have been faced before. Embedding long-term research in other activities may be one way of protecting their funding as it helps to increase their relevance further. Also, the value of long-term data is greatly increased when they are re-used by other scientists in different ways. While undoubtedly desirable this creates issues that need to be handled thoughtfully. For example, how is appropriate credit given to the scientists who collect the data so that there is an incentive for them to continue the collection in the first place? Furthermore, involving the original team is often necessary because they will have important insights into the data and the study system that are nearly impossible to capture in any metadata. One response to this is multi-author papers where the data contributors are recognised for their contribution and also have the opportunity to contribute their local expert knowledge and intellectual input. For example, the recent paper by Woolway et al. (2016) collated and analysed data from 100 lakes

around the world and had 28 authors on the paper. There is not yet a consensus on the best way forward on this and consequently it is an area of current hot debate (Mills et al. 2015, 2016; Whitlock et al. 2016). Nevertheless, the opportunities provided by growing internationalisation of science collaboration are transforming our global understanding of lake ecology and should be encouraged.

**Acknowledgements** We are grateful to the many individuals (too numerous to name here) who have participated in collecting the data and to the Freshwater Biological Association for its invaluable historical role in the production of the early long-term data. We dedicate this chapter to the late JWG Lund (1912–2015) and ED Le Cren (1922–2011) who were instrumental in starting much of the long-term research in Cumbria. This research, carried out today by CEH, is supported by the UK Natural Environment Research Council.

## References

- Allen KR (1935) The food and migration of the perch (*Perca fluviatilis*) in Windermere. *J Anim Ecol* 4:264–273
- Atkinson A, Harmer RA, Widdicombe CE et al (2015) Questioning the role of phenology shifts and trophic mismatching in a planktonic food web. *Prog Oceanogr* 137:498–512
- Bell VA, George DG, Moore RJ et al (2006) Using a 1-D mixing model to simulate the vertical flux of heat and oxygen in a lake subject to episodic mixing. *Ecol Model* 190:41–54
- Bernhardt J, Elliott JA, Jones ID (2008) Modelling the effects on phytoplankton communities of changing mixed depth and background extinction coefficient on three contrasting lakes in the English Lake District. *Freshw Biol* 53:2573–2586
- Birkhead T (2014) Stormy outlook for long-term ecology studies. *Nature* 514:405–405
- Bucke D, Cawley G, Craig J et al (1979) Further studies of an epizootic of perch *Perca fluviatilis* L., of uncertain aetiology. *J Fish Dis* 2:297–311
- Canfield DE, Brown CD, Bachmann RW et al (2002) Volunteer lake monitoring: testing the reliability of data collected by the Florida LAKEWATCH program. *Lake Reservoir Manage* 18:1–9
- Compton M, Barnaghi P, Bermudez L et al (2012) The SSN ontology of the W3C semantic sensor network incubator group. *J Web Semant* 17:25–32
- Conover H, Berthiau G, Botts M et al (2010) Using sensor web protocols for environmental data acquisition and management. *Ecol Inform* 5:32–41
- Craig JF, Fletcher JM, Winfield IJ (2015) Insights into percid population and community biology and ecology from a 70 year (1943 to 2013) study of perch *Perca fluviatilis* in Windermere, U.K. In: Couture P, Pyle G (eds) *Biology of Perch*. CRC Press, Boca Raton, pp 148–166
- Davison W, Hill M, Woof C et al (1994) Continuous measurement of stream pH. Evaluation of procedures and comparison of resulting hydrogen ion budgets with those from flow-weighted integrating samplers. *Water Res* 28:161–170
- Elliott JA (2010) The seasonal sensitivity of Cyanobacteria and other phytoplankton to changes in flushing rate and water temperature. *Glob Chang Biol* 16:864–876
- Elliott JA (2012) Predicting the impact of changing nutrient load and temperature on the phytoplankton of England's largest lake, Windermere. *Freshw Biol* 57:400–413
- Elliott JA, Bell VA (2011) Predicting the potential long-term influence of climate change on vendace (*Coregonus albula*) habitat in Bassenthwaite Lake, U.K. *Freshw Biol* 56:395–405
- Elliott JA, Reynolds CS (2010) Modelling phytoplankton dynamics in freshwaters: affirmation of the PROTECH approach to simulation. *Freshw Rev* 3:75–96

- Elliott JA, Thackeray SJ (2004) The simulation of phytoplankton in shallow and deep lakes using PROTECH. *Ecol Model* 178:357–369
- Elliott JA, Irish AE, Reynolds CS et al (2000a) Modelling freshwater phytoplankton communities: an exercise in validation. *Ecol Model* 128:19–26
- Elliott JM, Hurley MA, Maberly SC (2000b) The emergence period of sea trout fry in a Lake District stream correlates with the North Atlantic Oscillation. *J Fish Biol* 56:208–210
- Elliott JA, Jones ID, Thackeray SJ (2006) Testing the sensitivity of phytoplankton communities to changes in water temperature and nutrient load, in a temperate lake. *Hydrobiologia* 559:401–411
- Farman JC, Gardiner BG, Shanklin JD (1985) Large losses of total ozone in Antarctica reveal seasonal ClO<sub>x</sub>/NO<sub>x</sub> interaction. *Nature* 315:207–210
- Feuchtmayr H, Thackeray SJ, Jones ID et al (2012) Spring phytoplankton phenology - are patterns and drivers of change consistent among lakes in the same climatological region? *Freshw Biol* 57:331–344
- Foley B, Jones ID, Maberly SC et al (2012) Long-term changes in oxygen depletion in a small temperate lake: effects of climate change and eutrophication. *Freshw Biol* 57:278–289
- Frempong E (1983) Diel aspects of the thermal structure and energy budget of a small English lake. *Freshw Biol* 13:89–102
- Fretwell SD, Lucas HLJ (1969) On territorial behavior and other factors influencing habitat distribution in birds. 1. Theoretical development. *Acta Biotheor* 19:16–36
- George DG (2002) Regional-scale influences on the long-term dynamics of lake plankton. In: PJIW, Thomas DN, Reynolds CS (eds) *Phytoplankton productivity: carbon assimilation in marine and freshwater ecosystems*. Blackwell Science, Oxford, pp 265–290
- George DG, Taylor AH (1995) UK lake plankton and the Gulf Stream. *Nature* 378:139–139
- George DG, Maberly SC, Hewitt DP (2004) The influence of the North Atlantic Oscillation on the physical, chemical and biological characteristics of four lakes in the English Lake District. *Freshw Biol* 49:760–774
- George DG, Bell VA, Parker J et al (2006) Using a 1-D mixing model to assess the potential impact of year-to-year changes in weather on the habitat of vendace (*Coregonus albula*) in Bassenthwaite Lake, Cumbria. *Freshw Biol* 51:1407–1416
- Hänfling B, Lawson Handley L, Read DS et al (2016) Environmental DNA metabarcoding of lake fish communities reflects long-term data from established survey methods. *Mol Ecol* 25:3101–3119
- Hateley J, Claburn P, Drastik V et al (2013) Standardisation of hydroacoustic techniques for fish in fresh waters. In: Papadakis JS, Bjorno L (eds) *Proceedings of the first underwater acoustics conference*. FORTH Institute of Applied & Computational Mathematics, Heraklin, Greece, pp 1595–1600
- Haugen TO, Winfield IJ, Vollestad LA et al (2006) The ideal free pike: 50 years of fitness-maximizing dispersal in Windermere. *Proc R Soc B Biol Sci* 273:2917–2924
- Heaney SI, Lund JWG, Canter HM et al (1988) Population dynamics of *Ceratium* spp. in three English lakes, 1945–1985. *Hydrobiologia* 161:133–148
- Hurrell JW (1995) Decadal trends in the North Atlantic Oscillation: regional temperatures and precipitation. *Science* 269:676–679
- IPCC (2014) *Climate change 2014. Impacts, adaptation and vulnerability. Part A: Global and sectoral aspects. Contribution of working group II to the fifth assessment report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, Cambridge
- Jennings E, Jones SE, Arvola L et al (2012) Effects of weather-related episodic events in lakes: an analysis based on high-frequency data. *Freshw Biol* 57:589–601
- Jones ID, Winfield IJ, Carse F (2008) Assessment of long-term changes in habitat availability for Arctic charr (*Salvelinus alpinus*) in a temperate lake using oxygen profiles and hydroacoustic surveys. *Freshw Biol* 53:393–402
- Kipling C (1972) The commercial fisheries of Windermere. *Transactions of the Cumberland and Westmorland Antiquarian and Archaeological Society* 72: 156–204

- Lack TJ, Lund JWG (1974) Observations and experiments on phytoplankton of Blelham Tarn, English Lake District- Experimental tubes. *Freshw Biol* 4:399–415
- Lampert W, McCauley E, Manly BFJ (2003) Trade-offs in the vertical distribution of zooplankton: ideal free distribution with costs? *Proc R Soc B Biol Sci* 270:765–773
- Le Cren D (2001) The Windermere perch and pike project: an historical review. *Freshw Forum* 15:3–34
- Liboriussen L, Landkildehus F, Meerhoff M et al (2005) Global warming: design of a flow-through shallow lake mesocosm climate experiment. *Limnol Oceanogr Methods* 3:1–9
- Lottig NR, Wagner T, Henry EN et al (2014) Long-term citizen-collected data reveal geographical patterns and temporal trends in lake water clarity. *PLoS One* 9:e95769
- Maberly SC (1996) Diel, episodic and seasonal changes in pH and concentrations of inorganic carbon in a productive lake. *Freshw Biol* 35:579–598
- Maberly SC, Elliott JA (2012) Insights from long-term studies in the Windermere catchment: external stressors, internal interactions and the structure and function of lake ecosystems. *Freshw Biol* 57:233–243
- Mackereth FJH, Heron J, Talling JF (1978) Water analysis: some revised methods for limnologists. *Freshwater Biological Association Scientific Publication*: 1–120
- Meis S, Thackeray SJ, Jones ID (2009) Effects of recent climate change on phytoplankton phenology in a temperate lake. *Freshw Biol* 54:1888–1898
- Mills JA, Teplitsky C, Arroyo B et al (2015) Archiving primary data: solutions for long-term studies. *Trends Ecol Evol* 30:581–589
- Mills JA, Teplitsky C, Arroyo B et al (2016) Solutions for archiving data in long-term studies: a reply to Whitlock et al. *Trends Ecol Evol* 31:85–87
- Moss B (2012) Cogs in the endless machine: lakes, climate change and nutrient cycles: a review. *Sci Total Environ* 434:130–142
- Nisbet E (2007) Earth monitoring: cinderella science. *Nature* 450:789–790
- Norton L, Elliott JA, Maberly SC et al (2012) Using models to bridge the gap between land use and algal blooms: an example from the Loweswater catchment, UK. *Environ Model Softw* 36:64–75
- Ohlberger J, Langangen O, Edeline E et al (2011) Stage-specific biomass overcompensation by juveniles in response to increased adult mortality in a wild fish population. *Ecology* 92:2175–2182
- Ohlberger J, Thackeray SJ, Winfield IJ et al (2014) When phenology matters: age-size truncation alters population response to trophic mismatch. *Proc R Soc B Biol Sci* 281:20140938. doi:10.1098/rspb.2014.0938
- Paxton CGM, Winfield IJ, Fletcher JM et al (2004) Biotic and abiotic influences on the recruitment of male perch in Windermere, UK. *J Fish Biol* 65:1622–1642
- Pearsall WH, Pennington P (1947) Ecological history of the English Lake District. *J Ecol* 34:137–148
- Pickering AD (2001) Windermere: restoring the health of England's largest lake. *Freshwater Biological Association, Kendal*
- Read JS, Hamilton DP, Jones ID et al (2011) Derivation of lake mixing and stratification indices from high-resolution lake buoy data. *Environ Model Softw* 26:1325–1336
- Reynolds CS (1989) Physical determinants of phytoplankton succession. In: Sommer U (ed) *Plankton ecology: succession in plankton communities*. Springer, New York, pp 9–56
- Reynolds CS (1990) Temporal scales of variability in pelagic environments and the response of phytoplankton. *Freshw Biol* 23:25–53
- Reynolds CS (1997) Successional development, energetics and diversity in planktonic communities. In: Abe T, Levin SR, Higashi M (eds) *Biodiversity: an ecological perspective*. Springer, New York, pp 167–202
- Reynolds CS, Thompson JM, Ferguson AJD et al (1982) Loss processes in the population dynamics of phytoplankton maintained in closed systems. *J Plankton Res* 4:561–600

- Reynolds CS, Wiseman SW, Godfrey BM et al (1983) Some effects of artificial mixing on the dynamics of phytoplankton populations in large limnetic enclosures. *J Plankton Res* 5:203–234
- Reynolds CS, Irish AE, Elliott JA (2001) The ecological basis for simulating phytoplankton responses to environmental change (PROTECH). *Ecol Model* 140:271–291
- Reynolds CS, Maberly SC, Parker JE et al (2012) Forty years of monitoring water quality in Grasmere (English Lake District): separating the effects of enrichment by treated sewage and hydraulic flushing on phytoplankton ecology. *Freshw Biol* 57:384–399
- Roy HE, Preston CD, Roy DB (2015) Fifty years of the Biological Records Centre. *Biol J Linn Soc* 115:469–474
- Sand-Jensen K (1997) The origin of the Freshwater Biological Laboratory. In: Sand-Jensen K, Pedersen O (eds) *Freshwater Biology Priorities and development in Danish Research*. G.E.C. Gad, Copenhagen, pp 9–15
- Schindler DW (1990) Experimental perturbations of whole lakes as test of hypotheses concerning ecosystem structure and function. *Oikos* 57:25–41
- Shepherd TD, Litvak MK (2004) Density-dependent habitat selection and the ideal free distribution in marine fish spatial dynamics: considerations and cautions. *Fish Fish* 5:141–152
- Solomon CT, Bruesewitz DA, Richardson DC et al (2013) Ecosystem respiration: drivers of daily variability and background respiration in lakes around the globe. *Limnol Oceanogr* 58:849–866
- Straile D (2000) Meteorological forcing of plankton dynamics in a large and deep continental European lake. *Oecologia* 122:44–50
- Strong C, Maberly SC (2011) The influence of atmospheric wave dynamics on interannual variation in the surface temperature of lakes in the English Lake District. *Glob Chang Biol* 17:2013–2022
- Talling JF (1993) Comparative seasonal changes, and interannual variability and stability, in a 26-year record of total phytoplankton biomass in 4 English lake basins. *Hydrobiologia* 268:65–98
- Talling JF (1999) *Some English lakes as diverse and active ecosystems: a factual summary and source book*. Freshwater Biological Association, Kendal
- Talling JF (2008) The developmental history of inland-water science. *Freshw Rev* 1:119–141
- ten Brink H, Mazumdar AKA, Huddart J et al (2015) Do intraspecific or interspecific interactions determine responses to predators feeding on a shared size-structured prey community? *J Anim Ecol* 84:414–426
- Thackeray SJ, Jones ID, Maberly SC (2008) Long-term change in the phenology of spring phytoplankton: species-specific responses to nutrient enrichment and climatic change. *J Ecol* 96:523–535
- Thackeray SJ, Sparks TH, Frederiksen M et al (2010) Trophic level asynchrony in rates of phenological change for marine, freshwater and terrestrial environments. *Glob Chang Biol* 16:3304–3313
- Thackeray SJ, Henrys PA, Jones ID et al (2012) Eight decades of phenological change for a freshwater cladoceran: what are the consequences of our definition of seasonal timing? *Freshw Biol* 57:345–359
- Thackeray SJ, Henrys PA, Feuchtmayr H et al (2013) Food web de-synchronization in England's largest lake: an assessment based on multiple phenological metrics. *Glob Chang Biol* 19:3568–3580
- Thomcroft CD, Hoskins BJ, McIntyre MF (1993) Two paradigms of baroclinic-wave life-cycle behaviour. *Q J R Meteorol Soc* 119:17–55
- Trolle D, Hamilton DP, Hipsey MR et al (2012) A community-based framework for aquatic ecosystem models. *Hydrobiologia* 683:25–34
- Whitlock MC, Bronstein JL, Bruna EM et al (2016) A balanced data archiving policy for long-term studies. *Trends Ecol Evol* 31:84–85
- Williamson CE, Dodds W, Kratz TK et al (2008) Lakes and streams as sentinels of environmental change in terrestrial and atmospheric processes. *Front Ecol Environ* 6:247–254

- Winfield IJ, Fletcher JM, James JB (2007) Seasonal variability in the abundance of Arctic charr (*Salvelinus alpinus* (L.)) recorded using hydroacoustics in Windermere, UK and its implications for survey design. *Ecol Freshw Fish* 16:64–69
- Winfield IJ, Fletcher JM, James JB (2008a) The Arctic charr (*Salvelinus alpinus*) populations of Windermere, UK: population trends associated with eutrophication, climate change and increased abundance of roach (*Rutilus rutilus*). *Environ Biol Fish* 83:25–35
- Winfield IJ, James JB, Fletcher JM (2008b) Northern pike (*Esox lucius*) in a warming lake: changes in population size and individual condition in relation to prey abundance. *Hydrobiologia* 601:29–40
- Winfield IJ, Fletcher JM, James JB (2012) Long-term changes in the diet of pike (*Esox lucius*), the top aquatic predator in a changing Windermere. *Freshwater Biol* 57:373–383
- Woolway RI, Jones ID, Hamilton DP et al (2015) Automated calculation of surface energy fluxes with high-frequency lake buoy data. *Environ Model Softw* 70:191–198
- Woolway RI, Jones ID, Maberly SC et al (2016) Diel surface temperature range scales with lake size. *PLoS One* 11:e0152466
- Worthington EB (1942) The Windermere perch fishery, and possibilities of its expansion. *Fishing Gazette*, 3 Jan 1942
- Worthington EB (1950) An experiment with populations of fish in Windermere, 1939–48. *Proc Zool Soc London* 120:113–149
- Wuest A, Piepke G, Van Senden DC (2000) Turbulent kinetic energy balance as a tool for estimating vertical diffusivity in wind-forced stratified waters. *Limnol Oceanogr* 45:1388–1400