Novel approaches for modelling changes in phytoplankton diversity and lake ecosystem function

A thesis

submitted in fulfilment

of the requirements for the degree

of

Doctor of Philosophy in Science

at

The University of Waikato

by

Kohji Muraoka

THE UNIVERSITY OF
WAIKATO
Te Whare Wānanga o Waikato

2019

Abstract

Ecosystem function represents the collective outcome of many different processes. Function may be interrupted by events that originate from outside a system, influencing biological diversity dynamics. Difficulties in expressing how a system is functioning originate firstly from being able to define a normative status for a dynamic system and secondly from the accuracy of common metrics of biodiversity changes. In this thesis, I used a numerical model and high-frequency ecological observations to express functioning of a system. Chapter 2 used biogeochemical parameter perturbations in a lake ecological model to identify seasonal parameter sensitivity variabilities. A set of internal process parameters of calibrated shallow eutrophic Lake Waahi DYRESM-CAEDYM ecological model was used to apply Monte-Carlo perturbation. Analysis was conducted by examining the collective results variability, a "spread" of the ensemble results from the iteration. The results showed that the spreads were small when lake inflows had high discharge, suggesting that lake internal dynamics had lesser effect on water quality and inflows dominated the system dynamics. Due to the simplicity of the methods, regular use of perturbation methods is suggested to assess model uncertainty and to better understand the model. Chapter 3 used interdisciplinary methods to identify changes in dissolved oxygen (DO) observations caused by biological processes. DO in lakes is a key indicator of ecosystem function. Methods used in this chapter included expert panel decision making, Symbolic Aggregate approXimation (SAX) analysis, and text classification. The use of an expert panel was motivated by the common practice of DO data visual assessment. Variability in experts' boundaries for data quality were observed by data survey, reinforcing the necessity of robust and reproducible methods for unbiased analysis. Surface DO sensor data from 18 global lakes were used to create day-long data segments. The modelling framework successfully simulated the expert panel decisions on these segments, automatically labelling data to indicate when the signal is likely dominated by biological activities. In Chapter 4, species-neutral biological assemblage metrics were developed to account for phytoplankton changes associated with changes in species abundance. Every species' population changes were converted into binary metrics (i.e., increases or decreases) to identify the "constituents" of species richness, to allow robust assessments of population dynamics. Four lakes (Lakes Annie, Feeagh,

Esthwaite and Mendota) from different regions were analysed. The results showed several previously undocumented features. Species recruitment was proportional to the number of species that were increasing. The number of species that were decreasing did not immediately increase the number of species that went extinct. The rate of increase was logarithmically distributed from the fastest to the slowest growing species, with the distribution shape being strongly influenced by number of species that were increasing. Such species-neutral community metrics, along with abundance distribution and diversity, are helpful to assess mechanistic community ecology models. This thesis provides toolsets useful for future studies to understand relationships between forcing and functioning of ecosystems and changes in biodiversity, by providing means to assess ecosystem function and demonstrating examples of species-neutral community structural changes.

Acknowledgements

I cannot thank more my supervisor, Professor David Hamilton, for his supervision through my thesis. This thesis became very challenging due to number of reasons, but his patience during its progress was always helpful and his knowledge of wide variety of subjects was necessary for this interdisciplinary study. The chief supervisor, Dr Adam Hartland, had extraordinary ability to interpolate my chaotic thoughts into process understanding. I appreciate his time and effort devoted to my thesis conclusion. I also thank my other supervisors, Professor Paul Hanson, Professor Bas Ibelings, and Dr Piet Verburg, for their patience and input into a variety of aspects of my study. I thank colleagues from the Global Lake Ecological Observatory Network (GLEON), especially the original The Theory Group (TTG) supporting theory development and phytoplankton data generation, exploration and manipulation. I cannot thank more the data providers for this study, who let me use their published and unpublished data. I thank all the co-authors for useful inputs and their expertise on various matters, but I would especially like to thank Professor Eibe Frank for patiently going through data mining theories with me.

I thank Associate Professor Chris Hendy for giving me inspiration and motivation for this study. Without his teaching of undergraduate geochemistry to me, I would not be pursuing aquatic sciences research. I also like to thank many staff of Lake Ecosystem New Zealand (LERNZ) for their support, especially Professor Brendan Hicks, Chris McBride, Dr Grant Tempero, Professor Kevin Collier, Dr Moritz Lehmann, Dr Mat Allan, Dr Susie Wood, and Professor Troy Baisden. The university's staff, not limited to Ai-Phing Wood, Anthea Kivell, Carol Robinson, Dudley Bell, Gloria Edwards, Lee Laboyrie, Tanya Mete, Vicki Smith and Warrick Powrie, helped me in great deal during the process. I thank many friends and colleagues who helped me immensely, but who are too numerous to mention. Without them, I would not have been able to get through difficult times; thank you, I mean it.

Last but not least, I send special thanks and love to my family members for being so patient and supportive.

Table of contents

List of figures

List of tables

Chapter one

Introduction

1.1 Ecosystem functioning and environmental fluctuations

Responses of ecological variables to external environmental drivers have been a central theme of ecology (Yachi and Loreau, 1999). The responses are related to ecological function, which has several definitions: (1) direct interactions between two elements (biological or non-biological; function as processes), (2) state or trajectory of a system that enables it to be sustained (functioning of a system), (3) relationships between parts of a system and the whole system (function as a role), and (4) how a system contributes to human wellbeing through provision of ecosystem services (Jax, 2005). Biological communities are structured by the balance between biological organization and environmental fluctuations (Reynolds, 1993). Environmental fluctuations act as an external driver to influence well-being of a community according to the strength of the environmental interactions.

Species diversity is commonly used to measure community health or well-being. Conventional studies of diversity have related it to ecosystem processes such as stability or disturbance (Tilman and Downing, 1987). Others consider diversity changes (Hillebrand et al., 2018) and interactions with ecosystem functioning rather than processes (Jaillard et al., 2018). Ecosystem function generally needs to be evaluated relative to some normative status of the system (defined here as the state a system attains at equilibrium) (Jax, 2005).

1.2 The dynamic nature of phytoplankton community composition in lakes

Lake phytoplankton are microscopic algae suspended in the water column that often make up the majority of the aquatic productivity of these systems. Lakes provide conditions in which phytoplankton can grow and reproduce, as well as being removed through a range of processes. Hutchinson (1957) reiterated Forbes' (1887) famous quote "lakes as a microcosm", stating that lakes are a "whole series of comparable yet different" systems. In an ecological context, it is the vast gradient of physical, morphological and geochemical characteristics of lakes that make them unique, despite the commonality of ecological processes (Lewis, 2011). Lake phytoplankton comprise a unique assemblage but exhibit common patterns across communities, described through concepts like functional groups of phytoplankton (e.g. Reynolds et al., 2006) and their conceptualised seasonal behaviours (PEG model; Sommer et al., 2012). Ecologists have long hoped to explain such patterns based on common drivers and processes, combined with elementary theory such as competition and niche differentiation.

Plants in terrestrial ecosystems generally exhibit well-ordered patterns of species changes (succession) and ultimately may reach a steady state (climax community; Clements, 1936). Arguably, lake phytoplankton assemblages have a similar steady state phase that may be affected by environmental factors such as turbulence, nutrient concentrations and light exposure, each of which is strongly influenced by seasonal stratification (Estrada and Berdalet, 1997). Seasonal shifts in stratification may be likened to catastrophic events in terrestrial ecosystems, evidenced by the total reorganisation in community composition (Sommer et al., 2012). A disturbance can be defined as a change of environment at a scale relevant to the community, which halts, shifts or reverses the successional processes (Reynolds et

al., 1993). By contrast, climax communities arise in largely undisturbed conditions, when the environment allows biological processes to dominate the system.

When attempting to make sense of, and predict the dynamics of phytoplankton, high intensity spatial and temporal observations are required. For example, sudden physical changes such as seasonal mixing may occur quickly, within timeframes of hours to days, and phytoplankton are well adapted to respond quickly to these cycles (Adrian et al., 2009) compared with primary producers in terrestrial systems (Figure 1.1). Therefore, in lakes it is possible to witness multiple succession sequences with a few months of near-continuous observation.

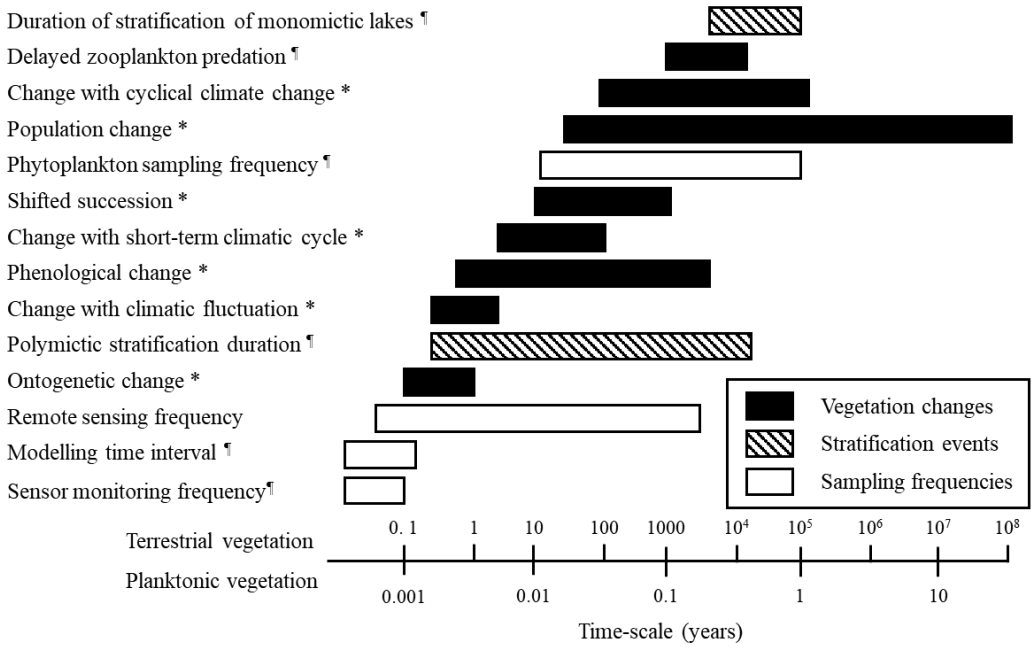


*Figure 1.1. Time-scales of change relevant to terrestrial and planktonic primary producers. The figure was modified from Reynolds (1995, original bars are indicated with *). The terrestrial time-scale was originally produced by Miles (1987). Lake stratification duration, lake water quality sampling and modelling time scales were added to compare with those of community change (lake-specific metrics are indicated by ¶).*

1.3 Conventional approaches to characterising phytoplankton community dynamics

Conventional lake monitoring often requires intense allocation of effort and resources. A typical sampling and analysis protocol may involve various locations and depths, acquiring sensor readings (e.g., as vertical profiles), analysing the composition of water quality samples (e.g., nutrients), manually identifying and counting phytoplankton and zooplankton species, and lab analysis of samples (Thornton et al., 1982). Rantajärvi et al. (1998) found that accurately predicting total phytoplankton biomass in the Baltic Sea would require weekly sampling. Successional shifts in phytoplankton assemblages take place much more quickly than total biomass changes (Figure 1.1), indicating that observations at higher frequency are likely to be required in order to accurately predict community dynamics. Furthermore, the phytoplankton community composition is much more difficult to observe than bulk estimates of phytoplankton biomass such as fluorescence and chlorophyll *a*. Recent developments of automated cell identification systems have the potential to revolutionize analysis of phytoplankton community composition by providing high frequency data (Thomas et al., 2018), however, deploying autonomous technology to identify and enumerate phytoplankton cells *in situ* has rarely been used and is not always applicable due to the high cost. Therefore, we are reliant upon use of (a) robust manual sampling schemes, (b) autonomous sensor readings when these are possible (mostly for physical variables) and (c) use of theories and models to compliment measurements (theory-guided data science; Karpatne et al., 2016).

1.4 Automated high resolution data acquisition

The Global Lake Ecosystem Observation Network (GLEON) is comprised of an *in situ* sensor network for lakes and has expanded rapidly since its formation in 2005 (Porter et al., 2009). Typical GLEON deployments consist of high resolution (HR; 1-30 minutes) meteorological observations (air temperature, wind speed/direction, shortwave radiation, humidity and surface photosynthetically active radiation), with water temperature profiles (various depths), and surface and bottom biogeochemical proxies (dissolved oxygen saturation and chlorophyll *a* fluorescence). Continuous HR observations from multiple lakes are a relatively recent advance in lake ecology research (Porter et al., 2005), and analytical tools have been actively developed to promote their use and to establish accepted methodologies to process the high volume of sensor data (e.g., physical stability - Read et al., 2011; environmental sensor observation quality assurance/quality control (QA/QC) - Horsburgh et al., 2015; energy flux - Woolway et al., 2015; Horsburgh et al., 2015; lake metabolism - Winslow et al., 2016). These directly- or indirectly-assessed metrics can help to understand biological processes by providing better understanding of physical conditions as well as continuous records of biogeochemical proxies.

1.5 Thermal stratification status

A key derivative of HR lake sensor observations is thermal stratification status. Many mathematical methods have been proposed to calculate density stratification from temperature profiles, such as Schmidt stability (Idso, 1973) or Wedderburn number (Thompson and Imberger, 1980). Read et al. (2011) laid the groundwork for expanding the capabilities of HR observations by standardizing and automating a set of equations for stability in a model, including a graphical user interface (GUI)

via a web application as well as a R language version (R package: Winslow et al., accessed in 2018). This tool has significantly reduced the time required to process lake thermal data and increased the ability of limnologists to comprehend and integrate lake physics indicators into ecological studies. The tool also has the potential to provide near real-time calculation of key lake stratification metrics.

The value of calculating lake thermal layer stability metrics can be reiterated in the context of climate change and increasing extreme weather events (Gallina et al., 2011; Brookes and Carey, 2011). Changes in stratification can radically alter the biogeochemistry and ecology of lakes (Sommer et al., 2012), influencing interactions amongst biota (e.g., grazing), nutrient levels and sediment diagenesis (Ostrovsky et al., 1996). Phytoplankton, which are short-lived compared to their terrestrial counterparts, are sensitive to subtle changes in physical conditions and environmental patterns, and respond quickly (Adrian et al., 2009). Therefore, analysis of physical conditions plays a critical role in understanding the dynamics of phytoplankton communities.

1.6 Free surface dissolved oxygen as a proxy of biological activities

The content of oxygen in lake water is influenced by whole ecosystem dynamics, including biological activities. In lakes, the surface oxygen concentration is regulated primarily by a dynamic balance between production (photosynthesis) and consumption (respiration) of the lake biota (Cole et al., 2000). Diurnal variations in oxygen concentrations can indicate the magnitude of the biological processes responsible for production and consumption of oxygen. By taking advantage of continuous HR observations of surface dissolved oxygen, a model has been developed to estimate ecosystem productivity (Hanson et al., 2003). Physical fluxes from air-water transfer of oxygen are included in the model, together with

biological processes related to oxygen production from photosynthesis and consumption from respiration. Data which cannot be explained from accounting for these processes in an ecosystem metabolism model are often manually or automatically removed (Staehr et al., 2010). This is achieved by investigating how the shape of the oxygen data compares with theoretical expectations. The facility to identify times when oxygen dynamics are primarily driven by biological activities allows 'clean', 'noise-free' data acquisition that better fits the ecosystem productivity model. Clean data are indicative of a less disturbed system, providing information to evaluate the behaviour of the community at times of low physical disturbance. Furthermore, the ecosystem metabolism model may be less well suited to the dynamic periods of changes in stratification that alter the phytoplankton community structure and dynamics.

1.7 Process-based modelling

Increased observation frequencies through autonomous sensor networks can enhance understanding of lake ecosystem dynamics (Porter et al., 2009), but additional processing may be required to support theories and provide deep understanding about the relevant processes. For this reason, numerical models have been designed to reflect real systems, through the application of a series of equations of dynamic processes occurring in lakes, and to study phenomena which cannot be assessed by conventional observations (Hamilton et al., 2015). Model uncertainty analysis, including parameter sensitivity analysis, is a methodology to understand uncertainty inherent in such numerical models (Sobol, 1993). Deterministic models can produce an array of results through different parameter combinations, and the application of a sufficiently large number of different parameter combinations allows an ensemble to be developed of model results

(Anderson and Anderson, 1999), so that the model output becomes quasi-probabilistic. The word quasi is used here to convey the meaning that model structure and parameter space is not necessarily fully explored, and there is only a limited space around the pre calibrated parameter values. These probabilistic results from ensemble model output can inform ecological research through understanding of model uncertainty, as opposed to a single deterministic model run. In addition, sensitivity of model output to manipulation of a collective of parameters that influence internal lake dynamics versus external forcing can provide information on critical gaps in measurement and understanding of the system.

1.8 Thesis outline

The objective of this thesis is to create new approaches that facilitate and enhance understanding of how physical events influence the dynamics of lake communities. Indices are developed and applied to understand how these events disrupt the normal community behaviour. The thesis has three research chapters which collectively test the hypothesis that the degree of disruption of lake functioning and biological communities can be explained retrospectively by key indices and metrics derived from HF data or model forcing variables. Susceptibility to disruption refers to an inverse transformation of resilience. A unique set of tools and methods is used to expand our understanding of the biophysical processes that dominate lake metabolism, phytoplankton succession and nutrient dynamics across a diverse set of lakes. The tools allow identification of when internal lake processes or biological equilibrium is overridden by external forcing factors. The chapters added valuable information on how the dominant lake processes can be identified in HF lake datasets and supporting models.

Chapter 2 uses a process-based model to identify the relative importance of external versus internal forces on lake ecosystem functioning. Using a calibrated shallow lake ecological model, 63 internal biogeochemical process parameters were perturbed from their calibrated values, while external forces were identical for all model iterations. Time series ensemble results were collated to assess the spread of results and identify the degree of parameter uncertainty. The extent of spread of daily ensemble results spreads was compared with external forcing causes of daily inflows to the lake. This chapter evaluates the value of parameter perturbation in deterministic numerical models and illustrates how this method can provide insights into lake dynamics that field observations or conventional deterministic model simulations cannot deliver.

Chapter 3 uses high-frequency dissolved oxygen observations from 18 lakes distributed across the globe, to assess the relative dominance of biological and non-biological processes. This assessment was made possible from the diurnal free surface dissolved oxygen concentration dynamics, based on the shape of changes in concentration within each day. The balance of community respiration and photosynthesis results in identifiable patterns of change in dissolved oxygen concentrations and a characteristic data 'shape' that is a function of time and light availability. The shape information is then used to identify days when lake metabolism is dominated by biological activities ("good data" for metabolism models) and allows information on the relative importance of biological versus non-biological factors affecting metabolism. A unique data mining method was used against an expert panel decision for the assessment of the value of automating the QA/QC process in dissolved oxygen data used in lake metabolism studies.

Chapter 4 uses conventional phytoplankton observations from four lakes to derive species diversity and population changes, and to understand community responses

to seasonal environmental perturbations. The analysis pairs consecutive samples in time and records species abundance changes in a binary form (i.e., increases or decreases) as well as for four constituents; increases, decreases, recruited or extinct. Relationships of each constituent, as well as temporal dynamics of constituent ratios, were assessed.

Each of the chapters in this thesis has been written in a manuscript format suitable for submission to a peer-reviewed scientific journal. Except when addressed or referenced, the analysis and results discussed in this thesis were produced from my own work under the supervision of Prof. David Hamilton (Griffith University, formerly the University of Waikato), with chief supervisor Dr Adam Hartland (the University of Waikato) and supervisory committee of Prof. Bas Ibelings (Université de Genève), Prof. Paul Hanson (University of Wisconsin-Madison), and Dr Piet Verburg (National Institute of Water and Atmospheric Research). Where data used in this research was provided externally, appropriate acknowledgement is made in each individual chapter. Where special assistance was received for individual chapters, acknowledgment sections in each chapter provide credit to co-authors as appropriate. At the time of submission of this thesis, Chapter 3 has been published in *Limnology & Oceanography: Methods*. Chapter 2 will be submitted to an appropriate journal. Chapter 4 is also formatted for submission to a journal but a decision has not been made on which journal at the time of thesis submittal.

1.9 References

Adrian, R., C. M. O. Reilly, H. Zagarese, and others. 2009. Lakes as sentinels of climate change. Limnol. Oceanogr. 54: 2283–2297. DOI: 10.4319/lo.2009.54.6_part_2.2283

Anderson, J. L., and S. L. Anderson. 1999. A Monte Carlo implementation of the nonlinear filtering problem to produce ensemble assimilations and forecasts. Mon. Weather Rev. 127: 2741–2758. doi:10.1175/1520-0493(1999)127<2741:AMCIOT>2.0.CO;2

Brookes, J. D., and C. C. Carey. 2011. Resilience to blooms. Science 334: 46–47. doi:10.1126/science.1207349

Cole, J. J., M. L. Pace, S. R. Carpenter, and J. F. Kitchell. 2000. Persistence of net heterotrophy in lakes during nutrient addition and food web manipulations. Limnol. Oceanogr. 45: 1718–1730. doi:10.4319/lo.2000.45.8.1718

Estrada, M., and E. Berdalet. 1997. Phytoplankton in a turbulent world. Sci. Mar. 61: 125–140

Forbes, S. A. 1887. The lake as a microcosm. Bull. Sci. Assoc. 77–87. doi:10.1353/rcr.2007.0015

Gallina, N., O. Anneville, and M. Beniston. 2011. Impacts of extreme air temperatures on cyanobacteria in five deep peri-alpine lakes. J. Limnol. 70: 186–196. doi:10.3274/JL11-70-2-04

Hamilton, D. P., C. C. Carey, L. Arvola, and others. 2015. A global lake ecological observatory network (GLEON) for synthesising high-frequency sensor data for validation of deterministic ecological models. Inland Waters 5: 49–56. doi:10.5268/IW-5.1.566

Hanson, P. C., D. L. Bade, S. R. Carpenter, and T. K. Kratz. 2003. Lake metabolism: Relationships with dissolved organic carbon and phosphorus. Limnol. Oceanogr. 48: 1112–1119. doi:10.4319/lo.2003.48.3.1112

Horsburgh, J. S., S. L. Reeder, A. S. Jones, and J. Meline. 2015. Open source software for visualization and quality control of continuous hydrologic and water quality sensor data. Environ. Model. Softw. 70: 32–44. doi:10.1016/j.envsoft.2015.04.002

Hutchinson, G. E. 1957. A Treatise On Limnology: Vol. 1: Geography, Physics And Chemistry. Wiley, New York

Idso, S. B. 1973. On the concept of lake stability. Limnol. Oceanogr. 18: 681–683. doi:10.4319/lo.1973.18.4.0681

Jax, K. 2005. Function and "functioning" in ecology: what does it mean? Oikos 111: 641–648. doi:10.1111/j.1600-0706.2005.13851.x

Jaillard, B., C. Richon, P. Deleporte, M. Loreau, and C. Violle. 2018. An a posteriori species clustering for quantifying the effects of species interactions on ecosystem functioning R. Chisholm [ed.]. Methods Ecol. Evol. 9: 704–715. doi:10.1111/2041-210X.12920

Karpatne, A., G. Atluri, J. H. Faghmous, et al. 2017. Theory-guided data science: a new paradigm for scientific discovery from data. IEEE Trans. Knowl. Data Eng. 29: 2318–2331. doi:10.1109/TKDE.2017.2720168

Lewis Jr, W. 2011. Global primary production of lakes: Inland Waters 1: 1–28. doi:10.5268/IW-1.1.384

Ostrovsky, I., Y. Z. Yacobi, P. Walline, and I. Kalikhman. 1996. Seiche-induced mixing: Its impact on lake productivity. Limnol. Oceanogr. 41: 323–332. doi:10.4319/lo.1996.41.2.0323

Porter, J. H., E. Nagy, T. K. Kratz, P. Hanson, S. L. Collins, and P. Arzberger. 2009. New eyes on the world: advanced sensors for ecology. Bioscience 59: 385–397. doi:10.1525/bio.2009.59.5.6

Porter, J., P. Arzberger, H.-W. Braun, et al. 2005. wireless sensor networks for ecology. Bioscience. doi:10.1641/0006-3568(2005)055[0561:WSNFE]2.0.CO;2

Rantajärvi, E., R. Olsonen, S. Hällfors, J. M. Leppänen, and M. Raateoja. 1998. Effect of sampling frequency on detection of natural variability in phytoplankton: Unattended high-frequency measurements on board ferries in the Baltic Sea. ICES J. Mar. Sci. 55: 697–704. doi:10.1006/jmsc.1998.0384

Read, J. S., D. P. Hamilton, I. D. Jones, K. Muraoka, L. A. Winslow, R. Kroiss, C. H. Wu, and E. Gaiser. 2011. Derivation of lake mixing and stratification indices from high-resolution lake buoy data. Environ. Model. Softw. 26: 1325–1336. doi:10.1016/j.envsoft.2011.05.006

Reynolds, C. S. 1995. The intermediate disturbance hypothesis and its applicability to planktonic communities: Comments on the view of Padisak and Wilson. N. Z. J. Ecol. 19: 219–225.

Reynolds, C. S. 2006. The Ecology of Phytoplankton, Cambridge University Press, Cambridge

Reynolds, C. S. 1993. Scales of disturbance and their role in plankton ecology. Hydrobiologia 249: 157–171. doi:10.1007/BF00008851

Sobol, I. M. 1993. Sensitivity estimates for nonlinear mathematical models. Math. Model. Comput. Exp. 1: 407–414

Sommer, U., R. Adrian, L. De Senerpont Domis, et al. 2012. Beyond the Plankton Ecology Group (PEG) Model: mechanisms driving plankton succession. Annu. Rev. Ecol. Evol. Syst. 43: 429–448. doi:10.1146/annurev-ecolsys-110411-160251

Spigel, R. H., and J. Imberger. 1987. Mixing processes relevant to phytoplankton dynamics in lakes. New Zeal. J. Mar. Freshw. Res. 21: 361–377. doi:10.1080/00288330.1987.9516233

Staehr, P. a, D. Bade, M. C. Van de Bogert, G. R. Koch, C. Williamson, P. Hanson, J. J. Cole, and T. Kratz. 2010. Lake metabolism and the diel oxygen technique: State of the science. Limnol. Oceanogr. Methods 8: 628–644. doi:10.4319/lom.2010.8.628

Thornton, K. W., R. H. Kennedy, A. D. Magoun, and G. E. Saul. 1982. Reservoir water quality sampling design. J. Am. Water Resour. Assoc. 18: 471–480. doi:10.1111/j.1752-1688.1982.tb00014.x

Tilman, D., and J. A. Downing. 1994. Biodiversity and stability in grasslands. Nature 367: 363–365. doi:10.1038/367363a0

Winslow, L. A., J. A. Zwart, R. D. Batt, H. A. Duggan, R. I. Woolway, J. R. Corman, P. C. Hanson, and J. S. Read. 2016. LakeMetabolizer: an R package for estimating lake metabolism from free-water oxygen using diverse statistical models. Inland Waters, 6(4), 622-636. doi:10.1080/IW-6.4.883

Woolway, R. I., I. D. Jones, D. P. Hamilton, S. C. Maberly, K. Muraoka, J. S. Read, R. L. Smyth, and L. A. Winslow. 2015. Automated calculation of surface

energy fluxes with high-frequency lake buoy data. Environ. Model. Softw. 70: 191–198. doi:10.1016/j.envsoft.2015.04.013

Yachi, S., and M. Loreau. 1999. Biodiversity and ecosystem productivity in a fluctuating environment: The insurance hypothesis. Proc. Natl. Acad. Sci. 96: 1463–1468. doi:10.1073/pnas.96.4.1463

Chapter two

Uncertainty assessment of a deterministic lake ecological model using parameter perturbations

2.1 Abstract

It is important to communicate assumptions and uncertainties in numerical lake ecological models, particularly to improve the quality of decision making. The objective of this study was to examine uncertainty in parameter values used in a deterministic lake ecological model and provide an ensemble of simulation outputs that reflect this uncertainty. I applied a Monte-Carlo parameter perturbation to a widely used lake ecological model (DYRESM-CAEDYM). The model parameters were calibrated for Lake Waahi, a eutrophic shallow lake in North Island, New Zealand. The distribution of most of the model outputs had stabilized in approximately 2,000 model runs, with additional model simulations having little effect on ensemble distribution statistics. Minor perturbations in parameter ranges (±5 and 10%) resulted in reasonably normal distributions around the base calibrated model, while greater perturbation ranges (±25 and 50%, which may include unrealistic parameter values) resulted in very wide output distributions. The spread of time series output varied seasonally, indicating shifts of the dominant processes between external forcing (e.g., flushing) and internal forcing (e.g., sediment nutrient release). Because parameter perturbation is commonly used to generate an of ensemble model results, an analysis of parameter sensitivity was made. The

results reiterated the importance of nutrient sediment diagenesis processes in key output variables including chlorophyll *a*. My method suggests that the conventional output from a deterministic model of a single time series based on a calibrated set of parameters could be replaced by a time series of probability densities that better capture observed data and convey parameter uncertainty effects. Current computing hardware and software make it possible to perform many hundreds of model simulations to quantify uncertainty and sensitivity of model output in this way. My approach could be used as standard practice for biogeochemical modelling.

2.2 Introduction

Lakes have great aesthetic, cultural and environmental value. Numerical models of lake water quality are especially useful where managers are required to evaluate past degradation as well as future scenarios arising from both planned changes (e.g., catchment management; Johnes et al., 1997; hydrological modifications; Liu et al., 2014) and expected environmental trajectories (e.g., climate change; Whitehead et al., 2009; Trolle et al., 2011). Creating realistic scenarios for unique and complex lake systems, in particular for water quality, is generally problematic because drivers of change are usually modelled or extrapolated, with a tendency for error to compound through the process (Knutti et al., 2012). Indeed, inadequate knowledge and assumptions about the processes can further compound such errors.

Hellweger (2017) listed six criticisms of numerical models: (1) a large number of assumptions, (2) risk of overfitting, (3) high levels of uncertainty, (4) inadequate knowledge, (5) insufficient observations for satisfactory calibration and validation, and (6) difficulties in development, operation, analysis and communication of output. Some of these issues stem from the deterministic nature of most process-based numerical models that generate a single output result with fixed parameter

values. These issues are why Oreskes et al. (1994) state that "verification and validation of numerical models of natural systems is impossible". In this context, numerical models that generate probabilistic results to express uncertainty can be of great value in communicating limitations of the model simulations to stakeholders and in conveying the accuracy of the simulations.

Deterministic models of lakes generally have fixed values of parameters based on single model realizations, and therefore do not provide explicit output of the associated uncertainties. They thus neglect the stochastic nature of natural ecosystems and uncertainty and the underlying errors in calibration, forcing variables (e.g., weather, nutrient loads; Refsgaard et al., 2006) and observations. Uncertainty analysis considers inputs (parameters and forcing variables) to the model for the potential errors that may be propagated through the simulation (USEPA, 2009). Sensitivity analysis is used to examine how the uncertainty in the model output may be attributed to changes made to a given input factor (i.e. parameters and forcing variables) (Saltelli et al., 2007; Pianosi et al., 2016). Uncertainty analysis and sensitivity analysis are two terms often used interchangeably, as many processes can be common between the two. Although sensitivity analysis is optional in modelling practice, it has been extensively applied in lake ecological modelling to validate models or test hypotheses (Makler-Pick et al., 2011), as well as informing the focus of future studies.

One way to represent model sensitivity is to rank parameters or forcing variables by their relative influence on model outputs (Pianosi et al., 2016), but the preferred method is a global sensitivity analysis, which involves testing the individual and combined effects of all factors over all of their possible ranges (Saltelli et al., 2007). For lake modelling, the factors may include forcing variables, parameters and initial conditions. In a local sensitivity analysis the analysis is limited to a specific range

or a subset of factors. If the aim of a study is to identify true sensitivity of a parameter, range selection must be made carefully. Two common sensitivity analysis methods are One At a Time (OAT) and All At a Time (AAT). An OAT analysis changes a single factor (e.g., one parameter) at a time. OAT is commonly practiced due to its simplistic nature, with a given change in a parameter being expressed by the relative change in model output. OAT does not provide information on the interactive influence with other parameters, which may be significant in a complex numerical model. The AAT method modifies multiple parameters simultaneously. A Monte-Carlo parameter selection is often used for AAT, and a scatter diagram that relates model output to factor (parameter) value changes is commonly used to assess the influence of the factor in the output, and whether the relationship is linear or nonlinear. Variance based analysis is a flexible sensitivity analysis method because no prior knowledge of linearity between the two variables and the output is required. For non-additive model equations, higher order parameter sensitivity analysis such as total effect indices (Sobol, 1993) may be necessary.

Addressing all factorial uncertainties may be challenging, but progress has been made using Markov Chain Monte Carlo based Bayesian techniques (e.g. Dietzel and Reichert, 2014; Couture et al., 2018) to capture some of the variability and stochasticity of parameters and models. Even simple Monte-Carlo based approaches can be useful, however, and these have been extensively used outside lake ecosystem models (e.g. Kroese et al., 2014). For example, daily weather forecasts use Monte Carlo parameter selection alongside multiple initial condition selections and model system setups to produce an ensemble of simulation outputs (Anderson and Anderson, 1999; Ollinaho et al., 2017). In lake modelling, multi-parameter selection using Monte Carlo techniques has been used to calibrate models

(Luo et al., 2018) or to test the sensitivity of parameters (Makler-Pick et al., 2011). Schlabing et al. (2014) produced model results using 500 different synthetic weather conditions, and presented the results as a time series in one figure to demonstrate the variability of model outputs caused by forcing factors. Ensemble results from a sensitivity analysis can be expressed as a probability density function (pdf).

The aim of this study was to incorporate parameter uncertainties into a deterministic model in order and examine ways to express multiple model outputs (model ensemble results). To achieve this, a Monte-Carlo parameter perturbation was carried out using a calibrated lake model. Calibrated parameter values were treated as a close reflection of those of the real system. An ensemble pdf of the model simulation output was used to express the variability of the output and examine the degree of alignment with observed values. The sensitivity of model outputs to each parameter was determined, including variations in sensitivity with season and ranking of parameters to explore processes most influential on the model output.

2.3 Methods

2.3.1 Study site and model setup

Lake Waahi (37°34'S, 175°91'E) is a eutrophic, medium sized (5.22 km^2), shallow (mean depth c. 2 m), turbid (Secchi depth < 0.5 m) lake, originally part of a floodplain connected to the Waikato River, in the central North Island of New Zealand. The lake has received diffuse and direct discharge from coal mining, contributing large quantities of sediment which has increased the turbidity of the lake. The catchment also has large areas of land used for pastoral grazing by dairy cows and drystock. Submerged macrophyte beds in the lake collapsed in the 1970s (Kingett, 1984), and their disappearance was followed by high rates of wind-induced sediment resuspension which has contributed high levels of suspended sediments. In addition, a minor decrease in the regulated minimum water level since 2013 has been associated with further increases in suspended sediments and declining water quality (Lehmann et al., 2017). The lake phytoplankton community is dominated by cyanophytes throughout the year.

In this study, a DYRESM-CAEDYM model (Hamilton and Schladow, 1997; Schladow and Hamilton, 1997; ver. 3.1.0-03) was applied to the lake. The model was developed by the Centre for Water Research at The University of Western Australia, and applied in many management and scientific research projects (e.g. Trolle et al., 2011). The model consists of one dimensional (1-D) hydrodynamic component (DYRESM) that calculates water density stratification and ecological component (CAEDYM) which simulates key water quality related biogeochemical variables, including TP, TN and phytoplankton biomass. The model was applied to Lake Waahi to examine different external and internal nutrient load scenarios that could be used to improve lake water quality (Lehmann et al., 2017). The model was set up for a five-year period from 2010 to 2015, and forced with meteorological

data from Ruakura meteorological station (approximately 20 km south of the lake). These data included air temperature (^{o}C), dew point temperature (^{o}C), relative humidity (%), precipitation (mm), wind speed (m s^{-1}), and shortwave radiation (W m^{-2}). Air temperature observations at Ruakura Station were recalibrated to available *in situ* high-frequency buoy observations (Eq. 1):

$$y = 1.267x - 1.097, \quad r^2 = 0.916, \quad p < 0.05, \quad \text{(Jan 2014 – Jun 2017)} \quad (1)$$

where y is the calibrated air temperature at the lake and x is the air temperature observed at Ruakura Station. Cloud cover, a proxy for long wave radiation input to the model, was calculated by assessing the fraction of observed shortwave radiation within the clear sky (=0) and cloudy (=1) shortwave radiation envelope. Total discharge to the lake, including seepage and surface tributaries, was estimated using a catchment water budget (CLUES, Elliot et al., 2016) and hydrological model (TopNET, Booker and Woods, 2014) operating at daily frequency. Monthly water quality observations for one (Awaroa Stream) of the two major surface inflows were applied to represent the composition of both inflows as the second inflow is not monitored but arises from a subcatchment of similar land use to Awaroa Stream.

Our main interest was on nutrient and phytoplankton dynamics, so emphasis was placed on bottom-up processes (nutrients – phytoplankton) in calibrating the model with observations of relevant physical, chemical and biological variables (such as dissolved oxygen and chlorophyll *a* concentrations). Sediment resuspension, thermal stratification and bottom-water anoxia strongly influence the nutrient dynamics in Lake Waahi (Lehmann et al., 2017). The process of validating input data and calibrating the model parameters involved first achieving a satisfactory water balance, followed by temperature and oxygen calibration, sediment and nutrients, and finally, chlorophyll *a*. A set of default parameters was available for Lake Ellesmere, Canterbury, New Zealand (Trolle et al., 2011) which is a similarly

shallow, eutrophic lake devoid of submerged macrophytes and with high rates of sediment resuspension. Parameters were adjusted in step-wise fashion, progressively reducing the root mean square error (RMSE) and improving the coefficient of determination in comparisons with observed variables for Lake Waahi.

Surface water concentrations of total phosphorus (TP), nitrogen (TN), chlorophyll *a* (TCHLA) and total suspended solids (TSS) were chosen as key output variables for further analysis. Median values of each of these output variables were calculated for the entire modelling period (~5 years). The time period is consistent with the New Zealand National Objective Framework (NOF) of the National Policy Statement for Freshwater Management (2014) that water quality is assessed as 3-5year period median values of surface TP, TN and TCHLA.

2.3.2 Monte Carlo parameter perturbation and All-At-a-Time (AAT) sensitivity analysis

A set of 63 parameters was perturbed simultaneously within the ranges of ±5, 10, 25 and 50% from the calibrated values. The perturbation used randomly distributed variations within the selected parameter perturbation ranges. Over 1000 perturbed model runs were produced and both time series and median values of surface TP, TN, TCHLA and TSS were recorded for each run. Any model simulation runs which did not fully complete (i.e., because extreme parameter ranges produced a 'crash' of the model) were disregarded from the analysis. For each variable, Kernel probability densities were calculated to describe the ensemble of results on each day of the model simulation. These results contain information of model iteration densities and spreads (ensemble spreads). The relative importance of the internal processes influenced by the perturbed parameters versus the external forcing

influenced by inflow volume was tested using standard deviations of the ensemble results, which were plotted against daily inflow volume.

The minimum number of model iterations necessary to provide adequate statistical representation of the variability was determined for each of the state variable outputs (5-year median TP, TN, TCHLA and TSS). The variance ratio (F) of a given number of iterations (j) was calculated as:

$$F_j = \frac{V(Y_1,Y_2,\ldots Y_j)}{V(Y_1,Y_2,\ldots Y_N)} \quad (2)$$

where Y_j is the j-th iteration model result (5-year median TP, TN, TCHLA and TSS; $j \leq N$), $V(\cdot)$ is the variance of the arguments $(\cdot)$, and N is the total number of iterations.

For each parameter, values of the four model output variables (Y axis) were plotted against the parameter value (x axis). Two methods were adopted to examine relationships between parameter values and model output: linear regression and variance-based analysis. For each model output (Y), the Pearson linear coefficient of determination was calculated as:

$$S_i = corr(x_i, Y)^2 \quad (3)$$

Where corr is the Pearson correlation coefficient function and x_i is the prescribed value of i-th parameter. A conditional variance-based analysis for Monte Carlo parameter distributions is described in Saltelli et al. (2007). This method slices the above-mentioned scatter diagrams vertically, with each slice containing an equal number of model results. For this study, ten slices were used. Mean Y values in each slice were then calculated, and variance across this mean Y value was calculated. The process can be expressed as:

$$S_i = \frac{V(E(Y|X_i))}{V(Y)} \quad (4)$$

where $E(\cdot)$ is the arithmetic mean of the argument $(\cdot)$, $V(\cdot)$ is the variance of $(\cdot)$, $Y|X$ describes multiple model outputs Y from a parameter set x, and x_i denotes parameter sets where the i-th parameter (x_i) is fixed while all other parameters are perturbed. In the conditional variance-based approach, the "conditional" term describes that x_i has a range within the limits of the slice to discretise the calculation process, instead of having an exact value.

2.3.3 One At a Time (OAT) parameter sensitivity analysis

OAT sensitivity analysis involved modifying each parameter value by ±5, 10, 25 and 50% of its calibrated value, while other parameters were kept identical to the base model. For all four variables, sensitivities (Si) of each parameter were calculated as:

$$S_i = \frac{abs(Y|x_{\sim i} - Y|x)}{Y|x} \quad (5)$$

where $x_{\sim i}$ is a base model parameter value set x, but only the i-th parameter is modified.

2.4 Results

The RMSE and coefficient of variation of the RMSE (CV(RMSE)) of the base model are shown in Table 2.1. Calibrated parameter values are listed in the Supplementary table.

Table 2.1: Model performance for Lake Waahi using root mean square error (RMSE) and coefficient of variation of root mean square deviation CV(RMSE) for high frequency buoy observations (n > 360; surface temperature, bottom temperature, surface dissolved oxygen, bottom dissolved oxygen and manual sampling results (n = 30; surface total nitrogen: TN, surface total phosphorus: TP, surface total suspended solids: TSS, surface total chlorophyll a: TCHLA).

Variable	RMSE	CV(RMSE)
Surface temperature (^{o}C)	0.93	0.05
Bottom temperature (^{o}C)	1.6	0.07
Surface dissolved oxygen (mg l^{-1})	1.02	0.11
Bottom dissolved oxygen (mg l^{-1})	1.11	0.13
TN (g m^{-3})	0.65	0.49
TP (g m^{-3})	0.03	0.39
TSS (g m^{-3})	46.26	0.51
TCHLA (mg m^{-3})	25.55	0.84

Stability in F was achieved for all four model output variables within 1000 iterations for the ±5%, 10% and 25% parameter perturbations. However, the ±50% parameter perturbations produced some instability, necessitating removal of extremely high values to achieve stability within a reasonable number of iterations (1000 – 2000). For this case, model output with simulated values exceeding both 90th and 80th percentiles was removed to test the results without extreme values.

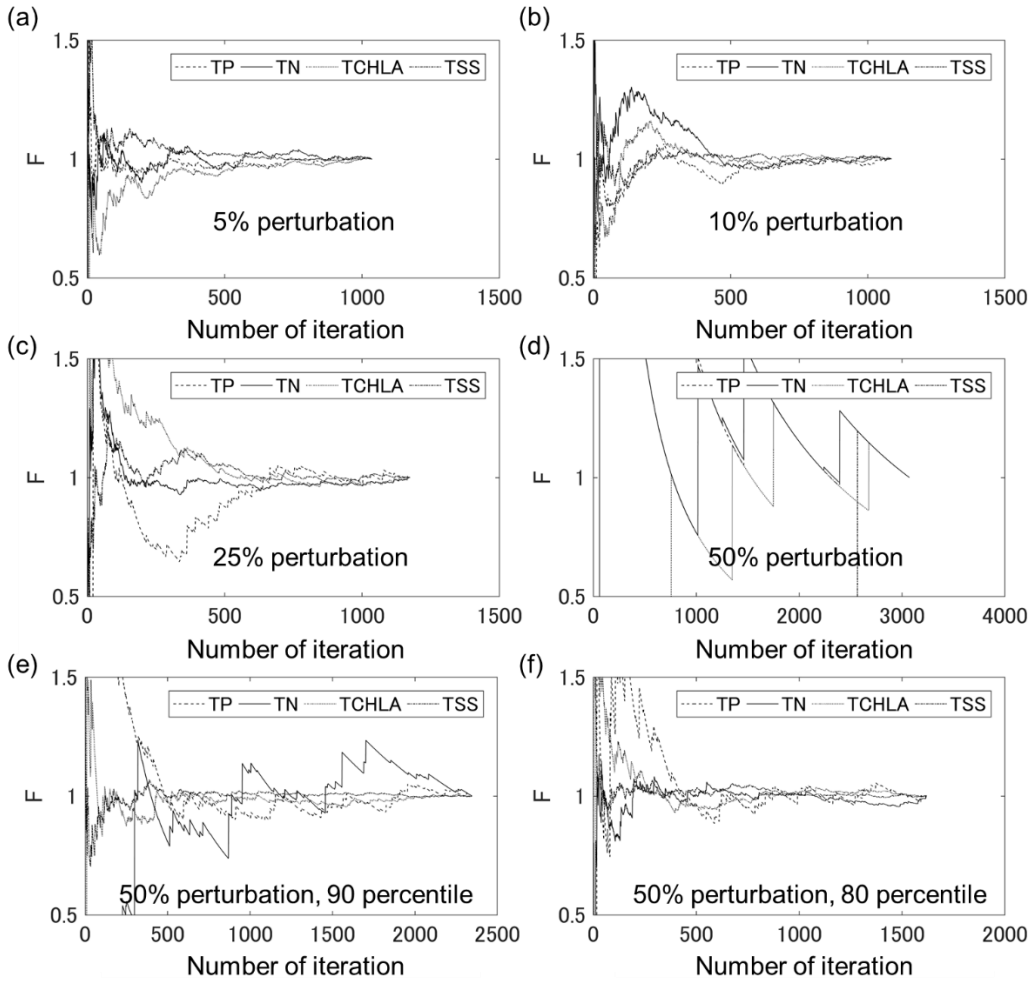


Figure 2.1: Variance ratio (F) and number of model iterations for four model output variables (5-year median of TP, TN, TCHLA and TSS) with ±5% (a), 10% (b), 25% (c) and 50% (d-f) parameter perturbations. (a) to (d) used all the model iterations. For (e) and (f), values were removed that respectively exceeded the 90th and 80th percentiles of the range.

The distributions of the four key model variables (TP, TN, TCHLA and TSS) based on five-year median values are illustrated in Figure 2.2. Model output for all four variables showed a normal distribution around the base model for both ±5% and ±10% parameter perturbations. The model output for the ±10% range was about twice as wide as that of ±5% range. For the ±25% and ±50% ranges, results were skewed towards the high end for TP, TCHLA and TSS and produced some extreme values, even though values exceeding the 80th percentile ranges had been removed. The TN distribution in these high ranges showed a bi-normal-like distribution

towards greater probability of low and high values, while TCHLA in many model simulations had collapsed, with values within the minimum assigned histogram bin.

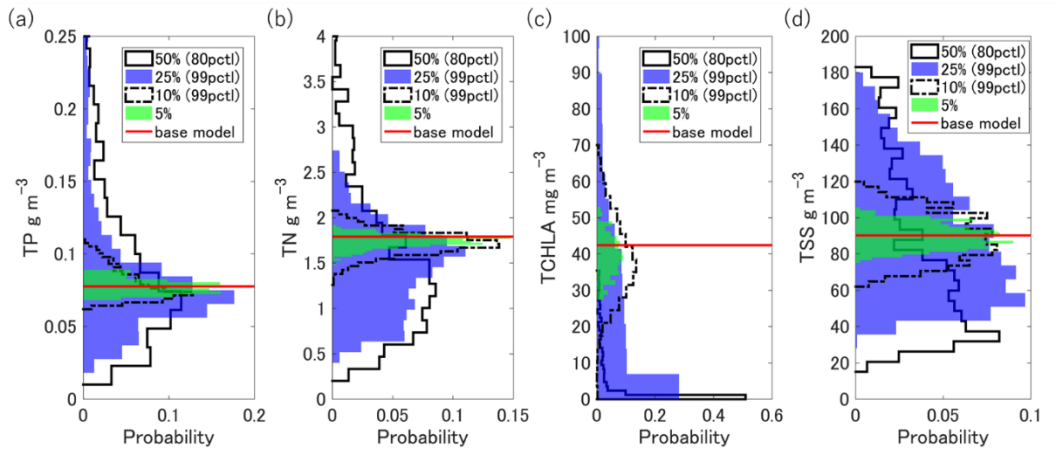


Figure 2.2: Histogram of parameter perturbed model results showing five-year median values of (a) TP, (b) TN, (c) TCHLA and (d) TSS. Probabilities were calculated as: (number of observations in the bin) / (total number of observations) where the bins were set to selected to provide adequate resolution of output data.

Figure 2.3 shows time series of probability density functions (pdfs) for TP (a, e, i, m), TN (b, f, j, n), TCHLA (c, g, k, o) and TSS (d, h, l, p) model output for the different parameter perturbation ranges. The base model outputs for which the model was calibrated, and the observed values, are also shown in Figure 2.3. Coloured envelopes (dynamic spreads) in Figure 2.3 show the limits of the ensemble model results, while the intensity of shading provides an indication of the frequency of occurrence. For the ±5% parameter perturbation (Figure 2.3a-d), pdf envelopes were small for all four variables. Throughout the five years of the model run, the distributions of ensemble results of output variables were reasonably normally distributed, with the base model being in the centre of the spreads of outputs. Envelope width thickening was observed during the summer periods for all four variables. Envelope sizes of TP and TN became markedly wider in 2011,

2013 and 2015 (Figure 2.3a-b). Some observations were not captured within the envelope, especially for TCHLA in 2013 (Figure 2.3c) and for TSS at various time points (Figure 2.3d). The ±10% parameter perturbation simulations (Figure 2.3e-h) gave similar but much wider envelope distributions than the ±5% perturbations ((Figure 2.3a-d). The results were still normally distributed throughout the model run, with the base model being in the centre of the ensemble result distributions. The TP envelope width increased in summer, and as a result, field observations were better captured than in the ±5% simulation envelopes. Envelope sizes expanded considerably for TP and TCHLA in the ±25% parameter range perturbation (Figure 2.3e, g). The TP (Figure 2.3i) and TSS (Figure 2.3l) results for this range showed high ensemble model results density around the base model results. TN and TCHLA output (Figure 2.3j, m) showed two regions of high density, with both pdf peaks occurring at lower values than the base calibration results. With ±50% parameter perturbation (Figure 2.3m-p), envelope thickness was significantly greater for TP, TN and TCHLA, despite the removal of state variables that exceeded the 80^{th} percentile values over the 5-year simulation period. While TP and TN densities were still centred around the base model (Figure 2.3m-n), TCHLA had high-density regions consistently near zero (Figure 2.3o), as reflected in the 5-year median values presented in Figure 2.2. In addition, high-density regions for TSS plots were also generally observed to be slightly below the base model (Figure 2.3p).

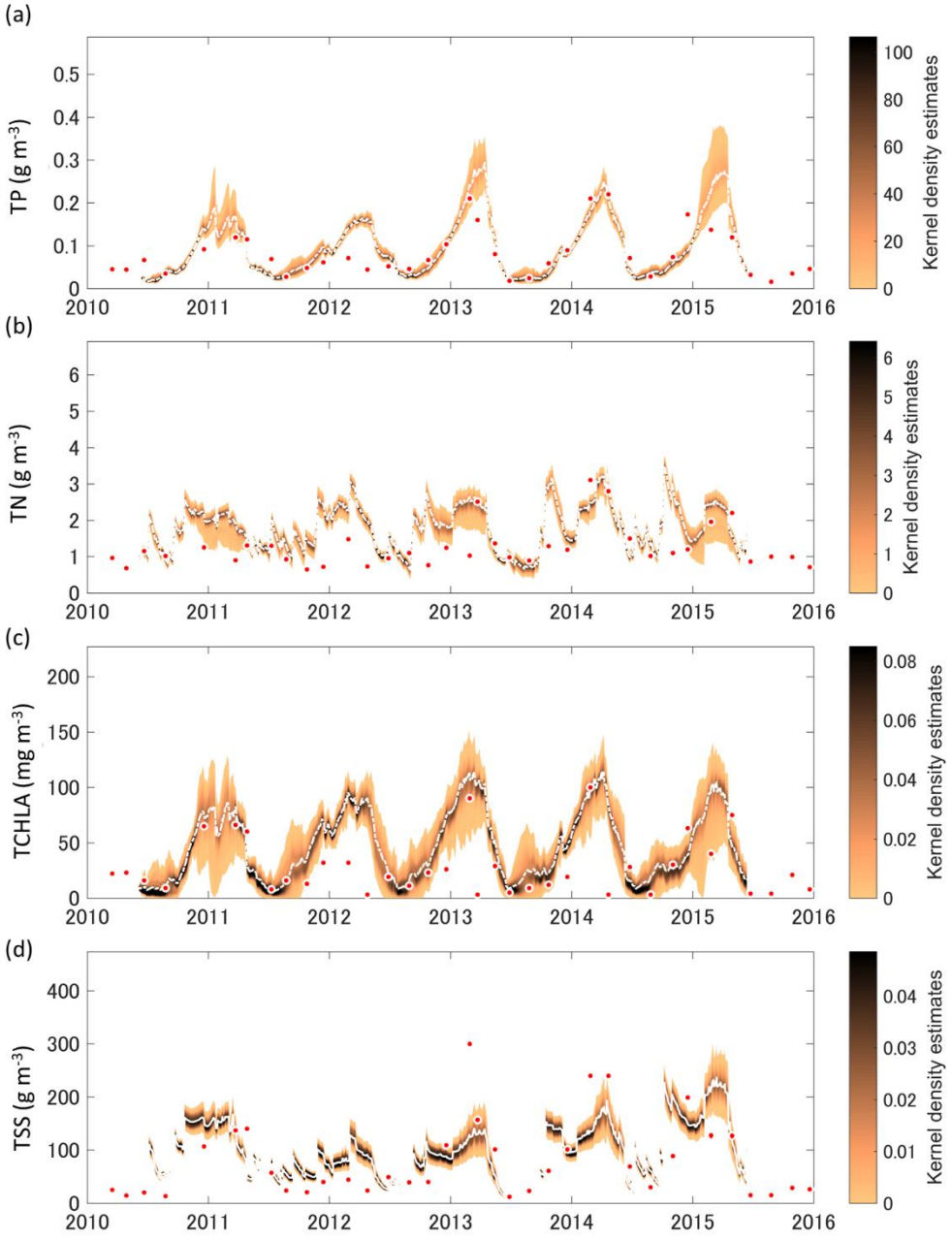


Figure 2.3. Time series model results with spreads from 63 perturbed parameters. a-d are from ±5% parameter perturbation (N = 1037), e-h are from ±10% perturbation (N = 1050), i-l are from ±25% perturbation (N = 1140), and m-p are from ±50% perturbation (N = 1620 with values exceeding 80 percentile of range removed). The results include surface total phosphorus (TP; a, e, i, m), total nitrogen (TN; b, f, j, n), total chlorophyll a *(TCHLA; c, g, k, o) and total suspended solids (TSS; d, h, l, p). Filled contours show Kernel probability density functions where occurrence increases from light to dark. The red dots indicate the observed values and white lines are the base model output.*

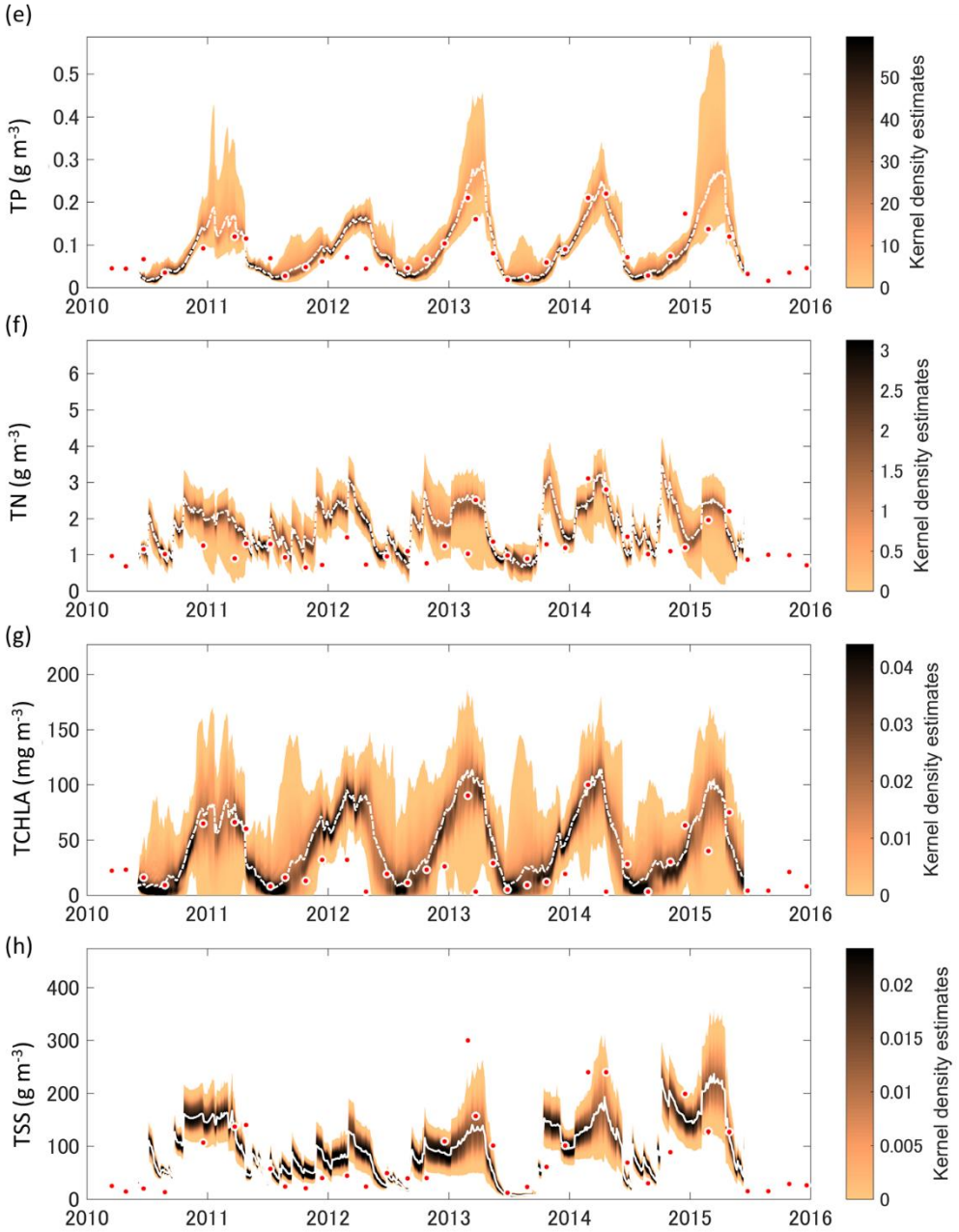


Figure 2-3. continued

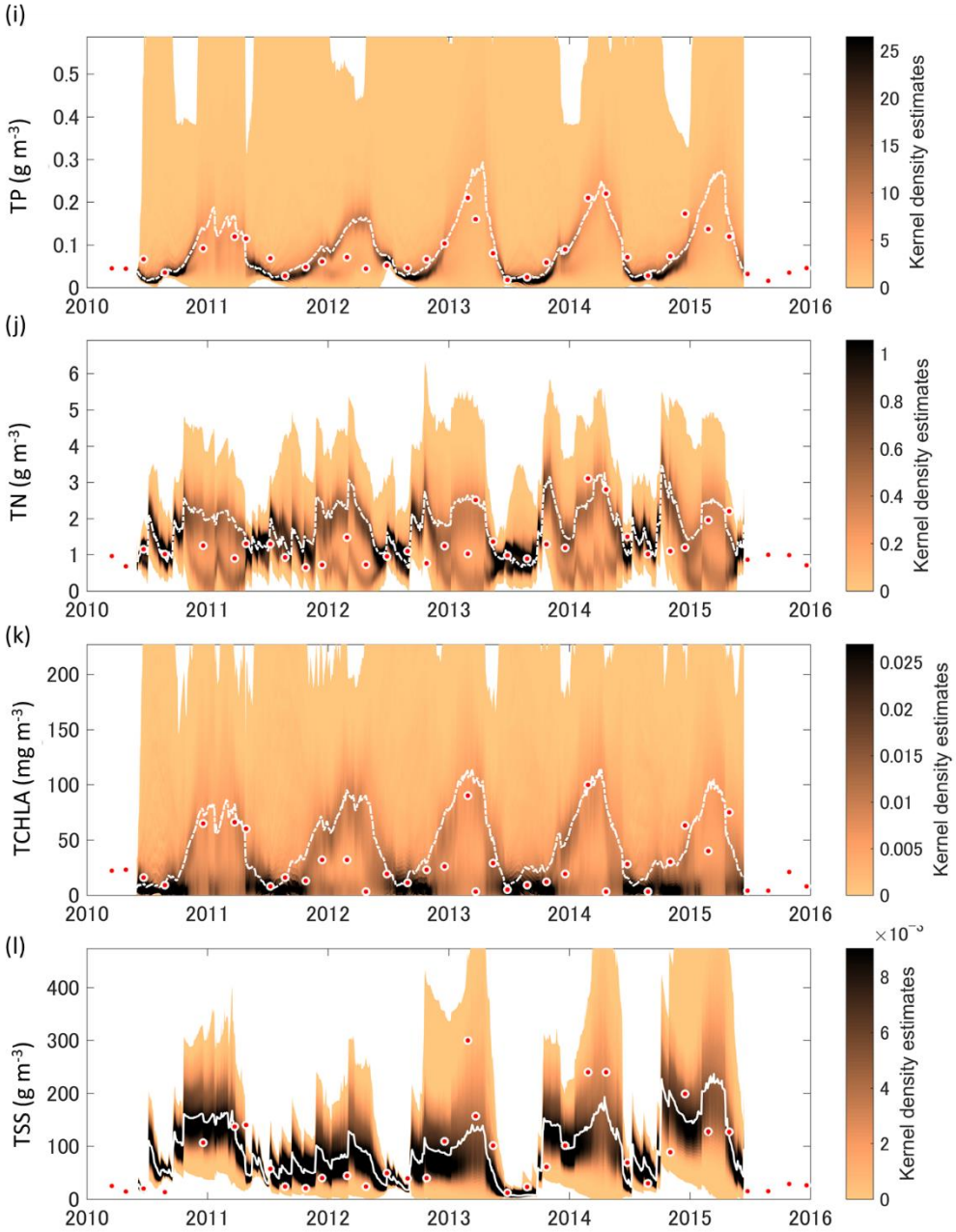


Figure 2-3. continued

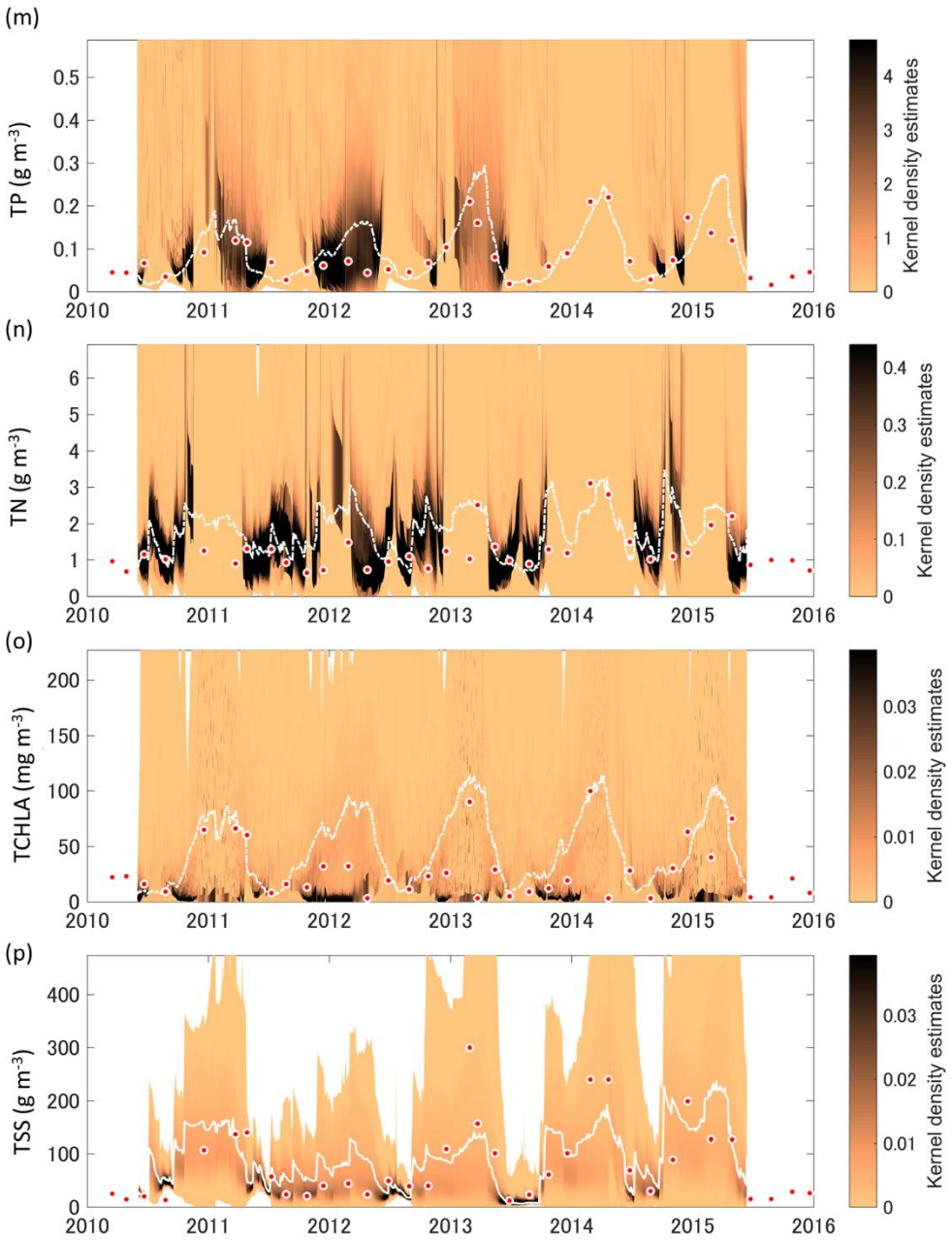


Figure 2-3. continued.

The range (spread) of the standard deviation of the variables from the model results increased when inflow volumes were relatively low (Figure 2.4). This trend was less apparent in the TCHLA results, especially in the ±25% perturbation ranges. The effect of seasonality (day of year, shown as colour gradient in Figure 2.4) was also visible, with a low standard deviation in the model outputs for the four variables occurring more frequently in winter.

Figure 2.5 illustrates the effects of one at a time (OAT) parameter modification on 5-year median outputs of TP, TN, TCHLA, and TSS from the model, with eight different proportional variations in parameters (5%, 10%, 25%, 50%, -5%, -10%, -25%, -50%). Most of the parameter changes influenced the results less than the proportion its parameter was modified (i.e. below 1:1 ratio lines). Furthermore, the majority of the changes in parameter values had a linear or no effect on the model output, but some parameters exhibited non-linear influence. The 25 to 50% parameter range changes resulted in more state variables being changed by 25 to 50%, respectively (i.e., exceeding the 1:1 ratio) but some parameters that directly altered TCHLA had reduced influence compared with the 5 or 10% parameter ranges changes (Figure 2.5-c).

Figure 2.6 shows summaries of model perturbation output (X-axis; 5-year median TP, TN, TCHLA and TSS) and selected parameter values (Y-axis; TP: temperature multiplier of sediment fluxes, TN: temperature multiplier for cyanophyte growth, TCHLA: temperature multiplier for cyanophyte growth, and TSS: critical shear stress of suspended solids). While the majority of the parameters had negligible effect on the model outputs, some showed a linear trend, while other parameters exhibited clear non-linear trends. The non-linear relationships tended to be more visible with wider parameter ranges, such as the 25% perturbation model output shown in Figure 2.6, especially with the parameters related to the temperature

multiplier for cyanophyte growth for TN and TCHLA simulation outputs. The Pearson coefficient of determination for these cases may have underestimated the significance of the relationships due to this aforementioned non-linearity (shown in Figure 2.6). Many models failed to run using vT (temperature multiplier for cyanophytes growth) as this parameter was given unrealistic values using the method of proportionately altering each parameter.

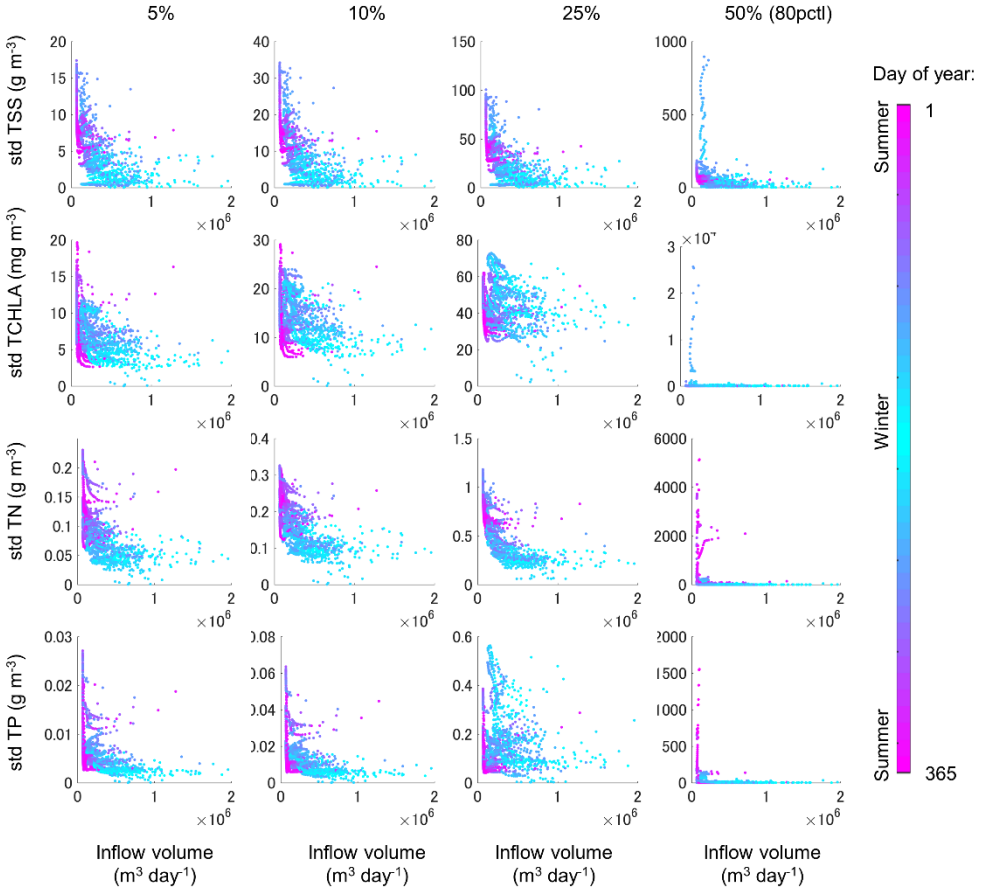


Figure 2.4. Ensemble model output (standard deviation of surface TP, TN, TCHLA and TSS, rows) versus total daily inflow volume (columns). Colour of each point corresponds to day of year (denoted by seasons in the colour gradients).

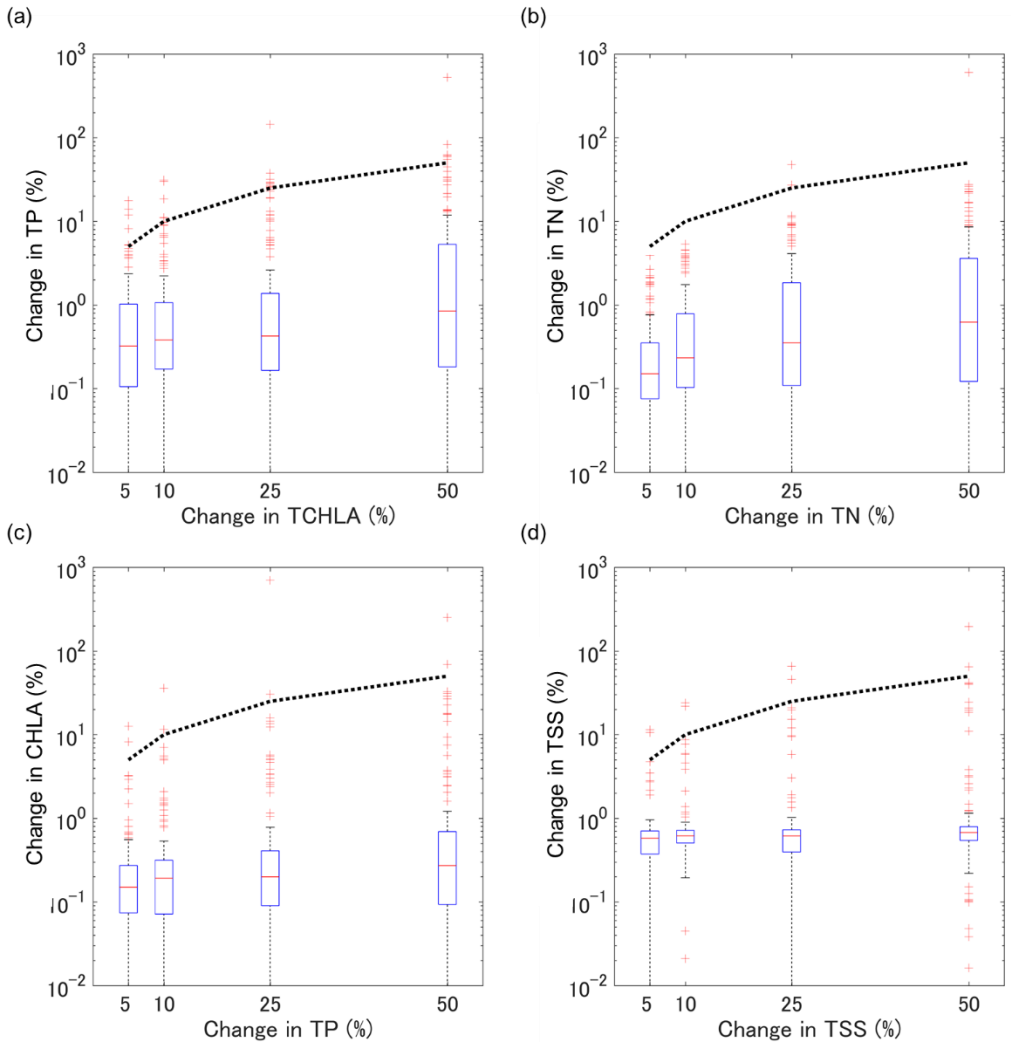


Figure 2.5. Box plot summary of influences of one-at-a-time (OAT) parameter changes (±5%, 10%, 25%, 50%) for the 5-year median surface-water model simulation outputs (a) TP, (b) TN, (c) TCHLA and (d) TSS. Changes in the results are normalised ratios using: $\frac{abs(Y|X_{\sim i}-Y|X)}{Y|X}$.

Each box represents 25th to 75th percentile results, and whiskers are theoretically 99.3% of the data in the normal distribution. The black dotted lines indicate the 1:1 ratio.

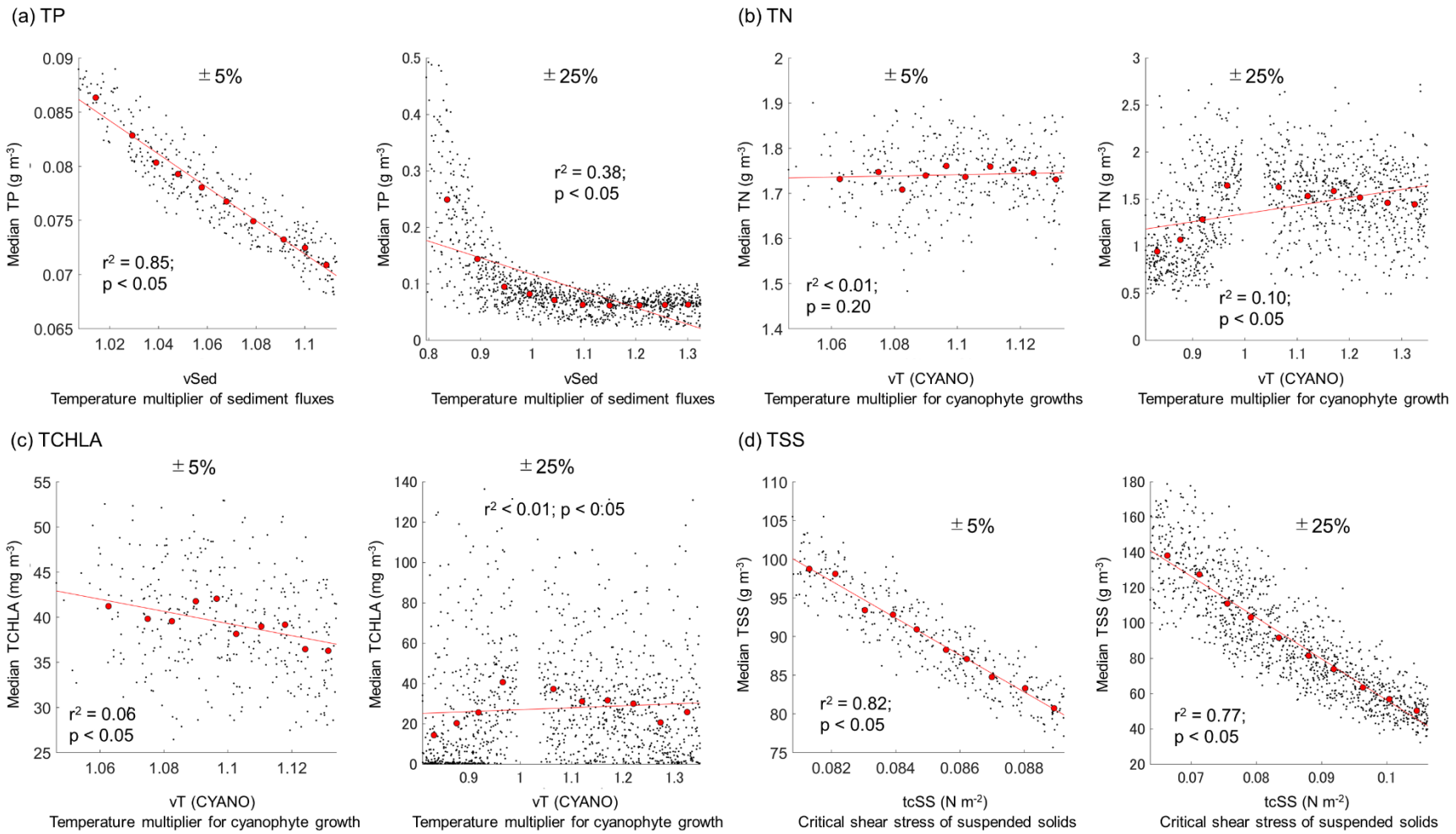


Figure 2.6: Scatterplots of ±5% and ±25% parameter perturbations versus simulation output for (a) TP against the temperature multiplier of sediment fluxes, (b) TN against the temperature multiplier for cyanophyte growth, (c) TCHLA and the temperature multiplier for cyanophyte growth, and (d) TSS and the critical shear stress for sediment resuspension. Red lines indicate least square best fit lines (and how these fail to fit in some cases), and red dots are the mean values for each ten slices which divides data points into ten equal occurrences.

Rankings of parameter sensitivity in both OAT and AAT are summarized in Figure 2.7. The ten most sensitive parameters are highlighted by colour. In the TP output, it was evident that the temperature multiplier for sediment fluxes (parameter #51) is the most sensitive parameter. In addition, almost all results except for the 50% perturbation showed that sediment PO_4 release (#54: release rate, and #55: half saturation for sediment fluxes) is strongly influences TP output. For TN, there were several sensitive parameters. While the temperature multiplier for sediment fluxes or composite resuspension rate (#51 and #62) was sensitive for smaller parameter changes, phytoplankton growth and loss processes became increasingly sensitive for larger parameter changes (±25, 50%). TCHLA results also showed that the temperature multiplier for cyanophyte fluxes was highly sensitive, as well as the temperature multiplier for cyanophyte (#15: growth, and #25: respiration) was also highly sensitive. The TSS sensitivity analysis indicated a strong effect of the suspended sediment properties (#47: density, and 48: diameter) and sediment resuspension parameters (#50: critical shear stress, and #62: composite resuspension rate).

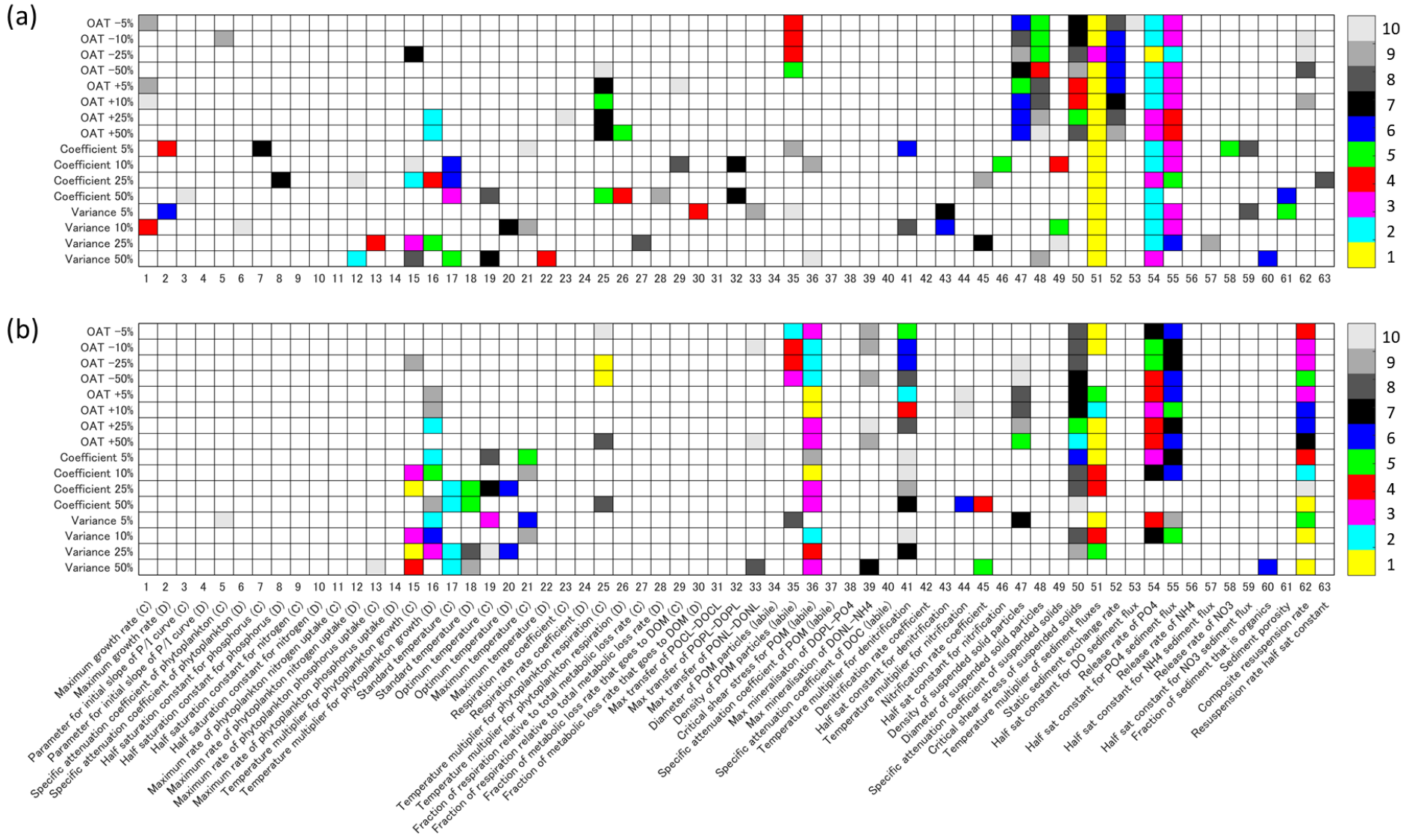


Figure 2.7: Sensitivity of 63 parameters which were adjusted within a range ±5%, 10%, 25% and 50% in simulations using one at a time (OAT) and all at a time (ATA) perturbation results in Pearson coefficient of determination and variance based impact assessment. Colours illustrate the rank of the parameter in terms of its sensitivity on the simulation output of (a) TP, (b) TN, (c) TCHLA, and (d) TSS.

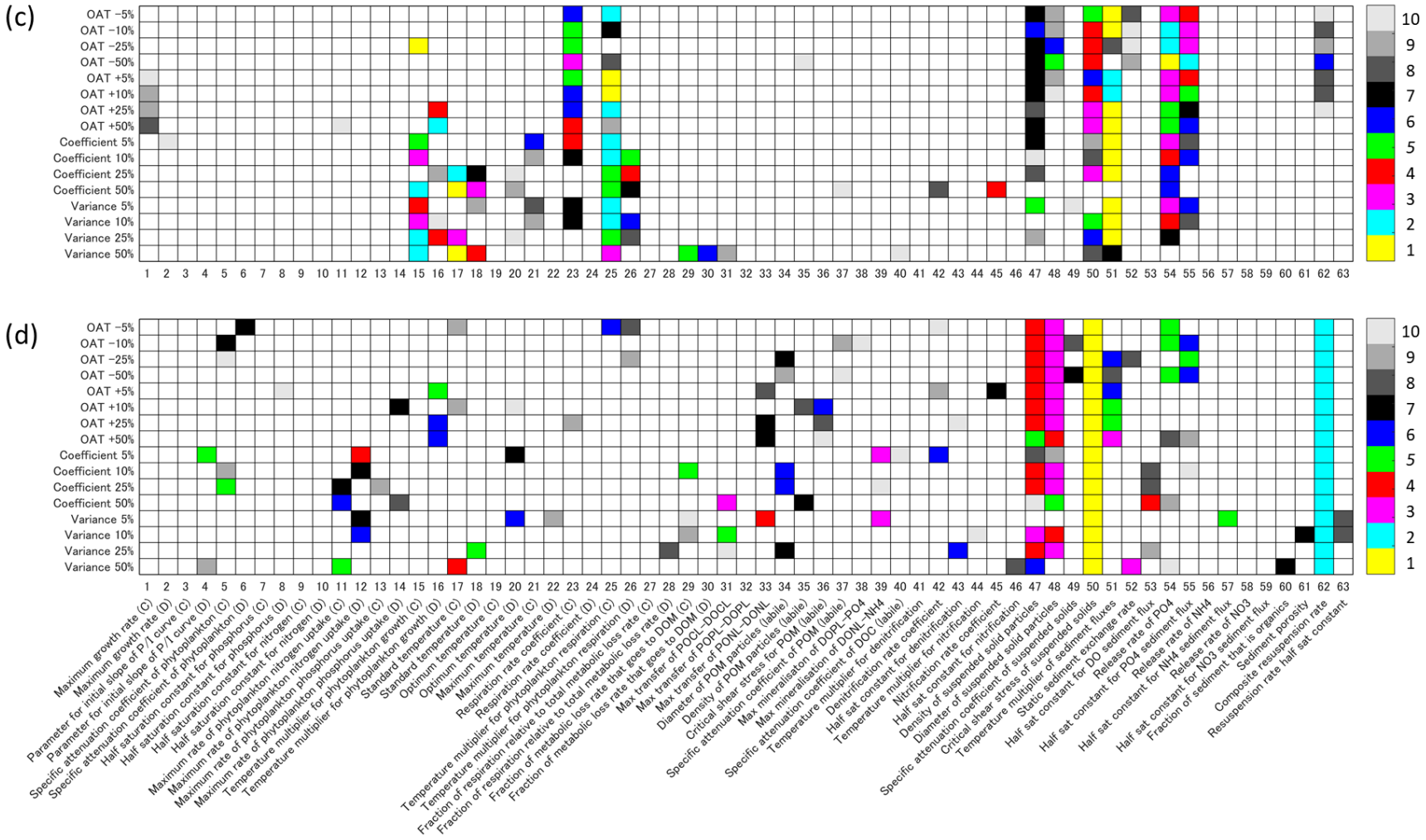


Figure 2-7: continued.

2.5 Discussion

For lake numerical ecological models to be useful forecasting tools, it is important to provide an indication of uncertainty in the model output and to demonstrate the sensitivity of the output to forcing input data and to model parameters. In a deterministic model, there is no statistical distribution but only a single model output when the parameters and input data are held constant, while an ensemble forecasting technique which deploys variable parameters offers a distribution of results designed to reflect in part the inherent elements of stochasticity and variability that are useful to demonstrate uncertainty in model output (Bellprat et al., 2011). Figure 2.8 illustrates how the recommended methods can be used in lake ecological modelling. After model development and corroboration of the validity of the model, parameters sensitivity analysis and uncertainty assessment can be performed. While the former evaluates direct contributions of each input factor to the model results, the latter evaluates results uncertainty generated by combined parameter variability or limitations of knowledge.

It is difficult to define specific parameter ranges for parameter perturbation as these can be highly dependent on the system, conditions and projects. Parameter ranges, however, are highly important in both sensitivity and uncertainty analysis. To test sensitivities of parameters, a starting point is to define minimum and maximum parameter values, which may be feasible ranges for a particular lake or project or based on theoretical or empirical studies (Makler-Pick et al., 2011; Pianosi et al., 2016). With a large number of parameters required to be calibrated in process-based lake ecological models, this is highly challenging and can become subjective. Even for a well-supported, physical process-oriented weather forecasting model, parameters and their range selections rely heavily on expert knowledge (Ollinaho et al., 2016). Considering the unique attributes of individual lake ecosystems (Hamilton et al. 2016), observations of parameters in a particular lake ecosystem will never be complete. Parameter ranges for sensitivity analysis can be defined by minimum and maximum values in literature (Figure 2.8 R1 parameter range; e.g,, Hamilton and Schladow 1997) or prior databases as in Robson et al. (2018). For parameters for which there are good information and literature resources, parameter probability densities may be created. While the ranges may be useful, model and algorithms can vary on the basis of different assumptions, thereby producing different outputs (e.g., Dugan et al., 2016). Thus, further developments may be

required to enrich the database and incorporate considerations of different modelling methods. Alternatively, small parameter perturbations from the calibrated values (Figure 2.8 B.2's R3 parameter range) can provide a satisfactory approach to address uncertainty, as I have demonstrated in this study. In the current study, parameter perturbation was expected to induce moderate shifts in output variables, for example, a slight shift of a dominating phytoplankton species/strain (which are aggregated in the model as single functional group), or or to reflect inherent environmental variability of sediment nutrient flux characteristics. Minor perturbations in parameter ranges may be sufficient to capture a realistic error envelope.

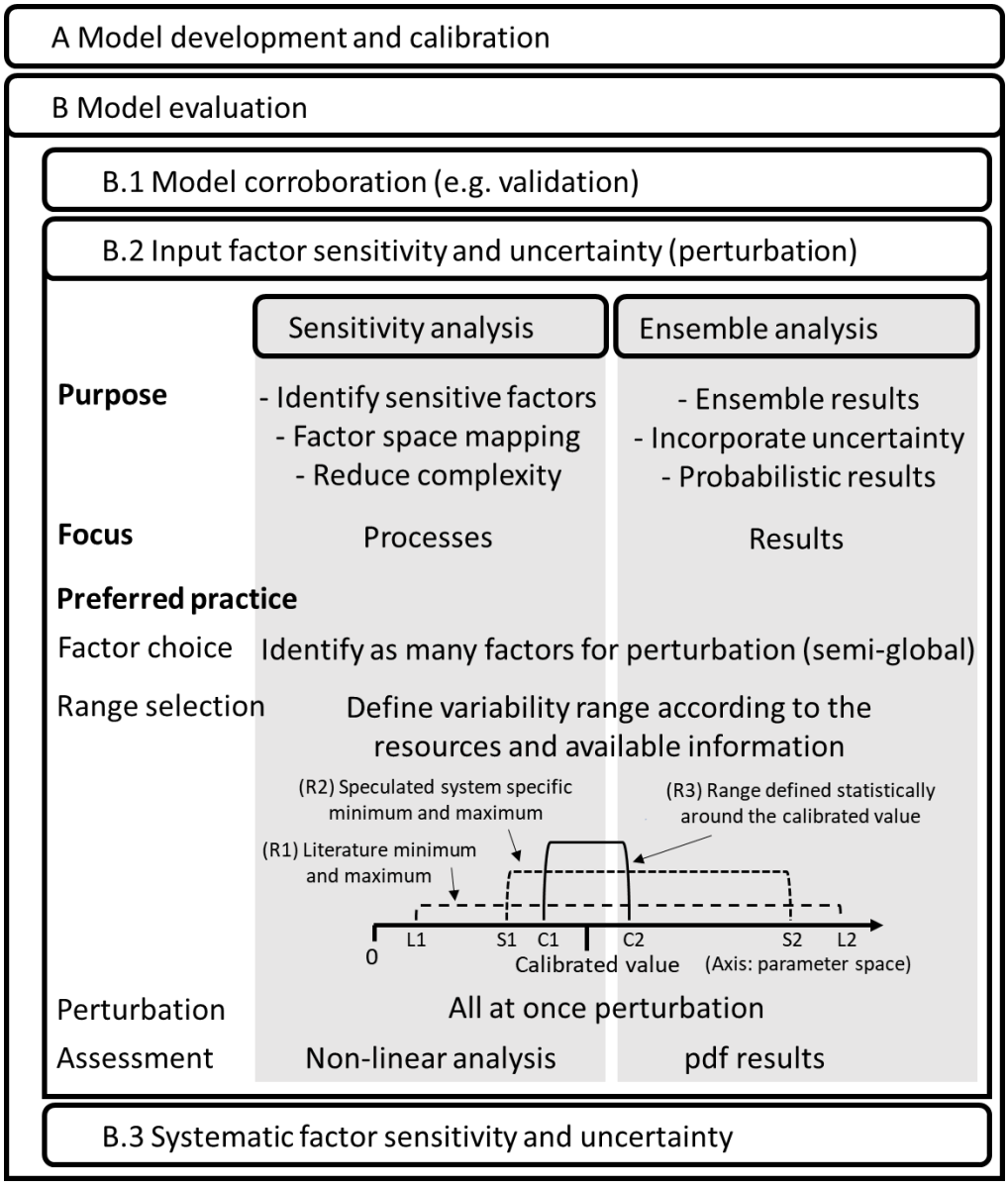


Figure 2.8: Schematic diagram to illustrate the use of parameter perturbation and sensitivity analysis as part of a modelling exercise. The ranges R1, R2 and R3 illustrate parameter variability definitions used in the analysis. If the project focus is a total sensitivity analysis, full ranges R1 or R2 should be used.

By presenting time series ensemble results with perturbed parameters, this study aimed to visualise uncertainty that occurs as a result of definition of parameter values. Seasonal trends of model outputs were preserved in the ±5% and ±10% parameter perturbation ensemble results. This indicated that external forcing factors maintained a strong influence in the model output responses to parameter changes. Parameter perturbations ≥25% produced substantially more variable results and seasonality was sometimes less evident in the output. However, the density of results (i.e., the pdf) was still useful in providing insights in these cases, e.g., the suggestion of alternative regimes of TN in the ±25% perturbation, shown as two high-density bands in the results. A high density of near-zero values of TCHLA in ±25% and 50% perturbations indicated potential for instability of the phytoplankton population with greater parameter variability.

For all output state variables and the different parameter ranges used, the temporal spread in model results was greatest when inflow volumes were lowest. This suggests that lake processes were directly or indirectly dominated by inflows during the higher flow periods and that parameters related to in-lake processes were effectively suppressed. This may be summarised as a smaller spread in the model output response to parameter perturbations when external forcing factors dominate the system. Future studies should consider including detailed analysis of inflow volumes and external nutrient concentrations, and perhaps add variation in climatic forcing, to compare and contrast sensitivity to parameters through periods of time with different environmental forcing factors.

Despite the variation in prescribed nutrient fluxes at the sediment-water interface, the pdf time series of the state variables (TP, TN, TCHLA, and TSS) did not indicate cumulative effects (e.g., accumulation of TP in the system). This is most likely because there is not a sediment nutrient storage compartment in DYRESM-CAEDYM, and residence time of Lake Waahi is short (annual average residence time ~ 0.07 years). Furthermore, even if there is accumulation of nutrients, one-off winter flushing should negate any long-term nutrient build up in the water column. For lakes with longer hydraulic residence time, internal processes influenced by model parameter values may take on greater relative importance than is smaller lakes with short residence times. For such cases, model needs to run long enough to reach spread equilibriums, and uncertainty should be assessed from this state.

Analysis of parameter sensitivity added value to the random parameter selection exercise, which was primarily carried out to acquire pdfs to represent model output uncertainty. I examined three different methods for testing parameter sensitivity. The simplest method was One at A Time (OAT) sensitivity analysis, and the others were two variations of All at A Time (AAT) sensitivity analysis based on assumptions of either linear or non-linear relationships between parameters and the model outputs. All analyses in this study were carried out locally, as I did not assess the entire theoretical range of parameter values or text the effect of extreme forcing variables. Since most complex numerical ecological models such as DYRESM-CAEDYM include non-additive equations, it would be useful to perform a higher-order sensitivity analysis (Sobol, 1993; Saltelli et al., 2007). To perform such a sensitivity analysis, parameter resampling methods would need to be carefully planned and many more model iterations may be required, or it may be necessary to create a non-linear meta-model (i.e., a model of a model) to produce these iterations. In either case, the focus of the parameter sensitivity analysis carried out in this study was to help understand the behaviour of parameters within given ranges to test what may have caused the results uncertainty.

The OAT sensitivity analysis using four different percentages of parameter perturbation revealed that the majority of the parameters had either a linear or limited effect on the state variable outputs from the model (i.e., on 5-year median surface TP, TN, TCHLA and TSS). Only a handful of parameters showed non-linear influences, approximating exponential or logarithmic relationships, on these state variables. In addition, only a few parameters influenced the outputs more than the degree to which the parameter was altered (in percentage terms) while in general, more parameters affected TCHLA results than any other output variable. The OAT and AAT parameter sensitivity rankings revealed that three parameters were highly sensitive.

In AAT analysis, the majority of the parameters had a linear effect on the state variable output, therefore the Pearson coefficient of determination and variance-based approaches produced similar results. However, positive and negative changes in parameter values exhibited different levels of sensitivity for several parameters. For example, for TP and TN, decreases in density of particulate organic matter (parameter #35, Figure 2.7-a) were more sensitive than increases. Additionally, the temperature multiplier for cyanophyte growth (parameter #15, Figure 2.6) had non-

linear effects on the phytoplankton state variable (TCHLA, Figure 2.7-c), clearly evident in the ±25% parameter range. Therefore, linearity of model output over the range of a parameter cannot be assumed and may require detailed evaluation of model responses to individual parameters.

No matter whether the focus is on input factor sensitivity uncertainty, it is important that part of the analysis should include modification of multiple factors at the same time (i.e., all at a time; AAT). An increasing number of publicly available lake ecological model applications facilitate iterative parameter selection (e.g. GLM-AED-GRAPLEr: Subratie et al., 2017; GOTM-PCLake-ACPy: http://bolding-bruggeman.com/portfolio/acpy/). Based on the outcomes of my study, I consider that model results can be highly informative to judge uncertainty with 5-10% perturbation of calibrated parameters and modest computational resources to support ~2000 model iterations. Considering the simplicity of the process, I argue that lake ecological modelling projects (especially those using 1-D models) should incorporate analysis of uncertainty and/or sensitivity as a standard operating procedure involving generating a 'cloud' of model output distributions.

Use of parameter perturbation is simple in concept (e.g. Stainforth et al., 2005; Weisheimer et al., 2011), and not very different from commonly practiced AAT parameter sensitivity analysis (e.g. Griensven et al., 2006). Uncertainty and sensitivity analysis are often interchangeably used, while lake ecological model uncertainty analysis often focuses on parameter sensitivity rather than its interactions with model forcing input. The density and envelope widths (spread; Eckel and Mass, 2005; Bellprat et al., 2012) need to be complimented with seasonal uncertainty distributions to provide a comprehensive analysis of ensemble parameter perturbation output. The analysis would provide a basis to investigate important scientific questions about lake function, productivity and metabolism, such as the relative importance of biological vs physical processes, and internal vs external drivers in the system.

Lake ecological numerical modelling needs to include aspects of uncertainty. This chapter contributed to a simple approach to include multiple parameter uncertainty. In this study, the parameter perturbation used DYRESM-CAEDYM (ver.3) model and parameters rather than alternative equations were used for the analysis. Weather and climate forecasting collates results evaluate modelling using potentially different sets of equations and model systems. Similar experimental attempts have

been made for lake ecological models in the past, but for ensemble results to be truly applicable (e.g. van Vliet et al., 2019), the model parameter calibration process needs to be objective and reproducible. With recent developments in autocalibration algorithms, an ensemble modelling approach to collectively explore parameter perturbation, model formulation and alternative equations, is foreseeable. Additionally, techniques to interpret multiple model output results using probability density functions will be valuable.

2.6 Acknowledgments

This chapter used Lake ecological Waahi model calibrated for Lehmann, M. K., D. P. Hamilton, K. Muraoka, G. W. Tempero, K. J. Collier, and B. J. Hicks. 2017. Waikato Shallow Lakes Modelling. ERI Report No. 94. Environmental Research Institute, The University of Waikato, NZ., while the report was prepared for the Waikato Regional Council and the Waikato River Authority. The model was originally prepared and calibrated by the thesis author. I aim to submit the chapter with co-authors of David Hamilton, Moritz Lehmann, Deniz Özkundakci, Piet Verburg, Liancong Luo, and Adam Harland with minor changes to the current chapter.

2.7 References

Anderson, J. L., and S. L. Anderson. 1999. A Monte Carlo Implementation of the Nonlinear Filtering Problem to Produce Ensemble Assimilations and Forecasts. Mon. Weather Rev. 127: 2741–2758. doi:10.1175/1520-0493(1999)127<2741:AMCIOT>2.0.CO;2

Booker, D. J., and R. A. Woods. 2014. Comparing and combining physically-based and empirically-based approaches for estimating the hydrology of ungauged catchments. J. Hydrol. 508: 227–239. doi:10.1016/j.jhydrol.2013.11.007

Couture, R. M., S. J. Moe, Y. Lin, Ø. Kaste, S. Haande, and A. Lyche Solheim. 2018. Simulating water quality and ecological status of Lake Vansjø, Norway, under land-use and climate change by linking process-oriented models with a Bayesian network. Sci. Total Environ. 621: 713–724. doi:10.1016/j.scitotenv.2017.11.303

Dietzel, A., and P. Reichert. 2014. Bayesian inference of a lake water quality model by emulating its posterior density. Water Resour. Res. 50: 7626–7647. doi:10.1002/2012WR013086

Elliott, A. H., A. F. Semadeni-Davies, U. Shankar, J. R. Zeldis, D. M. Wheeler, D. R. Plew, G. J. Rys, and S. R. Harris. 2016. A national-scale GIS-based system for modelling impacts of land use on water quality. Environ. Model. Softw. 86: 131–144. doi:10.1016/j.envsoft.2016.09.011

Flynn, K. J. 2005. Castles built on sand: Dysfunctionality in plankton models and the inadequacy of dialogue between biologists and modellers. J. Plankton Res. 27: 1205–1210. doi:10.1093/plankt/fbi099

Hamilton, D. P., K. J. Collier, and C. Howard-Williams. 2016. Lake Restoration in New Zealand. Ecol. Manag. Restor. 17: 191–199. doi:10.1111/emr.12226

Hamilton, D. P., and S. G. Schladow. 1997. Prediction of water quality in lakes and reservoirs. Part I - Model description. Ecol. Modell. 96: 91–110. doi:10.1016/S0304-3800(96)00062-2

Hellweger, F. L. 2017. 75 years since Monod: It is time to increase the complexity of our predictive ecosystem models (opinion). Ecol. Modell. 346: 77–87. doi:10.1016/j.ecolmodel.2016.12.001

Johnes, P. J., and a. L. Heathwaite. 1997. Modelling the impact of land use change on water quality in agricultural catchments. Hydrol. Process. 11: 269–286. doi:10.1002/(SICI)1099-1085(19970315)11

Kingett, P. D. 1984. Lake Waahi: an environmental history. A report prepared by Kingett and Associates for the Mines Division Ministry of Energy. 201 p.

Kroese, D. P., T. Brereton, T. Taimre, and Z. I. Botev. 2014. Why the Monte Carlo method is so important today. Wiley Interdisciplinary Reviews: Computational Statistics, 6(6), 386-392. doi:10.1002/wics.1314

Lehmann, M. K., D. P. Hamilton, K. Muraoka, G. W. Tempero, K. J. Collier, and B. J. Hicks. 2017. Waikato Shallow Lakes Modelling. ERI Report No. 94. Environmental Research Institute, The University of Waikato, NZ.

Liu, Y., Y. Wang, H. Sheng, and others. 2014. Quantitative evaluation of lake eutrophication responses under alternative water diversion scenarios: A water quality modeling based statistical analysis approach. Sci. Total Environ. 468: 219–227. doi:10.1016/j.scitotenv.2013.08.054

Makler-Pick, V., G. Gal, M. Gorfine, M. R. Hipsey, and Y. Carmel. 2011. Sensitivity analysis for complex ecological models – A new approach. Environ. Model. Softw. 26: 124–134. doi:10.1016/j.envsoft.2010.06.010

Malve, O., M. Laine, H. Haario, T. Kirkkala, and J. Sarvala. 2007. Bayesian modelling of algal mass occurrences-using adaptive MCMC methods with a lake water quality model. Environ. Model. Softw. 22: 966–977. doi:10.1016/j.envsoft.2006.06.016

Ollinaho, P., S. Lock, M. Leutbecher, and others. 2017. Towards process-level representation of model uncertainties: stochastically perturbed parametrizations in the ECMWF ensemble. Q. J. R. Meteorol. Soc. 143: 408–422. doi:10.1002/qj.2931

Oreskes, N., K. Shrader-Frechette, and K. Belitz. 1994. Verification, validation, and confirmation of numerical models in the earth sciences. Science 263.5147 (1994): 641-646.. doi:10.1126/science.263.5147.641

Pianosi, F., K. Beven, J. Freer, J. W. Hall, J. Rougier, D. B. Stephenson, and T. Wagener. 2016. Sensitivity analysis of environmental models: A systematic review with practical workflow. Environ. Model. Softw. 79: 214–232. doi:10.1016/j.envsoft.2016.02.008

Refsgaard, J. C., J. P. van der Sluijs, J. Brown, and P. van der Keur. 2006. A framework for dealing with uncertainty due to model structure error. Adv. Water Resour. 29: 1586–1597. doi:10.1016/j.advwatres.2005.11.013

Robson, B. J., G. B. Arhonditsis, M. E. Baird, and others. 2018. Towards evidence-based parameter values and priors for aquatic ecosystem modelling. Environ. Model. Softw. 100: 74–81. doi:10.1016/j.envsoft.2017.11.018

Saltelli, A., M. Ratto, T. Andres, F. Campolongo, J. Cariboni, D. Gatelli, M. Saisana, and S. Tarantola. 2007. Global Sensitivity Analysis. The Primer, John Wiley & Sons, Ltd. West Sussex, England.

Schlabing, D., M. A. Frassl, M. M. Eder, K. Rinke, and A. Bárdossy. 2014. Use of a weather generator for simulating climate change effects on ecosystems: A case study on Lake Constance. Environ. Model. Softw. 61: 326–338. doi:10.1016/j.envsoft.2014.06.028

Schladow, S. G., and D. P. Hamilton. 1997. Prediction of water quality in lakes and reservoirs: Part II - Model calibration, sensitivity analysis and application. Ecol. Modell. 96: 111–123. doi:10.1016/S0304-3800(96)00063-4

Sobol, I. M. 1993. Sensitivity estimates for nonlinear mathematical models. Math. Model. Comput. Exp. 1: 407–414.

Trolle, D., D. P. Hamilton, C. A. Pilditch, I. C. Duggan, and E. Jeppesen. 2011. Predicting the effects of climate change on trophic status of three morphologically varying lakes: Implications for lake restoration and management. Environ. Model. Softw. 26: 354–370. doi:10.1016/j.envsoft.2010.08.009

U.S. Environmental Protection Agency. 2009. Guidance on the Development, Evaluation, and Application of Environmental Models. USEPA Publ. EPA/100/K-: 90.

Van Vliet, M. T. H., Flörke, M., Harrison , J. A., Hofstra, N., Keller, V., Ludwig, F., Spanier, J. E., Strokal, M., Wada, Y., Wen, Y., Williams, R. J., 2019. Model inter-comparison design for large-scale water quality models. Curr. Opin. Env. Sust. 36, 59–67. doi: 10.1016/j.cosust.2018.10.013.

Whitehead, P. G., R. L. Wilby, R. W. Battarbee, M. Kernan, and A. J. Wade. 2009. A review of the potential impacts of climate change on surface water quality. Hydrol. Sci. J. 54: 101–121. doi:10.1623/hysj.54.1.101

Xiao, M., M. P. Adams, A. Willis, M. A. Burford, and K. R. O'Brien. 2017. Variation within and between cyanobacterial species and strains affects competition: Implications for phytoplankton modelling. Harmful Algae 69: 38–47. doi:10.1016/j.hal.2017.10.001

Supplementary Table 2-1

Parameter	Description	Units	Base Values
Pmax (CYANO)	Maximum growth rate	day^{-1}	0.9
Pmax (FDIAT)	Maximum growth rate	day^{-1}	1.1
IK (CYANO)	Parameter for initial slope of P/I curve	$\mu E\ m^2\ s^{-1}$	100
IK (FDIAT)	Parameter for initial slope of P/I curve	$\mu E\ m^2\ s^{-1}$	20
Kep (CYANO)	Specific attenuation coefficient of phytoplankton	$\mu g\ Chla\ L^{-1}\ m^{-1}$	0.014
Kep (FDIAT)	Specific attenuation coefficient of phytoplankton	$\mu g\ Chla\ L^{-1}\ m^{-1}$	0.014
KP (CYANO)	Half saturation constant for phosphorus	$mg\ L^{-1}$	0.0002
KP (FDIAT)	Half saturation constant for phosphorus	$mg\ L^{-1}$	0.001
KN (CYANO)	Half saturation constant for nitrogen	$mg\ L^{-1}$	0.017
KN (FDIAT)	Half saturation constant for nitrogen	$mg\ L^{-1}$	0.065
UNmax (CYANO)	Maximum rate of phytoplankton nitrogen uptake	$mg\text{-}N\ mg\text{-}Chla^{-1}\ day^{-1}$	3
UNmax (FDIAT)	Maximum rate of phytoplankton nitrogen uptake	$mg\text{-}N\ mg\text{-}Chla^{-1}\ day^{-1}$	3.3
UPmax (CYANO)	Maximum rate of phytoplankton phosphorus uptake	$mg\text{-}P\ mg\text{-}Chla^{-1}\ day^{-1}$	0.3
UPmax (FDIAT)	Maximum rate of phytoplankton phosphorus uptake	$mg\text{-}P\ mg\text{-}Chla^{-1}\ day^{-1}$	0.24
vT (CYANO)	Temperature multiplier for phytoplankton growth	no units	1.08
vT (FDIAT)	Temperature multiplier for phytoplankton growth	no units	1.06
Tsta (CYANO)	Standard temperature	ºC	20
Tsta (FDIAT)	Standard temperature	ºC	20
Topt (CYANO)	Optimum temperature	ºC	34
Topt (FDIAT)	Optimum temperature	ºC	29
Tmax (CYANO)	Maximum temperature	ºC	39
Tmax (FDIAT)	Maximum temperature	ºC	34
kr (CYANO)	Respiration rate coefficient	day^{-1}	0.07
kr (FDIAT)	Respiration rate coefficient	day^{-1}	0.16
vR (CYANO)	Temperature multiplier for phytoplankton respiration	no units	1.06
vR (FDIAT)	Temperature multiplier for phytoplankton respiration	no units	1.1
fres (CYANO)	Fraction of respiration relative to total metabolic loss rate	no units	0.1
fres (FDIAT)	Fraction of respiration relative to total metabolic loss rate	no units	0.7
fdom (CYANO)	Fraction of metabolic loss rate that goes to DOM	no units	0.1
fdom (FDIAT)	Fraction of metabolic loss rate that goes to DOM	no units	0.3
POC1max	Max transfer of POCL-DOCL	day^{-1}	0.01
POP1max	Max transfer of POPL-DOPL	day^{-1}	0.05
PON1max	Max transfer of PONL-DONL	day^{-1}	0.08
POMDia1	Diameter of POM particles (labile)	m	0.0000008
POMDensity1	Density of POM particles (labile)	$kg\ m^{-3}$	1050
tcPOM1	Critical shear stress for POM (labile)	$N\ m^{-2}$	0.03
KePOC1	Specific attenuation coefficient of POM (labile)	$mg\ L^{-1}\ m^{-1}$	0.047

DOD1max	Maximum mineralisaton of DOPL-PO4	day^{-1}	0.05
DON1max	Maximum mineralisation of DONL-NH4	day^{-1}	0.08
KeDOC1	Specific attenuation coefficient of DOC (labile)	mg L^{-1} m^{-1}	0.01
vN2	Temperature multiplier for denitrification	no units	1.07
KoN2	Denitrification rate coefficient	day^{-1}	0.25
KN2	Half saturation constant for denitrification	mg L^{-1}	6.5
vON	Temperature multiplier for nitrification	no units	0.1
KoNH	Nitrification rate coefficient	day^{-1}	0.1
KON	Half saturation constant for nitrification	mg-O L^{-1}	2.5
deSS	Density of suspended solid particles	kg m^{-3}	600
diaSS	Diameter of suspended solid particles	m	0.0000016
KeSS	Specific attenuation coefficient of suspended solids	mg^{-1} Lm^{-1}	0.01
tcSS	Critical shear stress of suspended solids	N m^{-2}	0.085
vSed	Temperature multiplier of sediment fluxes	no units	1.06
rSOs	Static sediment exchange rate	g m^{-2} day^{-1}	1.5
KSOs	Half saturation constant for DO sediment flux	mg L^{-1}	0.2
SmpPO4	Release rate of PO_4	g m^{-2} day^{-1}	0.02
KOxS-PO4	Half saturation constant for PO_4 sediment flux	g m^{-3}	1
SmpNH4	Release rate of NH_4	g m^{-2} day^{-1}	0.01
KDOS-NH4	Half saturation constant for NH_4 sediment flux	g m^{-3}	6.5
SmpNO3	Release rate of NO_3	g m^{-2} day^{-1}	-0.05
KDOS-NO3	Half saturation constant for NO_3 sediment flux	g m^{-3}	8.5
sedOrganicFrac	Fraction of sediment that is organic	no units	0.04
SedPorosity	Sediment porosity (porewater fraction)	no units	0.54
resusRate	Composite resuspension rate	g m^{-2} day^{-1}	0.055
resusKT	Resuspension rate half saturation constant	g	100000000

Chapter three

A data mining approach to evaluate suitability of dissolved oxygen sensor observations for lake metabolism analysis

This chapter was published as:

Muraoka, K., P. Hanson, E. Frank, M. Jiang, K. Chiu, and D. Hamilton. 2018. A data mining approach to evaluate suitability of dissolved oxygen sensor observations for lake metabolism analysis. Limnol. Oceanogr. Methods 16: 787–801. doi:10.1002/lom3.10283

4.1 Abstract

Despite rapid growth in continuous monitoring of dissolved oxygen for lake metabolism studies, the current best practice still relies on visual assessment and manual data filtering of sensor observations by experienced scientists in order to achieve meaningful results. This time-consuming approach is fraught with potential for inconsistency and individual subjectivity. An automated method to assure the quality of data for the purpose of metabolism modelling is clearly needed to obtain consistent results representative of collective expertise. I used a hybrid approach of expert panel and data mining for data filtration. Symbolic Aggregate approXimation (SAX) treats discretised numerical time series segments as symbolic indications, creating a series of strings which are literally comparable to human words and sentences. This conversion allows established text mining

techniques, such as classification methods to be applied to time series data. Half-hourly frequency surface dissolved oxygen data from 18 global lakes were used to create day-long segments of the original time series data. Three hundred sets of one-day measurements were provided to a group of seven anonymous experts, experienced in manual filtering of oxygen data for metabolism modelling studies. The collective results were treated as expert panel decisions and were used to rank the data by confidence level for use in metabolism calculations. While considerable variation occurred in the way the experts perceived the quality of the data, the model provides an objective and quantitative assessment method. The program output will assist the decision-making process in determining whether data should be used for metabolism calculations. An R version of the program is available for download.

4.2 Introduction

Ecosystem metabolism is an important and fundamental ecological concept. Many attempts have been made to numerically quantify its kefy components, productivity and respiration, for lake ecosystems across the world (Cole et al., 2000; Solomon et al., 2013). Ecosystem metabolism may be a proxy for trophic status and can be used to understand whether a lake is a source or sink of carbon (Hanson et al., 2003). As lake monitoring has become increasingly intensive and automated around the world (Weathers et al., 2013; Hamilton et al., 2015), the use of metabolism models to assess ecosystem functioning will likely grow.

Metabolism models in lakes typically assume that a change in free-water dissolved oxygen (DO) through time is driven primarily by the balance between photosynthesis (or primary production) and mineralization of organic carbon (often called 'respiration' for simplicity), as well as equilibration of DO with the atmosphere (Staehr et al. 2010). When these three processes are dominant, diel DO patterns will be nearly sinusoidal, with increases during daylight due to primary production exceeding respiration and decreases at night due to respiration. However, additional processes, such as inflow and outflow to and from a lake, vertical and horizontal mixing, and advective movement of water mass can affect the balance of DO within specific lake strata (Antenucci et al., 2013; Rose et al. 2014) or between littoral and pelagic zones (Lauster et al. 2006; Van de Bogert et al. 2007; Batt et al. 2012). When DO is measured using in situ sensors, these processes can impart patterns on the DO data that obscure the signal from biological processes and that, if left unaccounted, can introduce noise and bias into the estimate of metabolism (Rose et al. 2014).

Generalizable approaches are needed for separating signals due to biological processes from those derived from physical processes to better quantify and de-bias

lake metabolism estimates. For high frequency lake sensor observations, some attempts have been made to automate and standardize the methods of QA/QC (e.g. general QA/QC - Horsburgh et al., 2015) and calculation protocols (e.g., physical stability - Read et al., 2011; energy flux - Woolway et al., 2015; lake metabolism - Winslow et al., 2016). Experts have commonly removed data considered to be irrelevant noise or error, by visual assessment (e.g., Solomon et al., 2013), and in some cases have developed formalized approaches for evaluating uncertainty in metabolism predictions, as well as model parameters, and have identified the circumstances associated with those uncertainties (Rose et al. 2014, Cremona et al. 2014, Giling et al. 2017). While the aforementioned approaches have proven useful in evaluating metabolism predictions, they are subject to the overhead and constraints of coding parametric process-based models, and in some cases, the undocumented criteria of expert opinion. An alternative is to formalize the inclusion of expert knowledge on metabolism and use that knowledge, along with data-driven approaches, in efficient, flexible, and reproducible ways for data QA/QC.

Time series analysis, filtering and data mining offer a set of solutions that may be particularly useful for evaluation of DO data intended for metabolism modelling (Niennattrakul et al., 2009; Rakthanmanon et al., 2011). Preparation for time series analysis should be comprised of three components operating either independently or simultaneously: QA/QC, data dimensionality reduction, and data representation / approximation. Increasing dimensionality (information), which is inherent in increased sampling frequency from sensors, decreases performance of similarity or distance-based discovery algorithms (e.g. more difficult to build a robust model; Aggarwal et al. 2001; Zimek et al., 2012). This can be circumvented by removing some data or compressing the amount of information processed (Cannata et al., 2011) or by representing data in a simpler form (Keogh et al., 2001). Spectral

analysis, such as Discrete Fourier Transformation (DFT) and Discrete Wavelet Transformation (DWT) are two examples that have been used in recent limnological contexts (e.g., Cengiz, 2011; Kara et al., 2012; Cox et al., 2015). Techniques to accurately define "suitable data" have not been generalised but any methods needs to be robust and repeatable.

A promising technique that enables simplification of data while retaining key properties is Symbolic Aggregate approXimation (SAX; Lin et al. 2003). SAX has similarities to Piecewise Linear Approximation (PLA) and Piecewise Aggregate Approximation (PAA), which extract key information from complex time series data (Ratanamahatana et al., 2005). PLA and PAA divide time series into segments of equal or unequal length, and calculate segment trends or means for each segment. SAX uses arithmetic mean values of even length segments (PAA), and further bins the segmented values into defined categories, creating a series of discrete letter sequences (words) from the original numeric time series (Lin et al., 2003; Lin et al., 2007). The SAX transformation enables the user to create a dictionary of time series subsequences similar to DNA sequences, making it possible to rapidly search for coherence in the time series vocabulary space. This means SAX transformation allows typical time series numerical observations to be able to utilize well-developed text mining techniques, which have generally not been included within the domain of the environmental time series analysis discipline. SAX analysis, due to its piecewise approach, is suitable for noisy and/or variable time series data common in environmental settings. SAX has been used in multiple disciplines such as video detection (e.g. Ma et al., 2016) and has recently been used in limnology to identify fluorescence signal patterns (Ruan et al., 2017).

The main objective of this study is to identify procedurally meaningful DO time series patterns from high frequency sensor data and provide a filter enabling

identification and removal of complex data to improve the accuracy and consistency of lake metabolism calculations. The approach is designed to be reproducible and allow for automated classification of data quality that is consistent with expert opinion. To achieve this, the steps involved were: (1) generation of time series labels through expert evaluation, (2) transformation of time series data using the SAX method, and (3) supervised classification. I used a subsampled dataset from 18 lakes to generate and test the classification model.

4.3 Methods

Eighteen lakes with suitable datasets (e.g. high-frequency preliminary QA/QCed surface DO, temperature profiles and wind speed) for model training were selected from the Global Lake Ecological Observatory Network (GLEON) lakes. The majority of the data were reused from Solomon et al. (2013) (Table sA). The parent dataset contained 4,852 days with dissolved oxygen data, ranging from 132 – 434 days for individual lakes. To make labelling by experts feasible, random subsampling was used to obtain 300 days of data from the parent dataset, including 7 to 30 days (median 18 days) from individual lakes. For consistency, all time series data were downsampled to 30-min frequency.

Seven scientists at a conference were approached for their expertise in lake metabolism studies, i.e., experience with screening these datasets. The 300 days of subsampled data were provided to the experts as time series of DO over each day. Also included were supplementary figures comprising time series of water column temperature profile, wind speed, and photosynthetically active radiation (PAR), as well as the timing of sunrise and sunset, as these data could be used to further inform the experts about the quality of the data and the relevant processes. The group members were asked to evaluate which specific days of DO time series data were suitable for lake metabolism analysis, based on their experience and inspection of the visualised dataset. Three questions were asked of the experts in relation to each dataset: (Q1) "Would you use this DO data for metabolism studies?", (Q2) "Did biological processes dominate the metabolism signal represented in DO?", and (Q3) "Other than DO, what data influenced your Q1 decision?". Four choices were provided as options to Q1, namely [Yes], [Maybe Yes], [Maybe No], and [No]. The responses of the experts were aggregated and turned into labels for each day, based on eight possible classes: Y0, Y1, Y2, Y3, Y4, Y5, Y6, and Y7. For example, data

were labelled as Y7 (best class) if all seven scientists selected either options [Yes] or [Maybe Yes], and as Y0 (worst class) if all scientists selected [No] or [Maybe No]. This method was used to provide an independent quantitative expert evaluation of the level of confidence in the quality of the data. The survey results were analysed according to: (Q1) frequency of expert agreement, (Q2) whether usability of DO data was related to the dominance of biological processes in the DO signal, and (Q3) whether experts indicated additional data would have helped to refine Q1.

Labels Y0-Y7 from the expert panel assessment were used to build classification models after SAX transformation of each day of data. A diagram of this process is shown in Figure 4.1. The classification models used R libraries rWeka (ver. 0.4-34; Hornik et al., 2009 & Witten and Frank, 2005), RWekajars (ver. 3.9.1-3; Hornik, 2018), RJava (Simon, 2017) and shiny (ver. 1.0.3; Chang et al., 2017) on R (ver. 3.4.1). R was selected as the main framework since it is widely used in ecology and data mining disciplines and is open-source software. WEKA (Waikato Environment for Knowledge Analysis) is specialised data mining and machine learning software (Hall et al., 2009). Both rWeka and RWekajars are APIs (application program interface) in the R language platform that enable use of a variety of data mining resources through WEKA toolsets. Shiny is a library allowing the creation of a user-friendly front-end for the models.

To prepare data for the supervised classification, I followed the protocol described by Lin et al. (2007) for the SAX transformation. The SAX transformation combines two time series transformation methods that reduce the dimensionality of the data: piecewise averaging and data binning (Figure 4.2). The piecewise averaging method, also known as Piecewise Aggregate Approximation (PAA), segments the original time series data (measured at a 30-minute resolution) into n equal time

periods for which an average value is derived. For example, PAA applied to a 24-h DO dataset with n=4 will contain an average value for each of the four 6-h time segments. Similarly, data binning (into m bins) was used to segment DO values. For example, with m = 2 data bins, DO values can be defined as being ≥ or < a specified breakpoint value (Table sB). The binned data therefore holds ordinal information rather than nominal or numeric values. Each bin is represented by a letter in the processed version of the time series so that the original numeric times series becomes an alphabetic string.

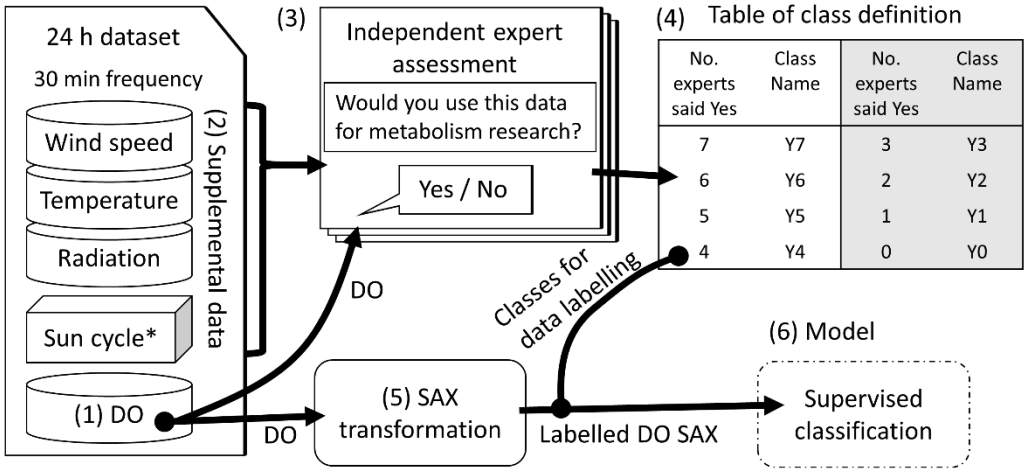


Figure 4.1 The workflow for generation of the classification model. Three hundred days of dissolved oxygen concentration (DO) at 30 min frequency were provided (1) to seven independent experts, along with supplementary data (2). Experts labelled the data (3), which was then collated and allocated according to classes (Y7 to Y0) representing the number of experts that said "Yes" to the data being useful (4; answers "maybe yes" and "maybe no" were aggregated to Yes and No respectively). The identical three hundred days of DO time series data were also transformed (5) by Symbolic Aggregate approXimation (SAX), and (6) a classification model was created using (5) to reproduce the labels (4). Sun cycle includes sunrise and sunset timing.

The SAX transformation was carried out after normalizing the original DO daily time series using a standard mean transformation:

$$DO\_norm = (DO - \mu) / \sigma \quad (1)$$

where DO_norm is the normalized DO time series, μ is the arithmetic mean of DO and σ is the standard deviation of DO for the day. Breakpoints were identified by splitting the DO_norm values into equal percentile probabilities assuming a standard normal distribution (see Table sB). Once the m-1 breakpoints were identified using PAA, and thus a mapping from DO_norm values to letters of an alphabet with m letters established, DO_norm was averaged for each of the n time periods and turned into alphabetic representation by looking up the appropriate bin in the list of m bins. The lowest numerical values of DO_norm were given the letter "a". Assuming, for example, m = 3, the largest numerical values would be given the letter "c". I express a SAX transformation with n PAA segments (corresponding to the size of the "words" that will represent each time series) and m bins (the size of the alphabet) as SAX(n,m). I deployed twenty-five SAX parameter sets (n = 2 – 6; m = 2 – 6) to examine performance of SAX against expert opinion. An R algorithm by Ruan et al. (2017), which uses the classic SAX technique, was used.

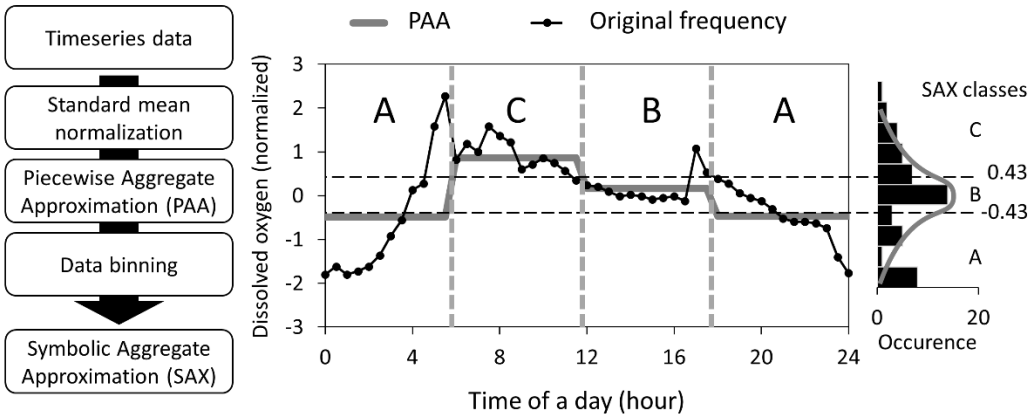


Figure 4.2: Schematic of the SAX transformation. The graph (middle) shows an example of normalised dissolved oxygen (DO_norm) data at 30 min intervals (black line with dots), its PAA results at 6 h intervals (thick vertical grey dashed lines) and SAX letters according to the breakpoints given in Supplementary table 4-2 (dashed lines; 0.43 and -0.43). In this example the SAX word length (n) is 4 and there are 3 letters (m) corresponding to the two breakpoints. The right histogram shows the distribution of the data with the grey line representing an idealized normal distribution. The SAX transformation processes are shown on the left-hand

side. In this case, the data consists of the following SAX letter combinations: [A, B, C, AC, BA, CB, ACB, CBA, ACBA].

In this study, I used the SAX-transformed time series to formulate a classification problem by associating each transformed series with one of the eight labels (Y0-Y7) generated from expert input. More specifically, I created an ordinal classification problem because the eight labels exhibit a natural order. Standard supervised learning algorithms cannot exploit this ordering information without converting the classes into numeric values. To overcome this issue, my model creates the following seven two-class problems: [Y0 | Y1 – Y7], [Y0 – Y1 | Y2 – Y7], [Y0 – Y2 | Y3 – Y7], [Y0 – Y3 | Y4 – Y7], [Y0 – Y4 | Y5 – Y7], [Y0 – Y5 | Y6 – Y7], and [Y0 – Y6 | Y7] where the threshold "|" separates the first and second binary class, i.e., unsuitable and suitable data respectively. For brevity, I use the notation Y0-1 to refer to the two-class problem [Y0 | Y1 – Y7], and so on for other classes. Based on this model setting, class probability estimates from the seven two-class models, one for each threshold, were combined to obtain multi-class probability estimates for all eight categories for each test sequence, assigning the sequence to the class with maximum probability. The method proposed by Frank and Hall (2001), in conjunction with the smoothing method from Schapire et al. (2002), was used to combine the two-class probability estimates into multi-class probability estimates. This process was implemented in the OrdinalClassClassifier procedure that is available in R via RWeka. To compare the sensitivity of SAX parameters to the model performance, I examined the model performance using the seven two-class problems.

Logistic regression, the classification technique I apply to my data, requires numeric input rather than strings of letters. I established the numeric features by computing subsequence frequencies for each sequence of letters to be classified. More specifically, for a SAX(n,m) model, which generates strings of length n

consisting of m letters, I count how often each of the $\sum_{i=1}^{n} m^i$ theoretically possible subsequences occurs in the sequence to be classified (as I only considered subsequences consisting of consecutive letters). The set of subsequence counts are used as the predictor variables in the logistic regression model.

Due to the available SAX parameter combinations, twenty-five candidate models were generated and tested for their performance. The model performance was evaluated in the form of the binary classes (suitable and unsuitable) for each of the seven two-class problems discussed above. To measure performance, I used Area Under the ROC Curve (AUC), and Matthews Correlation Coefficient (MCC). I also considered a confusion matrix for the classification problem to obtain additional insight. A confusion matrix is a frequency distribution table of the test data instances, illustrating how instances of class X are assigned to class Y by the classification model. A confusion matrix for a two-class problem shows the following frequencies: TP (True Positives), TN (True Negatives), FP (False Positives) and FN (False Negatives).

A receiver operating characteristics (ROC) curve shows the true positives rate (TPR = TP / (TP + FN)) and false positives rate (FPR = FP / (FP + TN)) in two-dimensional space (Witten et al., 2016; Bradley, 1996). Each TPR/FPR point in this space is obtained by applying a different classification threshold on the class probability estimates obtained from the classification model. To summarize the information in the curve, the area under the curve (AUC) is used as a performance measure. It can be shown that AUC corresponds to the estimated probability that a randomly chosen positive test instance is ranked above a randomly chosen negative test instance when the classifier's class probability estimates for the positive class are used to rank the test instances. AUC is less sensitive to the relative frequency of the two classes (positive and negative) than simple TPR or FPR measures,

allowing direct comparison across different threshold settings. Model performance is considered to be perfect if AUC = 1, and random if AUC = 0.5.

Matthews correlation coefficient (MCC, proposed by Matthews, 1975) is an alternative accuracy classification that is not affected by imbalanced class distributions. MCC is a discrete version of the Pearson correlation coefficient, varying from 1 (perfect fit) to 0 (no fit). Negative values are also possible if "anti-learning" has occurred. MCC is calculated as:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \qquad (2).$$

Both AUC and MCC information were used to determine appropriate models. Evaluation of AUC and MCC was carried out in a ten-fold cross-validation process to estimate performance of the full data model; i.e., they were estimated by the average of the 10 results obtained from a rotated 10% data split validation (Kohavi, 1995). These model evaluations were examined in the two-class models, while the full confusion matrix provided insights into the combined multi-class model.

4.4 Results

4.4.1 Data exploration and subsampling

The ten most frequently recurring daily DO SAX sequences of the parent dataset (4,582 days of DO data) are shown in Figure 4.3. The results are from the SAX(4,3) transformation ("candidate models" section explores different SAX transformation results), i.e., with 6-h resolution (n = 4) and two thresholds (m = 3). Recurrent patterns occur across several of the lakes and most of these patterns start with the letter 'a' (i.e., the bin with the lowest normalized DO). The sequence "aacc" (i.e., DO is low in the first half of the day (0–12 hr) and high in the second half of the day (12-24 hr)) is the most frequently occurring pattern in nine lakes and "abcb" (i.e., DO rises through the first three quarters of the day (0-18hr), and then decreases in the fourth quarter (18-24 hr)) is the most frequent pattern in three lakes. The letters are not randomly distributed, suggesting the feasibility of categorization of daily DO observations based on letter sequences alone. While the SAX(4,3) transformation theoretically results in $3^4 = 81$ possible full day sequences, the parent dataset includes only 54 (66.7%) of these patterns (Table 4.1A). Considering the substantial size of the parent data used, the parent data patterns in small SAX parameters thought to include all idealised DO curves driven by biological activities, and therefore those theoretical patters that did not appear in the parent datasets are primarily "noisy". This coverage decreases as the number of possible SAX strings increases. The lowest coverage is found in SAX(6,6), where 4% of the available sequences appeared in the parent data. An exception occurs for m = 2, where the parent dataset coverage generally increases when n increases. It is noteworthy that the differences in coverage are primarily determined by alphabet size rather than word length.

Table 4.1: (A) Percentages (%) of full day DO SAX(n,m) sequences that appeared in the parent dataset (N = 4582) in comparison to all the possible combination of letters (m^n) in various number of word size (n) and alphabet (m) settings. (B) Percentages of full day DO SAX(n,m) unique sequences that appeared in the training dataset in comparison to parent dataset patterns in various number of word size (n) and alphabet (m) settings. (C) Percentages of parent data incidents (i.e. number of days of N = 4582) covered by training dataset in terms of SAX sequence.

A: Parent dataset coverage (sequence)		Alphabet size				
		2	3	4	5	6
Word size	2	75.0	77.8	50.0	52.0	33.3
	3	75.0	70.4	39.1	43.2	27.3
	4	87.5	66.7	47.7	36.0	26.2
	5	93.8	63.8	38.4	22.2	13.1
	6	92.2	48.8	21.0	9.0	4.1
B: Training dataset coverage (sequence)		Alphabet size				
		2	3	4	5	6
Word size	2	66.7	71.4	62.5	61.5	50.0
	3	100.0	94.7	92.0	72.2	78.0
	4	100.0	72.2	57.4	45.3	38.2
	5	76.7	49.7	31.8	22.8	19.0
	6	61.0	28.7	18.4	15.3	12.5
C: Training dataset coverage (incidents)		Alphabet size				
		2	3	4	5	6
Word size	2	100.0	99.9	99.9	99.9	99.8
	3	100.0	99.8	99.6	96.7	96.3
	4	100.0	96.8	90.2	82.2	76.7
	5	97.7	87.9	75.4	58.5	52.6
	6	96.8	78.7	60.2	46.8	29.1

The coverage of sequences observed in the subsampled parent dataset (300 days) is summarised in Table 4.1 B. For SAX(4,3), 39 (72%) of the 54 sequences identified in the parent dataset are present. A higher proportion of parent data patterns are covered in the subsampled data when both SAX parameters are small. Both alphabet size and word length similarly affect the training (subsampled 300 days) data coverage of the parent data. Table 4.1 C shows the proportion of parent data full-day SAX sequences represented in the training data. For the SAX(4,3) setting, over 95% of parent data sequences are represented, leaving 149 instances not represented in the training data. Similarly, the majority of parent data sequences are

included for most of the SAX parameter settings, while two SAX parameter settings (SAX(6,5) = 47%; SAX(6,6) = 29%) fail to represent more than one-half of the instances.

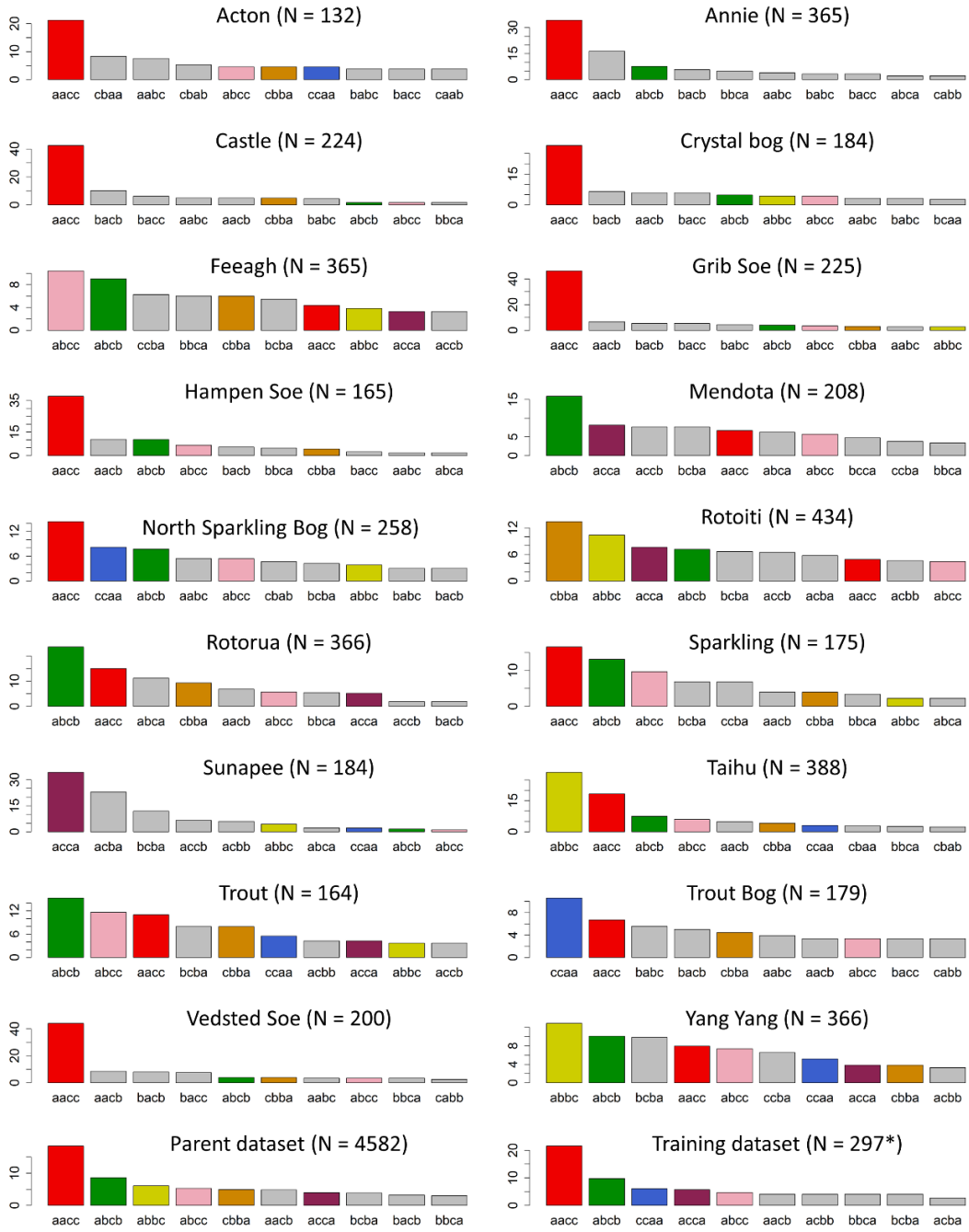


Figure 4.3: The ten most frequently recurring sequences of daily DO SAX letters from eighteen lakes as well as parent and training (subsampled 300 days) datasets are shown in proportion to the entire data used (Y axis: frequency of occurrence). For this, SAX transformation was parameterised with SAX(4,3); three letters (a, b, c) and 4 segments a day. Theoretically there are $3^4 = 81$ possible sequences. The seven sequences that occurred most frequently across the set are highlighted with colors to aid intuitive recognition of their frequency of detection (aacc-red; abcc-pink; abcb-green; cbba-orange; acca-violet; abbc-yellow; ccaa-blue). Parent (all lake) and training datasets are also shown. Two sequences abbc-yellow and cbba-orange that did not show up in the top ten training data have instances of seven and eight respectively appearing in the training data.

4.4.2 Survey results

The seven experts rated 34 to 80% (average 60%) of the 300 daily training data as suitable for lake metabolism analysis. For the threshold separation Y3-4 ([Y0 – Y3 | Y4 – Y7]), an average of 62% of the training data was labelled as "suitable" (Figure 4.4). The highest number of data instances was recorded in Y7 (n = 73), with 32 instances on average for the other classes (min = 23, max = 48, Figure 4.4). Figure 4.5 illustrates 'suitable' 30-min DO data according to the expert panel results and assigned thresholds. The available 'suitable data' reduces as the threshold level increases, but it is evident that noise in the data are filtered out through the expert panel evaluation process. The experts chose options [Yes] and [No] without "maybe" 85% of the time, while classes Y3 to Y5 contained more "maybe" responses. Survey results for Q1 (Would you use this DO data for metabolism studies?) and Q2 (Did biological processes dominate the metabolism signal represented in DO?) were strongly positively correlated. The survey results for Q3 (Other than DO, what data influenced your Q1 decision?) indicated that 29.6% of the time, the majority of the experts used one or more supplementary data sources for their assessment, but the type and number of supplementary data varied. The number of times the panel requested additional data was 20 for PAR, 37 for wind speed, 1 for diel solar radiation, 0 for surface temperature, 48 for temperature profile, and 0 for other information. On thirteen occasions, the panel cited two additional sources of supplementary data as being required to make their decision on Q1.

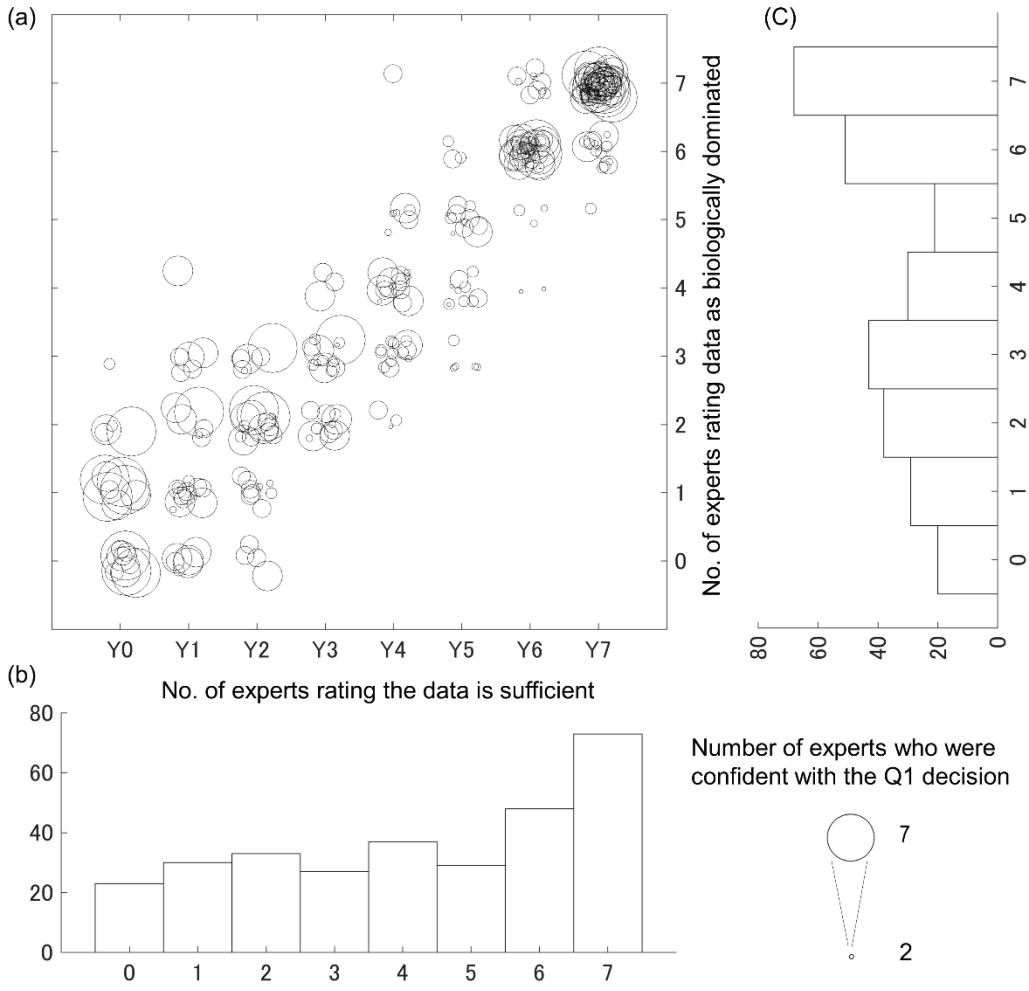


Figure 4.4: (a) Scatter diagram: number of experts indicating that daily data is biologically dominated vs data adequacy (Y0 to Y7) based on number of experts indicating 'Yes'. Circles are plotted with a small degree of randomness (0.25 jitter) to reduce visual data overlap of the discrete values, and size of the circles reflects the number of experts who were confident with their individual decision (Pearson's correlation coefficient: 0.87; $p < 0.01$). Histograms compliment the scatterplot to indicate frequency distribution of experts indicating that DO data were adequate (b) and that DO data were biologically dominated (c).

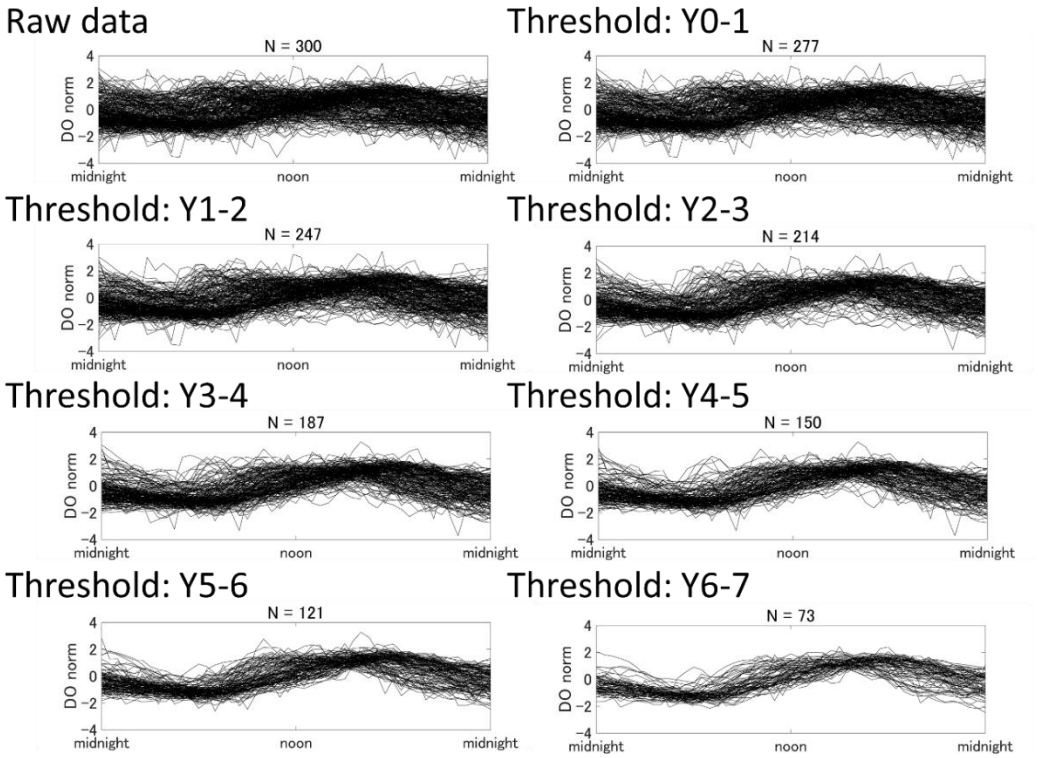


Figure 4.5: The 'good data' consisting of 30 min interval time series over one day and classified according to the expert panel decision. Seven thresholds are shown. N represents number of days that were classified as having "good data".

4.4.3 Candidate models

Parameter selection and threshold analysis were attempted through a model selection process. Twenty-five models were created based on different SAX parameter combinations (SAX(n, m)), i.e., n = 2 - 6 corresponding to 12 - 4 h intervals, respectively, and m = 2 – 6 corresponding to one to five threshold values to separate DO data. Table 4.2 - 4 show ten-fold cross validation results of the model performance using MCC and AUC metrics. I examined all possible combination of SAX parameters as well as thresholds, while shown here are what I consider as useful information. Table 4.2 gives the results of various models when the threshold was set to Y5-6. Table 4.3 provides the results with the SAX alphabet number fixed to 3, i.e., SAX(n, 3), and Table 4.4 indicates the results with SAX word length fixed to 4, i.e., SAX(4, m). For Y5-6 ([Y0-Y5 | Y6-Y7]), only SAX(4,3) appeared amongst the top five results based on both MCC and AUC analyses. For

the different threshold settings using SAX(n, 3), n = 4 performed better in both MCC and AUC analyses. For the models with SAX(4,m), m = 3 results ranked in top 5 performance in both MCC and AUC analysis.

Table 4.2: Ten fold cross validated model performances in terms of Mathews correlation coefficient (MCC, top) and area under the receiver operating characteristic curve (AUC, bottom) with various SAX word sizes and number of SAX alphabet, where threshold was fixed to Y5-6 ([Y0-Y5 | Y6-Y7]). Numbers in bold represent the top 5 results in the table.

MCC		Alphabet size				
		2	3	4	5	6
Word size	2	0.53	0.45	0.35	0.44	0.40
	3	0.28	0.41	0.44	0.41	0.34
	4	0.47	**0.63**	0.56	0.47	0.55
	5	0.53	**0.59**	0.51	0.51	**0.64**
	6	**0.64**	0.49	0.52	**0.58**	0.53
AUC		Alphabet size				
		2	3	4	5	6
Word size	2	0.72	0.73	0.76	0.74	0.75
	3	0.72	0.73	0.77	0.75	0.72
	4	0.79	**0.88**	**0.87**	0.80	0.82
	5	0.81	0.84	0.81	0.80	**0.88**
	6	0.83	**0.85**	**0.85**	0.82	0.84

Table 4.3: Ten fold cross validated model performances in terms of Mathews correlation coefficient (MCC, top) and area under the receiver operating characteristic curve (AUC, bottom) with various SAX word sizes and threshold settings, where size of SAX alphabet was fixed to 3. Numbers in bold represent the top 5 results in the table.

MCC		Threshold						
		Y0-1	Y1-2	Y2-3	Y3-4	Y4-5	Y5-6	Y6-7
Word size	2	0.13	0.34	0.49	0.51	0.43	0.45	0.00
	3	0.30	0.51	0.46	0.43	0.50	0.41	0.00
	4	0.29	0.50	0.52	**0.61**	**0.63**	**0.63**	0.36
	5	0.23	0.48	0.46	0.58	**0.59**	**0.59**	0.26
	6	0.34	0.42	0.47	0.59	0.50	0.49	0.27
AUC		Threshold						
		Y0-1	Y1-2	Y2-3	Y3-4	Y4-5	Y5-6	Y6-7
Word size	2	0.65	0.71	0.74	0.77	0.74	0.73	0.71
	3	0.74	0.81	0.78	0.75	0.72	0.73	0.67
	4	0.71	0.71	0.82	**0.84**	**0.86**	**0.88**	0.77
	5	0.71	0.77	0.77	0.83	0.81	**0.84**	0.76
	6	0.65	0.62	0.71	0.76	0.81	**0.85**	0.77

Table 4.4: Ten fold cross validated model performances in terms of Mathews correlation coefficient (MCC, top) and area under the receiver operating characteristic curve (AUC, bottom) with various number of SAX alphabet and threshold settings, where SAX word size was fixed to 4. Numbers in bold represent the top 5 results in the table.

MCC		Threshold						
		Y0-1	Y1-2	Y2-3	Y3-4	Y4-5	Y5-6	Y6-7
Number of alphabets	2	0.24	0.54	0.57	**0.59**	0.46	0.47	0.07
	3	0.29	0.50	0.52	**0.61**	**0.63**	**0.63**	0.36
	4	0.17	0.48	0.49	0.58	**0.67**	0.56	0.30
	5	0.24	0.37	0.42	0.52	0.49	0.47	0.36
	6	0.25	0.36	0.46	0.33	0.53	0.55	0.26
AUC		Threshold						
		Y0-1	Y1-2	Y2-3	Y3-4	Y4-5	Y5-6	Y6-7
Number of alphabets	2	0.62	0.72	0.78	0.82	0.77	0.79	0.74
	3	0.71	0.71	0.82	**0.84**	**0.86**	**0.88**	0.77
	4	0.63	0.71	0.82	**0.84**	0.83	**0.87**	0.79
	5	0.69	0.63	0.75	0.82	0.80	0.80	0.78
	6	0.72	0.74	0.76	0.75	0.80	0.82	0.74

An extended confusion matrix using SAX(4,3) is shown in Figure 4.6. For example, for the threshold Y3-4 ([Y0-Y3 | Y4-Y7]), 174 instances were correctly classified as 'suitable data' (TP) and 13 instances were wrongly classified as 'unsuitable data' (FN). This means that of the 187 instances of 'suitable data' (for the Y3-4 threshold), the model correctly labelled 93% instances. Conversely, TN = 85 and FP = 28, which means 75% of the 'unsuitable data' was correctly classified as unsuitable.

Figure 4.7 shows six normalized DO time series of instances with extreme errors (i.e., expert labels Y0 – Y2 were classified as Y7). The fact that the classifier mis-classifies these SAX sequences (aacc, abcb, bcca) implies that they appeared repeatedly in the training dataset and their corresponding DO time series were frequently identified by the expert panel as Y7. Inspection of the plots in Figure 4.7 shows that the likely reasons for the mis-classifications are: (1) appearance of repeated values over a part of the day, (2) low variations of DO values, and (3) obvious increase in DO before sunrise. Figure 4.8 illustrates DO data at 30-minute

intervals for days when data are classified as ‘suitable’ according to the SAX(4,3) model. The SAX(4,3) model generally overestimated the amount of ‘suitable data’ in each threshold compared with the expert panel labels (Figure 4.5). The number of instances of ‘suitable data’ classified into the higher threshold levels was greater than the expert panel decisions in favour of those thresholds (i.e. errors for Y5-6 and Y6-7 were 31 and 45, respectively), but the number of classifications is similar for the lower thresholds (mean error for Y0-1 to Y4-5 was 14).

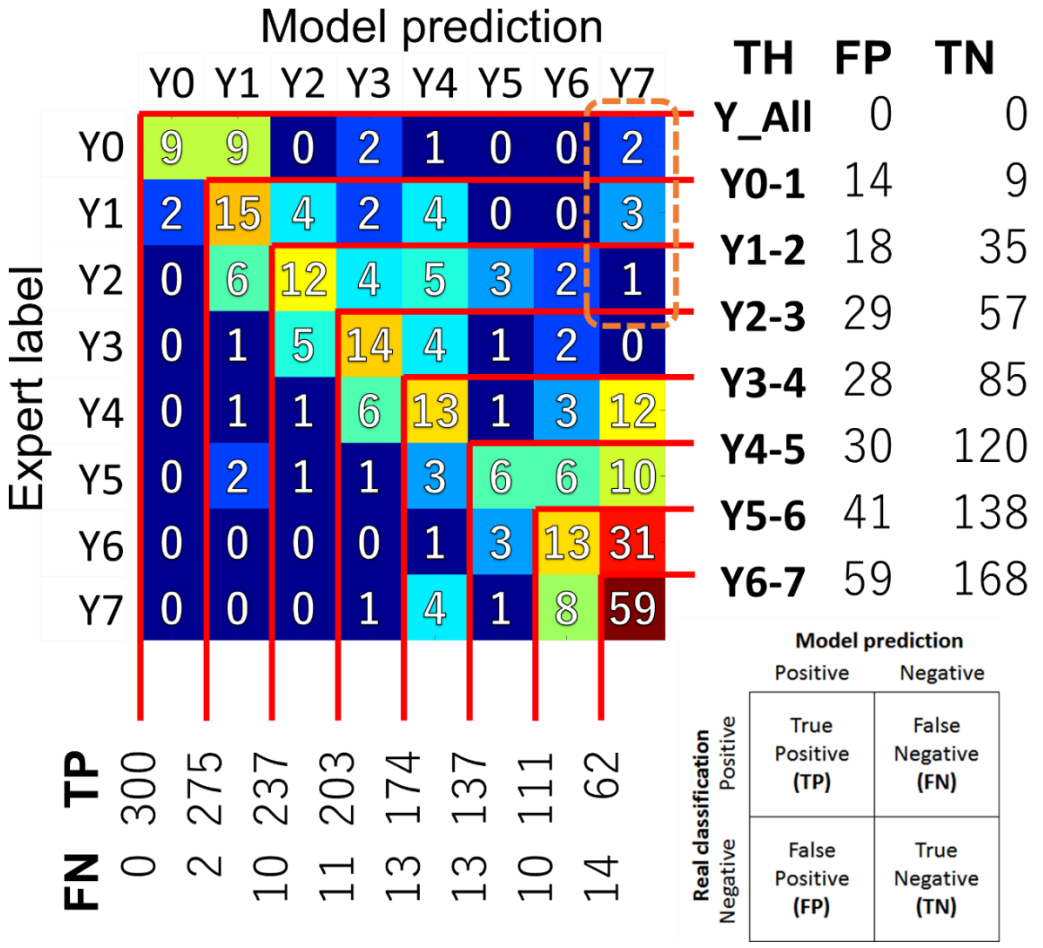


Figure 4.6: All training data model results for eight classes Y0 - Y7 in relation to the extent of expert agreement, where Y7 (y-axis) corresponds to the full consensus on the use of the data. Red lines illustrate the binary class threshold settings, and for each threshold, True Positive (TP), False Positive (FP), True Negative (TN), and False Negative were calculated. The schematic figure at the bottom right shows the basic structure of a confusion matrix for a two-class problem. TH stands for class threshold, and extreme errors (orange dashed box) are explored in Figure 4.7. The colour was added to provide visual realization of the number.

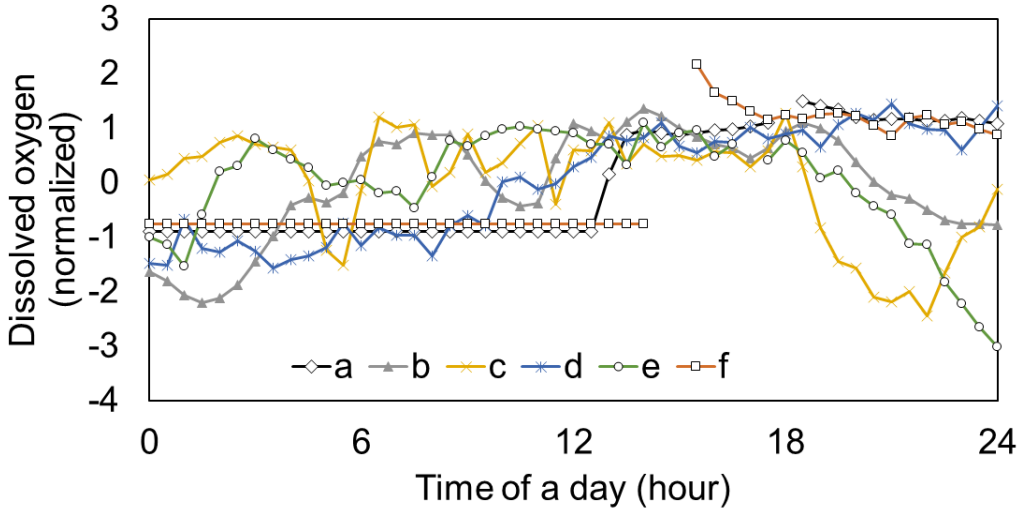


Figure 4.7: Six time series of normalized DO indicated as extreme errors in Figure 4.6. SAX(4,3) for each time series were: a) aacc; b) abcb; c) bcca; d) aacc; e) bcca; f) aacc. Variations of DO in mg L-1 (max - min) for each series were a) 0.28; b) 1.13; c) 0.38; d) 0.07; e) 0.90; f) 2.03, and standard deviations were a) 0.12; b) 0.32; c) 0.10; d) 0.02; e) 0.22; f) 0.69. Suspected causes of errors included: repeated values (a, f), increase of DO before the sunrise (b, e), and low variation of data (a, c, d).

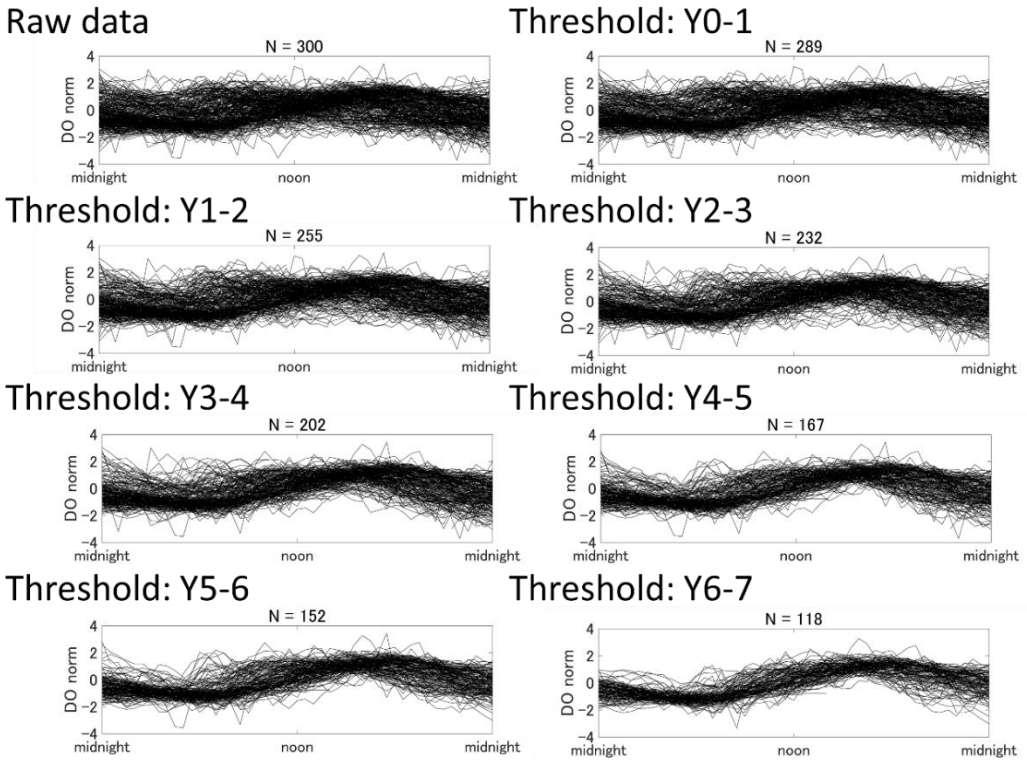


Figure 4.8: Data classified as 'good' for 30-min interval time series over one day according to the SAX(4,3) model results and with seven thresholds. N is the number of data classified as 'good data'.

4.5 Discussion

Humans have a great capacity for detecting visual patterns (e.g. Cox et al., 1997), and our approach to evaluating the suitability of DO data for metabolism models exploits that capacity. Simply increasing the amount of data used to estimate metabolism with HF observations reduces the risk of extreme and inaccurate estimations. However as previous work has shown, DO time series are often messy and have complex patterns, and teasing-apart the underlying causes of noise and bias in metabolism estimates made from DO signals is still challenging (Rose et al. 2014, Cremona et al. 2014, Giling et al. 2017). My method provides an alternative to the more classical approach of parametric analysis by basing the classification of DO signals on expert knowledge, as well as the patterns inherent in the DO data. The aforementioned approaches focus on the suite of processes, whereas my approach focuses on the suite of patterns. The SAX approach (1) assumes no particular metabolism model and therefore results do not depend on themodel or parameters used; (2) adopts pattern recognition and classification techniques as opposed to a mass balance approach for dissolved gasses that may not include all relevant fluxes; (3) uses the collective knowledge of experts to train the classifier based on a large dataset including other information besides DO, and (4) provides flexibility to apply the method for other purposes without further calibration.

4.5.1 A framework for labeling data

Our primary goal was to design a framework to provide simple labelling of suitable and unsuitable (i.e. suitable and unsuitable for free surface metabolism models) segments in high frequency autonomous DO data. Black-and-white expert panel decisions were not made, as the experts had different interpretations and expectations about the data provided to them. Consequently, my model provides a semi-qualitative, but informative judgement about: “how many experts would

support the quality of data", by reproducing labels Y0 to Y7. A user is then given the freedom to choose a threshold decision level based on their expectation of confidence in the data quality and the number of suitable data available for analysis. While this leaves a degree of variation in the data products, scientists may have different expectations for cleaning data. For a metabolism model study, for example, if a higher threshold (such as Y6-7) is used, the classification model may output data that has mostly idealised shapes of DO over a diurnal cycle (i.e., alternation of dominance by production and respiration). Such a case may be expected to occur where changes from transport and mixing are of lesser significance than biological processes. The selection of high levels of confidence might, however, restrict the number and frequency of data available for use by the metabolism model, and may well disregard specific features that occur in reality. Conversely, if a lower threshold such as Y3-4 is used, more data will be available, but the user will need to take a cautious approach to the interpretation of results since by definition, reducing the Y value reflects reduced levels of expert panel confidence in the data quality.

4.5.2 SAX as transformation for data QA/QC and analysis

With increasing use of autonomous in-situ sensors, and the resulting large volume and high complexity of observations, it is increasingly difficult to manage, archive and analyse data. Ecologists would therefore benefit from embracing approaches that meld simple models with machine learning. The conventional QA/QC approach for evaluating data from autonomous sensor networks is no longer practical due to the rapid expansion of sensors, networks and 'big data' generally (Campbell et al. 2013). Common QA/QC tasks typically focus on network or sensor malfunctions, such as missing values, sensor drift, or inconsistency. Observations of DO can be susceptible to these types of malfunctions, but further filtering is

necessary for more complex tasks, for example, assessment of lake metabolism. SAX is a simple transformation of time series data. The transformed data also provides a different way to think about environmental data as a sequence of words, and the unique approach opens up opportunities for additional analytical tools, such as a text sequence mining approach to analyse (dis)similarity of sequence patterns.

DO signal variations are known in both intra-lake data (driven mostly by seasonality) and inter-lake data (driven mostly by latitude, trophic status and geology) (e.g. Richardson et al., 2017). These variations might be evident as different magnitudes of variation and frequency of peaks and troughs in the data, and could also be influenced by differences in sunrise/sunset timing, or complex balances of productivity, respiration, transportation, diffusion and surface gaseous fluxes. The latter information (i.e., magnitude and rate of a change) was explicitly not used to filter out the data in my classification model, as this would potentially give bias to the metabolism model. Many other variables may also require complex QA/QC processes, as well as filtering, to make sense of the data. For example, phycocyanin sensors require consideration of factors that affect the assumed linearity between cyanobacteria biomass and phycocyanin, such as the proportion of colonial or filamentous populations, temperature or species-specific signals (Chang et al., 2012). In other words, relationships between environmental sensor readings and the data of interest may require specific knowledge or sensor conversions to assess and extract information relevant to the variable of interest (Kara et al., 2012).

Classification of time series data is challenging as the data usually exhibit high dimensionality and are inherently noisy (Keogh and Kasetty, 2003; Hanson et al., 2008). To overcome this, a classification model should only be provided with appropriate information extracted from the data. For my case, the essence of the

data is the shape of the time series. The variations of DO peaks and troughs and the timing of these requires a robust analytical procedure. The SAX transformation is simple in concept but has proven useful in many applications (Lin et al., 2007). In essence, SAX removes quantitative uncertainties and only preserves the general shape of time series data. Symbols, as a result of data normalisation and binning in SAX, are equiprobable. In other words, the probability of occurrence of each symbol is likely to be equal, on the assumption that the values in the time series are normally distributed. This provides good coverage when using SAX to detect shapes and trends in sequences by applying string sequence classification methods (Ratanamahatana et al., 2005). SAX symbols are robust and clean due to the segmented approximation process (PAA). As a result, most of the variability in the observed data can be represented semi-quantitatively. PAA also reduces the computational memory and run time used for classification, and thus allows for comparison of multiple models in a computationally efficient way. In the most extreme case in this study, data were reduced in number from 1440 samples per day to 4 (in case of SAX(4,3)).

4.5.3 Generalizability of the SAX and expert opinion approaches

The SAX transformation of DO for eighteen lakes resulted in discovery of relatively consistent diurnal DO sequences across most lakes. I emphasize that my evaluation is not of the lakes, per se, but of the collection of DO patterns likely to be encountered at the daily scale that have relevance to metabolism. To be clear, I am not evaluating the mean or variance of DO, but rather the specific patterns. Given a dimensionality of SAX(4,3) which was found to be most effective at reproducing expert classification, there are 81 possible patterns. Only about one half of the patterns were found in the data, and of these, 7-10 patterns accounted for most of the occurrences (Fig. 3). Put another way, relatively few patterns account for most

of those that are found in daily DO across a broad range of lakes, and the model training data cover most of the available patterns. Thus, I can expect that my analysis will apply to lakes not in this study if they present patterns that are in the diverse collection herein, which I feel is likely.

Having verified that SAX is appropriate for the globally-distributed lakes in my study, it may be appropriate in the future to examine the drivers of differences in DO patterns amongst lakes, using, for example, environmental and morphological drivers such as season, lake trophic state, climate, location, lake shape and depth. In addition, the data used in my study is localised to the level of time zone that a lake is within, but not precisely to geographical location. This may have caused a minor inconsistency in the temporal alignment of the patterns, and if a lake's longitude differs from the time zone longitude, a small adjustment to the observation time may be required to apply the model appropriately.

Like all methods, the combined use of expert opinion with SAX transformation has limitations. String similarity discovery models use multiple combinations of letters in a word as the model attribute. For example, in a four-letter word, one can examine the frequency of occurrence of two-letter (e.g., [a][a], [c][b]) or four-letter sequences (e.g., [a][a][c][b]) or the occurrences of two separated letter combinations in the word (e.g. [a][*][c]). When the size of the words or available alphabet size increases, the number of possible patterns used as model attributes increases exponentially. This results in a need for greater computational memory and runtime. To limit this from happening, models often deploy n-gram tokenizers that provide a minimum and maximum number of letter combinations for model inputs (Whitelaw et al., 2009). In my study, the SAX letter length and alphabet size for daily DO transformations did not create a major demand on computing resources, as the maximum resource demanded was for SAX(6,6) with

$\sum_{i=1}^{6} 6^i = 55{,}986$ predictor variables. For larger datasets, one would require further consideration of the maximum SAX sequences relevant to the frequency and length of the data of interest.

When there was unanimity (Y0 or Y7) amongst the experts, decisions were generally made without 'maybe'. This confirms similar underlying logic that the experts used to determine the data quality. The Y7 label (full consensus; indicating suitable-quality data) was by far the largest populated class, indicating a high occurrence of "textbook quality" data in the observations. The survey also revealed, however, that experts had different expectations about the quality of data. This was evidenced in the large variation between experts in the data that was selected to be suitable (ranging between 34% – 80%). Without a full consensus amongst the expert panel, it is difficult to make a strong judgement about what is suitable and unsuitable data. It is also unreasonable to disregard any expert's opinion simply because it is in the minority, hence ordinal type expert panel decisions Y0 to Y7 were created from the survey instead of a majority decision, to express confidence in the data by the expert panel.

Model performance assessment metrics require careful consideration. For example, for habitat niche distribution model evaluation practices, Lobo et al. (2007) suggest stating the true positive rate (TPR) and true negative rate (TNR) in addition to the AUC values, to further reduce the chance of class imbalance biases. I adapted two different metrics of AUC and MCC together when choosing the model, where the MCC method contains concepts of both TPR and TNR. Both metrics suggested that SAX(4,3) gives the appropriate set of transformation parameters, providing validation in the use of this model. With a training data set of 300 days, higher orders in the SAX transformation, e.g., SAX(6,6), would raise concerns regarding over-fitting of the model. While overfitting may be a concern even with SAX(4,3),

the 10 folds cross-validation suggests this complexity of SAX is a reasonable compromise between goodness-of-fit and generalizability. The confusion matrix provided useful insights about the model, and it also identified a few extreme error occurrences. The causes of extreme errors may be due to simple QA/QC type issues (e.g., appearance of partially repeated values and low levels of DO variation; Figure 4.7a, c, f), or lack of information to drive the model (e.g., DO increased before sunrise; Figure 4.7b, e). While the former QA/QC issue can be easily filtered out, the latter type of error requires additional information to help identify the specific nature of the problem. It should also be noted that the tool identifies repeated values or the occurrence of no data for an entire day as a specific sequence (i.e., aaaa), which is likely to be classified as Y0. One extreme error instance was not caused by such errors, but was most probably due to simplification made by SAX(4,3) (Figure 4.7d). Identification of extreme errors (if they exist) can be helped by comparing class differences between the model outputs of SAX(4,3) and other parameter sets such as SAX(6,6). Nevertheless, the occurrence of such extreme errors related to a lack of information or SAX simplification was minimal (<1%).

4.5.4 Conclusions

G.E. Hutchinson (1957) wrote that a "skilful limnologist can probably learn more about the nature of a lake from a series of oxygen determinations than from any other kind of chemical data. If these oxygen determinations are accompanied [by additional variables], a very great deal is known about the lake". This quote illustrates how a highly complex lake system can be evaluated with a series of dissolved oxygen observations, integrating both biogeochemical and physical processes. Most numerical modelling practices involve attempting to capture the majority of these processes using a highly complex set of equations and using a comprehensive dataset (e.g., Robson 2014), but such models themselves demand

considerable time, effort and expertise. Theory-guided data science (e.g., Karpatne et al., 2016) tries to capture the essence of the system (i.e., the information contained within the system) by using a less complex model structure, which may be appropriate when the dataset is incomplete. The DO data provided to the classification model in my study effectively represented a minimum level of information (DO shape) but the training dataset was complimented by an expert survey process which involved supplementary data. The use of classes together with additional information commonly requested by the experts (PAR, wind speed and surface water temperature) is supported by the model performance. The expert survey results tended to confirm that the removal of data was predominantly due to the influence of non-biological processes. Most lake metabolism models assume that biological processes of oxygen production and consumption dominate DO fluxes (e.g. Peeters et al., 2016). However, a variety of non-biological processes can be important in redistributing DO, and these difficult to distinguish phenomena can appear in sensor observations (e.g. Brand et al., 2008). A data mining procedure represents an intermediate level of complexity to capture ‘suitable’ and ‘unsuitable’ observations and the dominant biological and non-biological features of the high-frequency DO sensor data.

4.6 Acknowledgments

This research was supported by the New Zealand Ministry of Business, Innovation and Employment (UOWX1503; Enhancing the health and resilience of New Zealand lakes). This work benefited from participation in GLEON. The North Temperate Lakes LTER provided data and funding support for PCH. I thank multiple anonymous experts for providing their insights into data quality, and Chris Solomon and Peter Stæhr for letting us use their data products. I also thank the two reviewers for their detailed and constructive comments and suggestions. KM thanks Adam Hartland for his support for the manuscript writing.

The source code can be accessed at: https://github.com/kohjim/DOClassifier

4.7 References

Aggarwal, C. C., A. Hinneburg, and D. A. Keim. 2001. On the surprising behavior of distance metrics in high dimensional space. Database Theory – ICDT 2001 **434**: 420–434. doi:10.1007/3-540-44503-X_27

Antenucci, J. P., K. M. Tan, H. S. Eikaas, and J. Imberger. 2013. The importance of transport processes and spatial gradients on in situ estimates of lake metabolism. Hydrobiologia **700**: 9–21. doi:10.1007/s10750-012-1212-z

Batt, R. D., and S. R. Carpenter. 2012. Free-water lake metabolism: Addressing noisy time series with a Kalman filter. Limnol. Oceanogr. Methods **10**: 20–30. doi:10.4319/lom.2012.10.20

Batt, R. D., and S. R. Carpenter. 2012. Free-water lake metabolism: addressing noisy time series with a Kalman filter. Limnol. Oceanogr. Methods 10: 20–30. doi:10.4319/lom.2012.10.20

Bradley, A. 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognit. **30**: 1145–1159. doi:10.1016/S0031-3203(96)00142-2

Brand, A., D. F. McGinnis, B. Wehrli, and A. Wüest. 2008. Intermittent oxygen flux from the interior into the bottom boundary of lakes as observed by eddy correlation. Limnol. Oceanogr. **53**: 1997–2006. doi:10.4319/lo.2008.53.5.1997

Campbell, J. L., L. E. Rustad, J. H. Porter, and others. 2013. Quantity is Nothing without Quality. Bioscience **63**: 574–585. doi:10.1525/bio.2013.63.7.10

Cannata, A., P. Montalto, M. Aliotta, C. Cassisi, A. Pulvirenti, E. Privitera, and D. Patanè. 2011. Clustering and classification of infrasonic events at Mount Etna using pattern recognition techniques. Geophys. J. Int. **185**: 253–264. doi:10.1111/j.1365-246X.2011.04951.x

Cengiz, T. 2011. Periodic structures of Great Lakes levels using wavelet analysis. J. Hydrol. Hydromechanics **59**: 24–35. doi:10.2478/v10098-011-0002-z

Chang, D. W., P. Hobson, M. Burch, and T. F. Lin. 2012. Measurement of cyanobacteria using in-vivo fluoroscopy - Effect of cyanobacterial species,

pigments, and colonies. Water Res. **46**: 5037–5048. doi: 10.1016/j.watres.2012.06.050

Cole, J. J., M. L. Pace, S. R. Carpenter, and J. F. Kitchell. 2000. Persistence of net heterotrophy in lakes during nutrient addition and food web manipulations. **45**: 1718–1730. doi: 10.4319/lo.2000.45.8.1718

Cox, K. C., S. G. Eick, and R. J. Brachman. 1997. Brief Application Description; Visual Data Mining: Recognizing Telephone Calling Fraud. Data Min. Knowl. Discov. 1: 225–231. doi:http://dx.doi.org/10.1023/A:1009740009307

Cox, T. J. S., T. Maris, K. Soetaert, J. C. Kromkamp, P. Meire, and F. Meysman. 2015. Estimating primary production from oxygen time series: A novel approach in the frequency domain. Limnol. Oceanogr. Methods **13**: 529–552. doi:10.1002/lom3.10046

Cremona, F., A. Laas, P. Nõges, and T. Nõges. 2014. High-frequency data within a modeling framework: On the benefit of assessing uncertainties of lake metabolism. Ecol. Modell. **294**: 27–35. doi:10.1016/j.ecolmodel.2014.09.013

Frank, E., and M. Hall. 2001. A simple approach to ordinal classification. Mach. Learn. ECML 2001 **2167**: 145–156. doi:10.1007/3-540-44795-4_13

Giling, D. P., P. A. Staehr, H. P. Grossart, and others. 2017. Delving deeper: Metabolic processes in the metalimnion of stratified lakes. Limnol. Oceanogr. **62**: 1288–1306. doi:10.1002/lno.10504

Hall, M., E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. 2009. The WEKA data mining software. ACM SIGKDD Explor. Newsl. **11**: 10. doi:10.1145/1656274.1656278

Hamilton, D. P., C. C. Carey, L. Arvola, and others. 2015. A global lake ecological observatory network (GLEON) for synthesising high-frequency sensor data for validation of deterministic ecological models. Inl. Waters **5**: 49–56. doi:10.5268/IW-5.1.566

Hanson, P. C., D. L. Bade, S. R. Carpenter, and T. K. Kratz. 2003. Lake metabolism: Relationships with dissolved organic carbon and phosphorus. Limnol. Oceanogr. **48**: 1112–1119. doi:10.4319/lo.2003.48.3.1112

Hanson, P. C., S. R. Carpenter, N. Kimura, C. Wu, S. P. Cornelius, and T. K. Kratz. 2008. Evaluation of metabolism models for free-water dissolved oxygen

methods in lakes. Limnol. Oceanogr. Methods **6**: 454–465. doi:10.4319/lom.2008.6.454

Horsburgh, J. S., S. L. Reeder, A. S. Jones, and J. Meline. 2015. Open source software for visualization and quality control of continuous hydrologic and water quality sensor data. Environ. Model. Softw. **70**: 32–44. doi:10.1016/j.envsoft.2015.04.002

Hutchinson, E. 1957. MLA Hutchinson, G. Evelyn. "A Treatise on." Limnology 1, John Wiley & Sons, Inc.

Kara, E. L., P. Hanson, D. Hamilton, and others. 2012. Time-scale dependence in numerical simulations: Assessment of physical, chemical, and biological predictions in a stratified lake at temporal scales of hours to months. Environ. Model. Softw. **35**: 104–121. doi:10.1016/j.envsoft.2012.02.014

Karpatne, A., G. Atluri, J. Faghmous, and others. 2016. Theory-guided data science: a new paradigm for scientific discovery. arXiv 1–14. doi:10.1109/TKDE.2017.2720168

Keogh, E., K. Chakrabarti, M. Pazzani, and S. Mehrotra. 2001. Dimensionality reduction for fast similarity search in large time series databases. Knowl. Inf. Syst. **3**: 263–286. doi:10.1007/PL00011669

Keogh, E., and S. Kasetty. 2002. On the need for time series data mining benchmarks. Proc. eighth ACM SIGKDD Int. Conf. Knowl. Discov. data Min. - KDD '02 102. doi:10.1145/775047.775062

Kohavi, R. 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection 2 methods for accuracy estimation. Proc. of IJCAI'95 1137–1145. doi:10.1067/mod.2000.109031

Lauster, G. H., P. C. Hanson, and T. K. Kratz. 2006. Gross primary production and respiration differences among littoral and pelagic habitats in northern Wisconsin lakes. Can. J. Fish. Aquat. Sci. **63**: 1130–1141. doi:10.1139/f06-018

Lin, J., E. Keogh, S. Lonardi, and B. Chiu. 2003. A symbolic representation of time series, with implications for streaming algorithms. SIGMOD Work. Res. Issues Data Min. Knowl. Discov. 2–11. doi:10.1145/882082.882086

Lin, J., E. Keogh, L. Wei, and S. Lonardi. 2007. Experiencing SAX: A novel symbolic representation of time series. Data Min. Knowl. Discov. **15**: 107–144. doi:10.1007/s10618-007-0064-z

Ma, Y., X. Meng, and S. Wang. 2016. Parallel similarity joins on massive high-dimensional data using MapReduce. Concurr. Comput. Pract. Exp. **28**: 166–183. doi:10.1002/cpe.3663

Matthews, B. W. 1975. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. BBA - Protein Struct. **405**: 442–451. doi:10.1016/0005-2795(75)90109-9

Niennattrakul, V., P. Ruengronghirunya, and C. A. Ratanamahatana. 2009. Exact Indexing for Massive Time Series Databases under Time Warping Distance.

Peeters, F., D. Atamanchuk, A. Tengberg, J. Encinas-Fernández, and H. Hofmann. 2016. Lake Metabolism: Comparison of lake metabolic rates estimated from a diel CO2-and the common diel O2- Technique. PLoS One **11**: 1–24. doi:10.1371/journal.pone.0168393

Rakthanmanon, T., E. J. Keogh, S. Lonardi, and S. Evans. 2011. Time series epenthesis: Clustering time series streams requires ignoring some data. Proc. - IEEE Int. Conf. Data Mining, ICDM 547–556. doi:10.1109/ICDM.2011.146

Ralanamahatana C.A., Lin J., Gunopulos D., Keogh E., Vlachos M., Das G. 2005. Mining Time Series Data. In: Maimon O., Rokach L. (eds) Data Mining and Knowledge Discovery Handbook. Springer, Boston, MA. doi: 10.1007/0-387-25465-X_51

Read, J. S., D. P. Hamilton, I. D. Jones, K. Muraoka, L. A. Winslow, R. Kroiss, C. H. Wu, and E. Gaiser. 2011. Derivation of lake mixing and stratification indices from high-resolution lake buoy data. Environ. Model. Softw. **26**: 1325–1336. doi:10.1016/j.envsoft.2011.05.006

Richardson, D. C., C. C. Carey, D. A. Bruesewitz, and K. C. Weathers. 2017. Intra- and inter-annual variability in metabolism in an oligotrophic lake. Aquat. Sci. **79**: 319–333. doi:10.1007/s00027-016-0499-7

Robson, B. J. 2014. State of the art in modelling of phosphorus in aquatic systems: Review, criticisms and commentary. Environ. Model. Softw. **61**: 339–359. doi:10.1016/j.envsoft.2014.01.012

Rose, K. C., L. A. Winslow, J. S. Read, E. K. Read, C. T. Solomon, R. Adrian, and P. C. Hanson. 2014. Improving the precision of lake ecosystem metabolism estimates by identifying predictors of model uncertainty. Limnol. Oceanogr. Methods **12**: 303–312. doi:10.4319/lom.2014.12.303

Ruan, G., P. C. Hanson, H. A. Dugan, and B. Plale. 2017. Mining lake time series using symbolic representation. Ecol. Inform. **39**: 10–22. doi:10.1016/j.ecoinf.2017.03.001

Sadro, S., G.W. Holtgrieve, C.T. Solomon, and G.R. Koch. 2014. Widespread variability in overnight patterns of ecosystem respiration linked to gradients in dissolved organic matter, residence time, and productivity in a global set of lakes. Limnology and Oceanography 59: 1666-1678.

Schapire, R. E., P. Stone, D. McAllester, M. L. Littman, and J. A. Csirik. 2002. Modeling auction price uncertainty using boosting-based conditional density estimation. Mach. Learn. Work. Then Conf. 546–553.

Solomon, C. T., D. a. Bruesewitz, D. C. Richardson, and others. 2013. Ecosystem respiration: Drivers of daily variability and background respiration in lakes around the globe. Limnol. Oceanogr. **58**: 849–866. doi:10.4319/lo.2013.58.3.0849

Staehr, P. a, D. Bade, M. C. Van de Bogert, G. R. Koch, C. Williamson, P. Hanson, J. J. Cole, and T. Kratz. 2010. Lake metabolism and the diel oxygen technique: State of the science. Limnol. Oceanogr. Methods **8**: 628–644. doi:10.4319/lom.2010.8.0628

Van de Bogert, M. C., S. R. Carpenter, J. J. Cole, and M. L. Pace. 2007. Assessing pelagic and benthic metabolism using free water measurements. Limnol. Oceanogr. Methods **5**: 145–155. doi:10.4319/lom.2007.5.145

Weathers, K., P. C. Hanson, P. Arzberger, and others. 2013. The Global Lake Ecological Observatory Network (GLEON): the evolution of grassroots network science. Limnol. Oceanogr. Bull. **22**: 71–73. doi: 10.1002/lob.201322371

Whitelaw, C., B. Hutchinson, G. Y. Chung, and G. Ellis. 2009. Using the Web for Language Independent Spellchecking and Autocorrection. Proc. 2009 Conf. Empir. Methods Nat. Lang. Process. 890–899. doi:10.3115/1699571.1699629

Winslow, L. A., J. A. Zwart, R. D. Batt, H. A. Dugan, R. I. Woolway, J. R. Corman, P. C. Hanson, and J. S. Read. 2016. LakeMetabolizer: an R package for estimating lake metabolism from free-water oxygen using diverse statistical models. Inl. Waters 6: 622–636. doi:10.1080/IW-6.4.883

Witten, I. H., E. Frank, M. A. Hall. and C. Pal 2016. Data mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann. San Francisco, CA.

Woolway, R. I., I. D. Jones, D. P. Hamilton, S. C. Maberly, K. Muraoka, J. S. Read, R. L. Smyth, and L. A. Winslow. 2015. Automated calculation of surface energy fluxes with high-frequency lake buoy data. Environ. Model. Softw. **70**: 191–198. doi:10.1016/j.envsoft.2015.04.013

Zimek, A., E. Schubert, and H.-P. Kriegel. 2012. A survey on unsupervised outlier detection in high-dimensional numerical data. Stat. Anal. Data Min. **5**: 363–387. doi:10.1002/sam.11161

R Packages

R Core Team. 2017. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna Austria. URL https://www.R-project.org/.

Simon Urbanek. 2017. rJava: Low-Level R to Java Interface. R package version 0.9-9. https://CRAN.R-project.org/package=rJava

Winston Chang, Joe Cheng, JJ Allaire, Yihui Xie and Jonathan McPherson. 2017. shiny: Web Application Framework for R. R package version 1.0.5. https://CRAN.R-project.org/package=shiny

Kurt Hornik. 2018. RWekajars: R/Weka Interface Jars. R package version 3.9.2-1. https://CRAN.R-project.org/package=RWekajars

Hornik, K., C. Buchta, and A. Zeileis. 2009. Open-source machine learning: R meets Weka. Comput. Stat. 24: 225–232. doi:10.1007/s00180-008-0119-7

4.8 Supplementary tables

Supplementary table 4-1: List of lakes used to create training dataset and their locations, elevations (Elv), maximum depths (Zmax), mean depths (Zmean), surface area and TP. Values were retrieved from Solomon et al. (2013), Staehr et al. (2010; Castle lake), and (North Sparkling Lake)

Lake Name	Latitude	Longitude	Elv (m asl)	Zmax (m)	Zmean (m)	Area (km^2)	TP ($\mu g\ L^{-1}$)
Acton	39.58	-84.75	263	8	4	2.53	114
Annie	27.21	-80.65	3.7	9	4	0.365	4.3
Castle*	56	12	0	9	3.5	0.223	156.7 (summer)
Crystal Bog	46.01	-89.61	503	2.5	2	0.005	27
Feeagh	53.95	-8.42	0	45	14	4	7.3
Gribsø	55.98	12.30	50	12	5	0.1	69
Hampensø	56.01	9.36	79	14	4	0.76	22.7
Mendota	43.11	-88.58	259	25	13	39.4	85
North Sparkling Bog	46.00	-89.71	497	4.3	N/A	0.46	31.8
Rotoiti	-37.92	176.27	279	125	31	34.6	30.3
Rotorua	-37.92	176.27	280	24	11	79.8	32.7
Sparkling	46.01	-89.70	497	20	11	0.64	10
Sunapee	43.38	-71.94	333	32	10	16.7	5.3
Taihu	31.17	120.15	3	3	2	2338	186
Trout	46.04	-89.67	495	36	15	16.1	13
Trout Bog	46.04	-89.69	495	7.9	6	0.011	29
Vedstedsø	55.20	9.36	25	12	5	0.09	19.5
Yuan Yang	24.58	121.40	1670	4.5	1.7	0.036	6.4

Supplementary table 4-2. Two to ten bin breakpoints (β) for a standard normal distribution, which splits data probability evenly (Adapted from Lin et al., 2003).

Number of alphabets	Breakpoints (β)								
	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9
2	0								
3	-0.43	0.43							
4	-0.67	0	0.67						
5	-0.84	-0.25	0.25	0.84					
6	-0.97	-0.43	0	0.43	0.97				
7	-1.07	-0.57	-0.18	0.18	0.57	1.07			
8	-1.15	-0.67	-0.32	0	0.32	0.67	1.15		
9	-1.22	-0.76	-0.43	-0.14	0.14	0.43	0.76	1.22	
10	-1.28	-0.84	-0.52	-0.25	0	0.25	0.52	0.84	1.28

Chapter four

Developing a mechanistic understanding of aquatic biodiversity using species richness constituents

5.1 Abstract

Conventional phytoplankton cell count monitoring data from four lakes (Lakes Annie, Feeagh, Esthwaite and Mendota) is used in this study to analyse the constituents of species richness. The constituents include number of species recruited, extinguished, and increasing or decreasing in abundance. Each constituent is expressed as a relative proportion of the total species richness. The number of species increasing and decreasing was negatively correlated in the four lakes, and the number of recruited species correlated with the ratio of increases in species to richness in the previous sample in three of the study lakes. The proportion of increasing and recruited species to the total number of species provided insights into environmental forcings that affected species richness. During thermal stratification, this proportion decreased, likely due to limited resources and intense competition. The techniques used in this study have application for the verification of population models and provide insights into the environmental drivers of species richness.

5.2 Introduction

Biodiversity in aquatic ecosystems is widely regarded as one of the most impacted by anthropogenic stressors of any of major ecosystem type (Dudgeon et al., 2006;

Williamson et al., 2009). Indicators of biodiversity have an important role in allowing the frequency and magnitude of a disturbance to be assessed and understanding the origin of coexistence. Conventional measures like species richness (i.e. how many unique species are in a sample) or evenness (i.e. a measure of equitability; Whittaker, 1972) have commonly been used to assess biodiversity. While it is well known that biodiversity is in decline globally, these measures can fail to demonstrate such a trend at a local scale due to the lack of mechanistic information within a single observation (Hillebrand et al., 2018).

Phytoplankton generally respond rapidly to environmental stimuli (Adrian et al., 2009) compared with longer-lived communities that change more slowly. Therefore, it is logistically difficult to capture the drivers of changes in phytoplankton community. To describe why phytoplankton species richness changes over time, diversity measures can be complimented with information such as the temporal dynamics of abundance changes (Hillebrand et al., 2018; Jones et al., 2018). Seasonal variability in different phytoplankton functional groups has been described theoretically in the Phytoplankton Ecology Group model (Sommer et al., 1993; Sommer et al., 2012). According to the model, in a simple case with a meromictic lake, fast growing phytoplankton species increase their population in spring due to increased light and temperature. After the spring bloom, the population collapses due to zooplankton grazing, followed by gradual increase because of decreased zooplankton abundance. During stratification, populations of phytoplankton typically are controlled by nutrient limitation, and increase their abundance in autumn due to mixing and increased resources, then gradually decrease due to light limitation and temperature decrease (Sommer et al., 2012).

When it comes to species level diversity, Hutchinson (1961) found unexpectedly large numbers of species can coexist in lake phytoplankton despite a small number

of limiting resources, in what has sometimes been termed a supersaturated state (Schippers et al., 2001; Roelke et al., 2008). Many theories have been proposed to explain this phenomenon, including species being confined to distinct niches spatially and temporally (Richerson et al., 1970), the occurrence of frequent environmental fluctuations (Grenny et al., 1973; Descamps-Julien and Gonzalez, 2005), selective grazing (Haberman et al., 2003) and non-hierarchal competition and chaos (Huisman and Weissing, 1999). More recent studies have used concepts of neutral theory, where differences in niches among species are essentially ignored (Rosindell et al., 2011), or the lumpy coexistence concept (e.g. Withrow et al., 2018) where similar functional species coexist within clusters, but competitive exclusion occurs across clusters. Total or functional group biomass can be explained using resource competition theory (e.g. Page et al., 2018), but prediction of species richness is not easy due to potentially similar, high-dimensional niches.

A well-documented facet of community structure is relative species abundance distribution, which assigns abundance ranks to multiple species (e.g. Whittaker 1972). Interestingly, a recent meta-analysis showed that species with low abundance are similarly rare (i.e., only few species have low populations in a sample) to species with high abundance (Alroy 2015). This implies that large numbers of species are found in groups at moderate relative abundance, and therefore the dynamics of species in these groups are critical to understand species richness, the potential of these species to sustain their populations, and species richness at equilibrium. Using the concept of abundance distribution, Hillebrand et al. (2018) proposed that studies of community dynamics could use of relative (proportional) abundance of species between two consecutive samples. However, a challenge of *in situ* analysis of phytoplankton population changes is the potential to bias the analysis by sampling interval. For example, a long sampling interval might

completely miss peaks in the abundance of species. A more robust approach is required to reduce the potential for sampling bias and to better understand community dynamics.

Studies have been carried out associating environmental factors, commonly disturbance, to species richness (e.g., Wilson, 1994). Disturbance can be defined as fluctuations in the environment caused by external forces whose magnitude and frequency exceed the capacity of the existing community to absorb them and remain largely unaffected (Reynolds et al., 1993). A disturbance halts, shifts or reverses successional processes, in a sense reversing the supersaturated status (Reynolds et al., 1993). A disturbance may exclude parts of an existing population and open up niches for new individuals or species (Kondoh, 2001). Quantifying disturbance has historically been problematic because the time scales of monitoring do not necessarily coincide with those of abrupt disturbance events (Klug et al., 2012). Recently, high-frequency observations of variables and proxies of physical (e.g. temperature and derived variables), biophysical (e.g. dissolved oxygen and derived ecosystem metabolism), and biological (chlorophyll fluorescence) variables have expanded rapidly (Hamilton et al., 2014). Relating such high-frequency sensor observations to phytoplankton community dynamics may help to understand the mechanisms leading to species coexistence and succession, even though phytoplankton phenology measurements are commonly made at lower resolution.

A major challenge in phytoplankton ecology is to quantify the scale of disturbance that is relevant to the community, in terms of its magnitude and variability in time and space (Levin, 1976; Reynolds, 1995). In lakes, for example, wind events may alter vertical migration of phytoplankton (Pesant et al., 2002) and heat waves affect phytoplankton succession and biomass (Gallina et al., 2011). Moreover, mixing processes are critical for the transport of phytoplankton and their ability to access

resources (Estrada and Berdalet, 1997). In lakes that thermally stratify, the extent of mixing is inversely related to the Schmidt stability (*SSt*) (Idso, 1973); the surface energy required to completely mix a lake (Read et al., 2011). Decreases in *SSt* are associated with increases in surface mixed layer depth and turbulence (Imberger and Patterson, 1989), which can lead to increases in nutrient concentrations and changes in phytoplankton species assemblages and diversity (Estrada and Berdalet, 1997).

Species richness is frequently used as a simple diversity measure in community ecology. However, changes in richness in consecutive samples can tell very little about mechanisms of diversity change, due to a lack of any means to examine population dynamics. On the other hand, trends in populations (e.g. rate of increase) cannot be trusted when there are long time intervals between successive samples, since the timescale of phytoplankton growth rates are very small (i.e. population sampling time intervals of 2-8 weeks most likely miss growth peaks and troughs). Therefore, a robust measure is required that bridges gaps between quantitative measures of population change and qualitative indicators of species diversity. In this chapter, species richness of paired samples was subdivided into constituents depending on whether the species in a population increased (i.e. due to dominance of newly recruited species) or decreased (i.e. due to dominance of species extinctions). This binary measure (increase or decrease) provides a better measure of population changes that conventional measures (species richness). The aim of this chapter was to test the constituent method of species richness between consecutive samples to examine if the changes in populations constituting communities reflects overall species richness changes. The chapter also explores relationships among the growth/loss constituents from paired phytoplankton samples to test whether there are distinct patterns in these indices.

5.3 Methods

5.3.1 Study sites

I chose four lakes where: (a) depth-integrated phytoplankton cell densities were recorded at species level by a single observer for at least two periods of seasonal thermal stratification, and (b) high-frequency observations of water column temperature were made over the same period. Lakes Annie (USA), Esthwaite (UK), Feeagh (Ireland; de Eyto et al., 2016) and Mendota (USA) were selected on the basis of these prerequisites (Table 5.1). In each lake, surface integrated water samples (Annie: 0-10 m; Esthwaite: 0-5 m; Feeagh: 0-15 m; Mendota: 0-8 m) were collected routinely at a mid-lake station for the phytoplankton analysis and high-frequency (<1 h) profiles of water temperature were also available from this station. For phytoplankton observations, Lake Annie and Esthwaite had consistent sampling intervals of 30 days and 14 days, while lakes Feeagh and Mendota had basic sampling intervals of 14 days with variation in sampling intervals mainly during the winter. To maintain the consistency of data, observations at greater than three-week intervals were removed from pair-wise analysis, except those from Lake Annie. Phytoplankton were identified to the lowest possible taxonomic resolution (i.e. usually species level), and abundance (cell density; N) was counted for each species.

*Table 5.1. Information on the four lakes used in the study including latitude (Lat), Longitude (Long), maximum depth (Z_{max}), mean depth (Z_{mean}), surface area (A), depth interval of thermistor measurements (dZ), summer phytoplankton sampling intervals (f_phyto), and buoy temperature observation intervals (f_temp). * Thermistor intervals varied in Loch Feeagh (2.5, 5, 8, 11, 14, 16, 18, 20, 22, 27, 32, 42 m). f_phyto were consistent throughout the study periods in Lakes Annie and Esthwaite, while Lakes Feeagh and Mendota increased sampling intervals during winter.*

	Lat (°).	Long (°).	Z_{max} (m)	Z_{mean} (m)	A (km^2)	dZ (m)	f_phyto (days)	f_temp (min)
Annie	27.2	281.4	21	9	0.4	1	~30	15
Esthwaite	54.2	2.6	15.5	6.4	1.1	1	~14	60
Feeagh	53.9	29.6	45	14	4	*	~14	2
Mendota	43.1	289.7	25	13	39.4	1	~14	1

5.3.2 Species richness and evenness

Species richness (S) was determined by counting the number of unique species in a sample. Abundance (cell density; N) of each species (i) in a sample at time (t) was used to calculate evenness (E), which describes the distribution of abundance:

$$E_{(t)} = \frac{\sum_{i=1}^{S_t} N_{(i,t)} \ln(N_{(i,t)})}{\ln(S_t)} \quad (1)$$

For a sample at time *t*, abundance was ranked for all species i. The change in N was categorised as increasing (Si) or decreasing (Sd), as well as newly recruited (Sr) or extinct (Se) when a species first appeared or disappeared from the sample (Table 5.2). The proportion of species in each of the four categories of a sample was calculated. Species richness based increasing species ratio was also determined as the number of species which were increasing in abundance (i.e., growth + recruitment) at time t:

$$ISR_{(t)} = (Sr_{(t)} + Si_{(t)}) / Sc_{(t)} \quad (2)$$

Values of the Pearson correlation coefficient were calculated amongst each of the four categories at time t and for species richness between the current ($S_{(t)}$) and previous time step ($S_{(t-1)}$). Furthermore, in order to examine the influence of sampling interval, Lake Esthwaite observations (bi-weekly samples) were re-analysed by randomly selecting the data at two, four, six and eight-week intervals in an iterative process.

Table 5.2. Summary of species components categorised by their population dynamics between two consecutive samples.

Notations	Equal to	Description
$Sr_{(t)}$		Number of newly recruited species from sample at time t
$Si_{(t)}$		Number of species with increasing cell counts in sample at time t
$Sd_{(t)}$		Number of species with decreasing cell counts in sample at time t
$Se_{(t)}$		Number of species extinct since time t-1
$S_{(t-1)}$	$Si_{(t)}+Sd_{(t)}+Se_{(t)}$	Number of species at time t-1
$S_{(t)}$	$Sr_{(t)}+Si_{(t)}+Sd_{(t)}$	Number of species at time t
$Sc_{(t)}$	$Sr_{(t)}+Si_{(t)}+Sd_{(t)}+Se_{(t)}$	Total number of species for paired sample t and t-1
$ISR_{(t)}$	$(Sr_{(t)}+Si_{(t)}) / Sc_{(t)}$	Proportion of species increasing in abundance between time t-1 and t

5.3.3 Transformation of phytoplankton cell densities

A per capita rate of increase (R) for each species i was calculated between consecutive samples:

$$R_{(i,t)} = \frac{N_{(i,t-1)} - N_{(i,t)}}{N_{(i,t-1)}} \cdot \frac{1}{dt} \qquad (3)$$

Species classified as recruited (Sr) or disappearing (Se) are not included in this analysis. Values of R were ranked for each sample and compared with species abundance ranks (t). The five fastest growing species were compared to their

abundance ranks at times t-1 (previous abundance rank) and t (current abundance rank).

Values of R (natural log-transformed) were ranked and regression relationships were developed between ln(R) and the ranking number. Times when there were less than 5 species growing were removed from the analysis so that the regression relationship was statistically meaningful. The slope of the regression lines (β) was then used as a growth rate distribution function across the community, being mindful that it may be influenced by sampling interval. A summary of the process is given in the diagram in Figure 5.1.

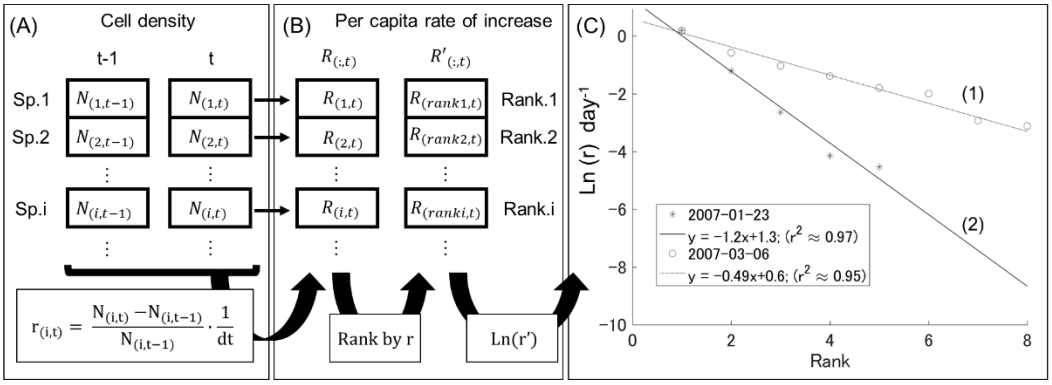


Figure 5.1. Schematic of the methodology used to transform phytoplankton cell density observations to apparent per capita rate of increase distribution in a paired sample. Phytoplankton cell densities (A) were used to calculate R, "apparent" per capita rate of population increase (change) for each species (B; R). Data were then sorted by their rate of increase (B; R'), natural log transformed, and a regression relationship was developed for rank-of-R (x-axis) and R (y-axis) for each sample occasion (C). From each regression equation, the slope of the line (β) was recorded, which corresponds to a growth rate distribution function across the community.

5.3.4 Lake stability calculation

Time series of water column temperature observations for each lake were converted to daily averaged Schmidt water column stability (SSt) using the Lake Analyzer program (available at lakeanalyzer.gleon.org; Read et al., 2011):

$$SSt = \frac{g}{A_{surf}} \int_0^{z_{max}} (z - z_V) \rho_z A_z \delta_z \quad (4)$$

where g is the acceleration due to gravity, A_{surf} is the lake surface area, A_z is the area at depth z, ρ_z is the density at depth z, z_{max} is the maximum water depth, and z_V is the depth to the centre of volume. The physical stability is intended to indicate seasonality, specifically thermal stratification of the lake.

5.4 Results

5.4.1 Time series analysis of species richness and constituents of richness

Time series of species richness (S) and four categories of constituents of richness (Sr, Se, Si, and Sd) are shown in Figure 5.2 over periods ranging from two to three years for the four lakes. The highest variability of richness (S) was found in Lake Esthwaite. The maximum S was > 40 in lakes Esthwaite and Feeagh and the minimum was 10 in Lake Mendota (Table 5.3). Values of S varied little seasonally in Lake Annie compared with the other lakes but a decline of S in this lake in autumn coincided with low values of Sr (Figure 5.2a). In Lake Annie the combined values of Sr and Se (proportion recruited + extinct) were generally greater than those of Si and Sd (proportion increasing + decreasing) compared with the other lakes. Lake Esthwaite (Figure 5.2b) exhibited a decline in S followed by peaks in spring and late summer. Observed declines in S in Esthwaite generally coincided with periods of high extinction and low recruitment. In Lake Feeagh there were peaks in S in summer of 2008 and 2009 (Figure 5.2c). In Lake Mendota S was generally higher in summer than in winter, although there was no distinct peak in 2008.

Table 5.3. Maximum and minimum species richness (S) and standard deviation (SD) in the four study lakes.

	Annie	Esthwaite	Feeagh	Mendota
max S	35	44	48	36
min S	18	12	19	10
SD	5.0	8.4	6.7	6.5

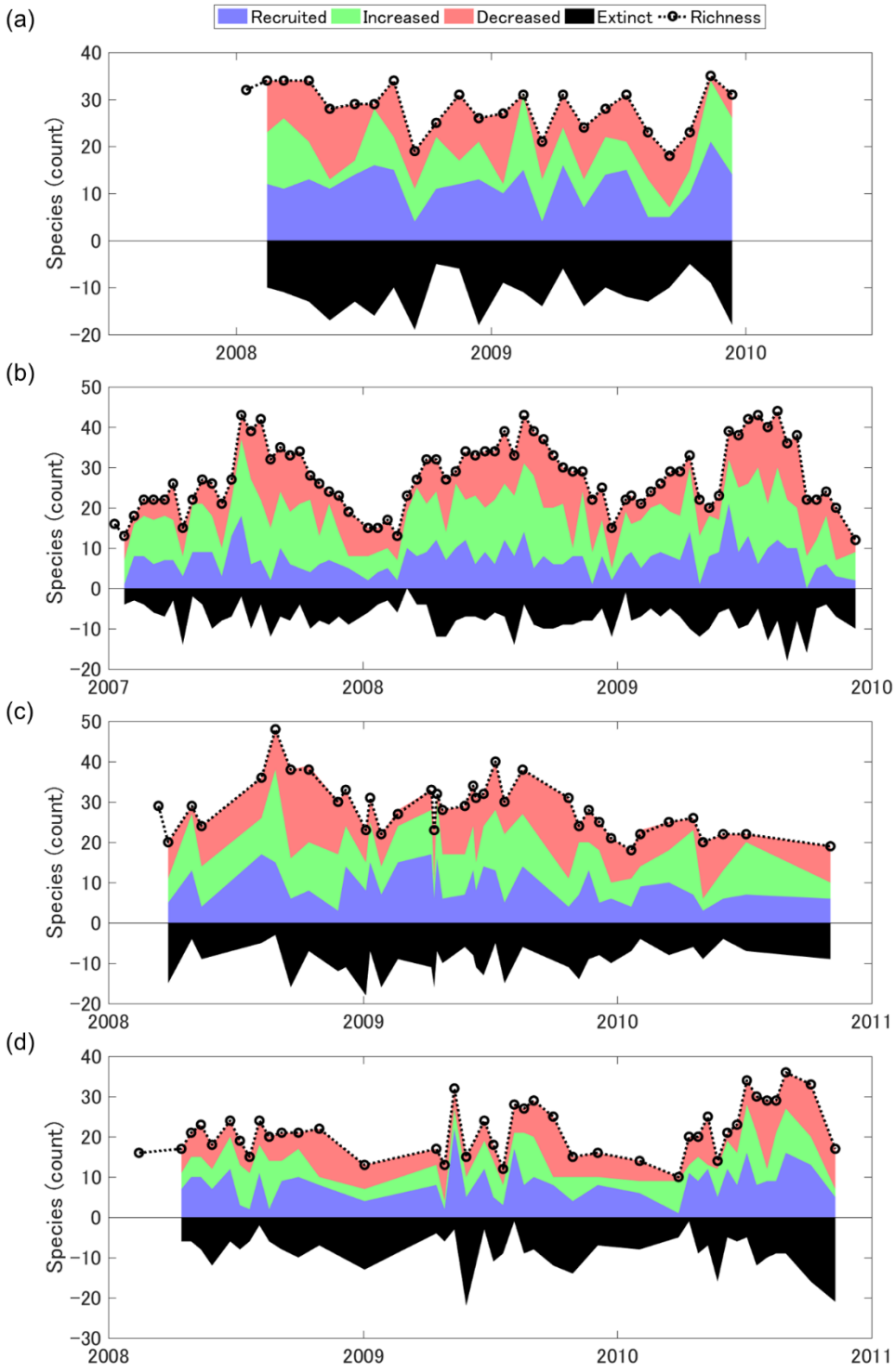


Figure 5.2. Species richness and constituents of richness: newly recruited, extinct, increasing cell concentration and decreasing cell concentration. (a) Lake Annie, (b) Lake Esthwaite, (c) Lake Feeagh and (d) Lake Mendota.

Values of the Pearson's correlation coefficient of the four constituents of species richness, and $S_{(t-1)}$, $S_{(t)}$ and change in S ($S_{(t)}$ - $S_{(t-1)}$) in the four lakes are shown in Table 5.4. Sample intervals of >3 weeks were removed from lakes Esthwaite, Feeagh and Mendota data for consistency.

$S_{(t)}$ was mostly positively correlated with Sr in the lakes, and with Si, though the correlation was weak in Lake Annie and not significant in Lake Mendota. Values of $S_{(t-1)}$ were positively correlated with Se except for Loch Feeagh where the correlation was not significant. As would be expected, the proportion of recruitment (Sr/Sc) was negatively correlated with the proportion of extinction (Se/Sc), and the proportion of increases in cell concentration (Si/Sc) was negatively correlated with the proportion of decreases in cell concentration (Sd/Sc) in all lakes. Similarly, the proportion of species that were increasing, ISR = (Si+Sr)/Sc, was positively correlated with Si/Sc and Sr/Sc, and negatively correlated with Sd/Sc and Se/Sc. In addition, ISR was negatively correlated with $S_{(t-1)}$ (not significant in Loch Feeagh) and positively correlated with $S_{(t)}$ (not significant in Lake Esthwaite). Change in S ($S_{(t)}$-$S_{(t-1)}$) was strongly positively correlated with Sr, Se/Sc and ISR, and strongly negatively correlated with Se and Se/Sc.

Sensitivity of the calculations to sampling interval was examined by artificially manipulating the observation intervals for one of the lakes, Esthwaite, with the results summarised in Table 5.5. This lake was chosen due to its continuous sampling at bi-weekly intervals. Significant correlative relationships were generally preserved for all four or eight-week intervals, as well as randomised sampling intervals, implying that the relationships related to constituents of species richness were not sensitive to sample interval duration for the intervals used in the four lakes.

Table 5.4. Correlation coefficient of previous ($S_{(t-1)}$) and current ($S_{(t)}$) species richness values, change in S ($S_{(t)}$ - $S_{(t-1)}$), constituents of richness (Sr, Si, Sd & Se) as well as proportion of these constituents (Sr/Sc, Si/Sc, Sd/Sc, Se/S & (Si+Sr)/Sc = ISR). For lakes Feeagh and Mendota, observation intervals greater than three weeks were removed from the analysis. Only values with $p < 0.05$ are shown.

Lakes		Variables	$S_{(t)}$		$S_{(t)}$ - $S_{(t-1)}$		Sr		Si		Sd		Se		Sr/Sc		Si/Sc		Sd/Sc		Se/Sc		ISR	
Annie	Esthwaite	$S_{(t-1)}$	-	0.76	-0.66	-0.37	-	-	-	0.56	-	0.75	0.82	0.63	-0.58	-0.47	-	-	-	0.33	0.66	-	-0.42	-0.34
Feeagh	Mendota		-	-	-	-0.60	-	-	0.48	0.56	-	-	-	0.75	-0.48	-0.60	-	-	-	-	-	0.52	-	-0.43
Annie	Esthwaite	$S_{(t)}$			0.65	0.32	0.82	0.58	0.44	0.72	-	0.55	-	-	0.64	-	-	-	-	-	-0.55	-0.41	0.57	-
Feeagh	Mendota				0.74	0.64	0.74	0.77	0.79	-	-	0.46	-0.51	-	0.51	0.53	0.62	-	-	-	-0.75	-0.62	0.74	0.56
Annie	Esthwaite	$S_{(t)}$ - $S_{(t-1)}$					0.79	0.79	-	-	-	-0.30	-0.76	-0.75	0.93	0.88	-	0.33	-	-0.40	-0.92	-0.84	0.75	0.78
Feeagh	Mendota						0.88	0.87	0.46	-	-	-	-0.83	-0.83	0.87	0.91	-	-	-	-	-0.91	-0.92	0.83	0.80
Annie	Esthwaite	Sr							-	0.30	-	-	-	-	0.91	0.82	-	-	-0.53	-0.42	-0.53	-0.50	0.74	0.58
Feeagh	Mendota								-	-	-	-	-0.47	-0.44	0.94	0.91	-	-	-0.57	-	-0.66	-0.67	0.81	0.70
Annie	Esthwaite	Si									-0.77	-	-	-	-	-	0.95	0.75	-0.84	-0.39	-	-0.40	0.75	0.51
Feeagh	Mendota										-	-	-	-	-	-	0.95	0.87	-0.63	-0.49	-0.54	-	0.77	-
Annie	Esthwaite	Sd											-	0.39	-	-0.47	-0.80	-0.50	0.94	0.83	-	-	-0.76	-0.63
Feeagh	Mendota												-	-	-0.45	-	-0.52	-0.43	0.89	0.81	-	-	-0.63	-
Annie	Esthwaite	Se													-0.52	-0.52	-	-0.44	-	-	0.91	0.82	-0.43	-0.62
Feeagh	Mendota														-0.54	-0.61	-	-	-	-	0.94	0.91	-0.60	-0.66
Annie	Esthwaite	Sr/Sc															-	-	-0.44	-0.59	-0.73	-0.60	0.79	0.76
Feeagh	Mendota																-	-0.45	-0.51	-	-0.65	-0.74	0.76	0.73
Annie	Esthwaite	Si/Sc																	-0.81	-0.67	-	-0.53	0.78	0.78
Feeagh	Mendota																		-0.66	-0.49	-0.52	-	0.78	-
Annie	Esthwaite	Sd/Sc																			-	-	-0.79	-0.82
Feeagh	Mendota																				-	-	-0.76	-0.42
Annie	Esthwaite	Se/Sc																					-0.62	-0.73
Feeagh	Mendota																						-0.76	-0.77

Table 5.5. Pearson's correlation coefficient of previous ($S_{(t-1)}$) and current ($S_{(t)}$) species richness values, change in S ($S_{(t)}$ - $S_{(t-1)}$), constituents of richness (Sr, Si, Sd & Se) as well as proportion of these constituents (Sr/Sc, Si/Sc, Sd/Sc, Se/S & (Si+Sr)/Sc = ISR). Values are only shown for $p < 0.05$. All results are from Lake Esthwaite, but with various intervals including two (actual sampling frequency), four and eight weeks. Pearson's correlation coefficient values are given for randomly selected intervals (from two, four, six and eight weeks), and averages of ten iteration results are shown (denoted as "Random").

Intervals (weeks)		Variables	$S_{(t)}$		$S_{(t)}$ - $S_{(t-1)}$		Sr		Si		Sd		Se		Sr/Sc		Si/Sc		Sd/Sc		Se/Sc		ISR	
Two	Four	$S_{(t-1)}$	0.76	0.61	-0.37	-0.50	-	-	0.56	0.39	0.75	0.80	0.63	0.67	-0.47	-0.58	-	-	0.33	0.49	-	0.34	-0.34	-0.50
Eight	Random		-	-	-0.68	-0.68	-0.40	-	-	-	0.79	0.64	0.80	0.54	-0.74	-0.51	-0.32	-	0.60	0.51	0.59	0.36	-0.69	-0.53
Two	Four	$S_{(t)}$			0.32	0.38	0.58	0.60	0.72	0.72	0.55	0.43	-	-	-	-	-	-	-	-	-0.41	-0.43	-	0.26
Eight	Random				0.56	0.56	0.72	0.65	0.67	0.67	-	-	-	-	0.42	-	0.26	-	-	-	-0.54	-0.51	0.43	0.38
Two	Four	$S_{(t)}$ - $S_{(t-1)}$					0.79	0.82	-	0.34	-0.30	-0.46	-0.75	-0.81	0.88	0.90	0.33	0.43	-0.40	-0.55	-0.84	-0.87	0.78	0.87
Eight	Random						0.88	0.70	0.37	0.45	-0.57	-	-0.85	-0.57	0.94	0.69	0.47	0.40	-0.62	-0.49	-0.91	-0.69	0.91	0.73
Two	Four	Sr							0.30	0.31	-	-0.25	-	-0.33	0.82	0.86	-	-	-0.42	-0.51	-0.50	-0.53	0.58	0.64
Eight	Random								0.34	-	-0.44	-	-0.50	-0.36	0.89	0.89	-	-	-0.61	-0.55	-0.67	-0.55	0.75	0.68
Two	Four	Si									-	-	-	-0.24	-	-	0.75	0.77	-0.39	-0.36	-0.40	-0.51	0.51	0.54
Eight	Random										-	-	-0.30	-	-	-	0.81	0.81	-0.40	-	-0.50	-0.50	0.53	0.53
Two	Four	Sd											0.39	0.52	-0.47	-0.57	-0.50	-0.48	0.83	0.87	-	0.27	-0.63	-0.69
Eight	Random												0.55	0.45	-0.69	-0.59	-0.54	-0.45	0.92	0.89	0.43	-	-0.75	-0.69
Two	Four	Se													-0.52	-0.60	-0.44	-0.61	-	0.38	0.82	0.89	-0.62	-0.79
Eight	Random														-0.74	-0.60	-0.62	-0.57	0.46	-	0.92	0.90	-0.83	-0.76
Two	Four	Sr/Sc															-	-	-0.59	-0.65	-0.60	-0.63	0.76	0.78
Eight	Random																0.36	-	-0.75	-0.67	-0.80	-0.66	0.90	0.82
Two	Four	Si/Sc																	-0.67	-0.58	-0.53	-0.64	0.78	0.75
Eight	Random																		-0.58	-0.52	-0.65	-0.62	0.72	0.70
Two	Four	Sd/Sc																			-	0.34	-0.82	-0.80
Eight	Random																				0.47	0.33	-0.82	-0.79
Two	Four	Se/Sc																					-0.73	-0.83
Eight	Random																						-0.89	-0.84

The correlation coefficients of Sr (i.e., number of species recruited) with Si, and Sr with the ratio $Si/S_{(t-1)}$ (proportion of species from the previous sample that had increased) are shown in Figure 5.3 for the four lakes. Significant positive correlations were found in three of the lakes but not in Lake Mendota ($p = 0.59$). Correlation coefficient values were low ($r \leq 0.59$) across the four lakes.

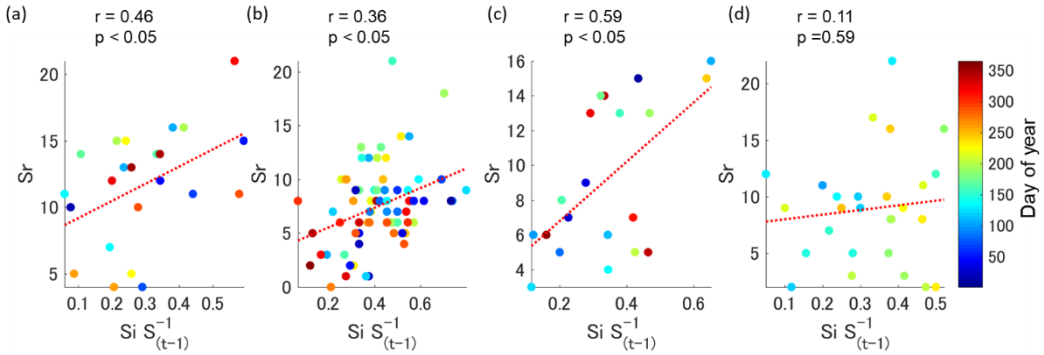


Figure 5.3. Scatter diagrams showing relationships between the proportion of species richness in the current sample, Si, and in the previous sample, S(t-1), in relation to Sr (number of newly recruited species since the last sample). Colours show day of year. Data are from Annie (A), Esthwaite (B). Feeagh (C), and Mendota (D).

5.4.2 Relating species dynamics to abundance ranks

Figure 5.4a-d visualises the normalized rate of population increase or decrease for each species in relation to their abundance ranks at time t for Lake Esthwaite. This lake was chosen to examine seasonality due to its constant and frequent sampling. Results for the other three lakes are shown in Appendix B. The results suggest that most of the species which were increasing had lower abundance ranks. Abundant species tended to decline from summer to late autumn and increase from late winter and early spring. Lake Feeagh (Appendix B) and Lake Mendota (Appendix B) showed similar seasonal patterns to Lake Esthwaite.

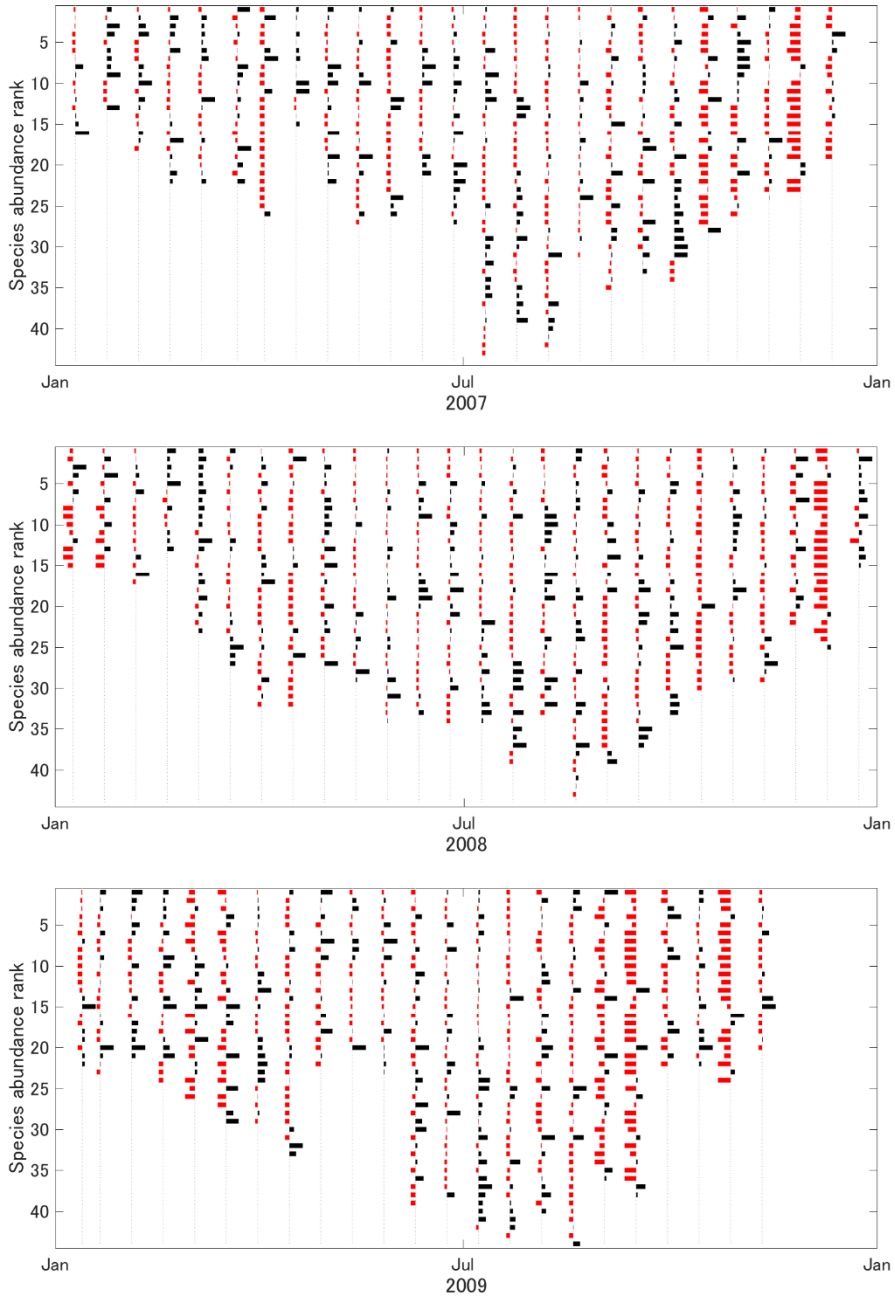


Figure 5.4. Time series of Lake Esthwaite cell concentration increases (black) and declines (red) between two consecutive samples in 2007, 2008 and 2009, in relation to their previous (t-1) abundance ranks (y axis). The bar heights were log normalized to the maximum rate of increase or decrease of the sample.

A summary of species' abundance ranks between consecutive samples (times t and t-1) is shown in Figure 5.5. Species with a high abundance ranking at time t-1 had a significantly higher chance of remaining abundant at time t. This probabililty appears incongruent to the findings presented in Figure 5.4, that fast growing, less abundant species had a higher rank, but many of the most abundant species remained abundant, despite population declines. The results suggest increases in

the less abundant species do not always result in exclusion of the dominant species. At the same time, species that were previously ranked below fifth most abundant had low probability of becoming the most abundant species in all lakes.

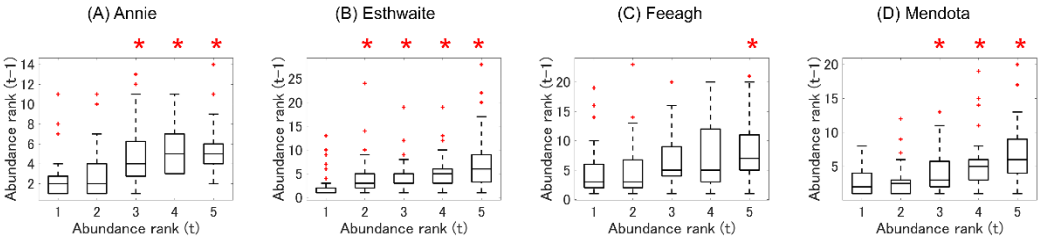


Figure 5.5. Comparisons of previous (t – 1) and current (t) five most highly ranked species abundance across the four lakes: Annie (A), Esthwaite (B), Feeagh (C) and Mendota (D). Abundance was derived from biweekly phytoplankton samples. Box illustrates the 25th and 75th percentile boundaries, and the in-box line is the median value. Whiskers are the 99% probability boundaries, whereas red markers in the figures are outliers. Red asterisks above figures differences (one-way ANOVA) between the sample associated with the box plot underneath and the most abundant sample of the same plot.

Figure 5.6 shows pair-wise comparisons between species abundance ranks and ranking of the per capita rate of increase for the five species with the highest rate of per capita increase. As expected from Figure 5.4, the fast-growing species most frequently originated from species of low to medium abundance in the previous sample (row I, Figure 4-6), but they had shifted to higher in the abundance ranks in the current sample (row II).

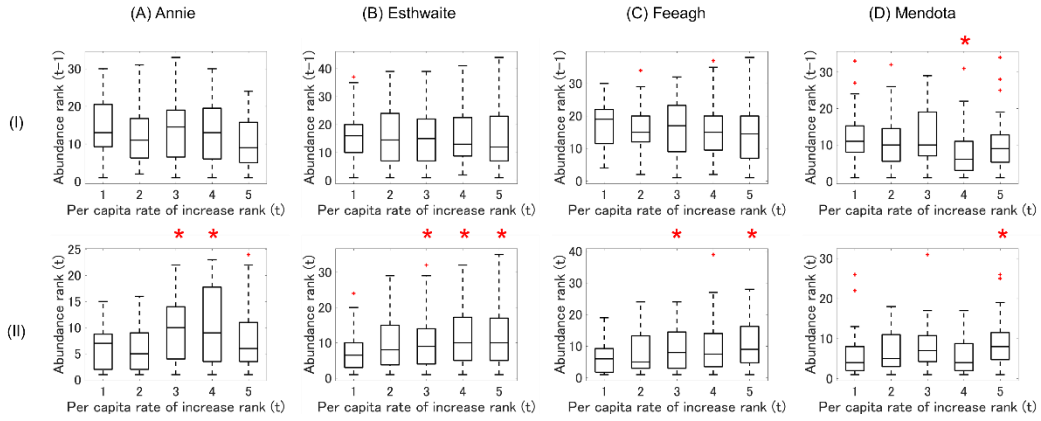


Figure 5.6. Comparisons of previous (t – 1) (I) *and current (t)* (II) *abundance rank against per capita rate of increase of rank for the five species with the highest per capita rate of increase in lakes Annie (A), Esthwaite (B), Feeagh (C) and Mendota (D). Box illustrates the 25th and 75th percentile boundaries, in-box line is the median value, whiskers are statistically determined 99% boundaries, and red markers in the figures are outliers. Red asterisks above figures indicate significant differences (one-way ANOVA) between the sample associated with the boxplot underneath and the highest ranked sample of the same plot.*

5.4.3 Seasonal behaviour of species richness, evenness and proportion increasing

Time series of species richness (S), cell density evenness (E), and the proportion of species that was increasing (ISR) were compared in Figure 5.7a-d for the four lakes. For all lakes, the shape of S was similar to ISR but with short time lags, notably in spring. In Lake Annie, the shape and magnitude of evenness (E) did not correspond to either S or ISR (Figure 1-7a). In Lake Esthwaite, ISR had the same number of peaks as S until early summer (Figure 5.7b) but peaks in ISR tended to occur before peaks of S. Large decreases in E occurred between late winter and early spring in Lake Esthwaite. Loch Feeagh also exhibited similar patterns between S and ISR until late summer (Figure 1-7c). Patterns of E in this lake also did not match those of S or ISR. Lake Mendota had peaks in S in spring and summer of 2009 and 2010, and these typically followed ISR peaks.

Table 5.6 shows the linear correlation coefficient for E, S, ISR and SSt, as well as change in S ($S_{(t)} - S_{(t-1)}$), all of which are smoothed monthly for the four lakes. SSt was significantly positively correlated with S and ISR in all of the lakes except Annie. ISR was reasonably well correlated with S (r = 0.50) in Lake Annie, reflecting strong influence of Sr and Se on ISR in this lake. For other three lakes, ISR was not very well correlated with S, suggesting that internal population dynamics are not immediately reflected in S. ISR was strongly correlated with $S_{(t)} - S_{(t-1)}$ in all lakes.

Table 5.6 Pearson's correlation coefficient for evenness (E), richness (S), proportion of species that were increasing (ISR), Schmidt Stability (SSt) and change in S ($S_{(t)} - S_{(t-1)}$). All values are smoothed monthly time series used to analyse frequency. Values shown are significant ($p < 0.05$).

		Variable	**S**		**ISR**		**SSt**		**S(t) – S(t-1)**	
Annie	Esthwaite	E	0.33	-	-	-0.25	-0.29	-	0.08	-0.21
Feeagh	Mendota		0.18	0.26	-	-0.25	-0.23	-0.34	-0.14	-0.12
Annie	Esthwaite	S			0.50	0.16	-0.41	0.67	0.52	0.25
Feeagh	Mendota				0.09	-	0.48	0.34	0.28	0.17
Annie	Esthwaite	ISR					-0.41	0.16	0.79	0.86
Feeagh	Mendota						0.30	0.52	0.79	0.82
Annie	Esthwaite	SSt							-0.25	0.33
Feeagh	Mendota								0.39	0.32

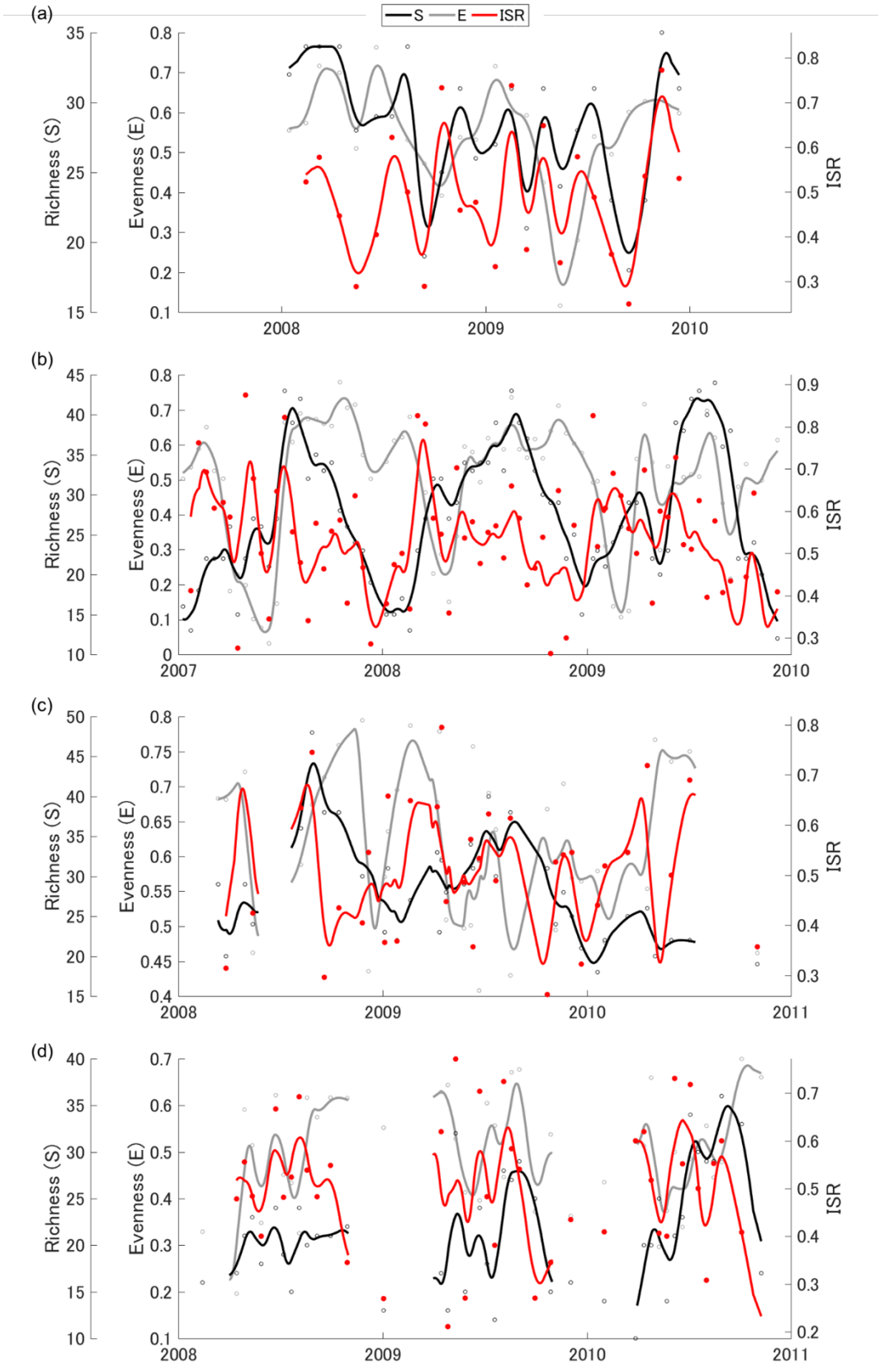


Figure 5.7. Species richness (S; black), evenness (E; grey), and proportion that were increasing (ISR) for (a) Lake Annie, (b) Lake Esthwaite, (c) Lake Feeagh and (d) Lake Mendota. Circles are raw data points and lines are smoothed data (monthly linear interpolation).

Time series of the proportion of species that were increasing (ISR) are plotted against Schmidt stability (*SSt*) in Figure 5.8 a-d. ISR values generally started to increase well before the onset of seasonal stratification in the lakes. Lake Annie showed strong seasonal stratification and *SSt* therefore had a distinct annual peak. ISR decreased in early spring of 2008 and 2009, as *SSt* increased, but increased until mid summer when *SSt* peaked. ISR values then started to decrease towards end of the stratification period.

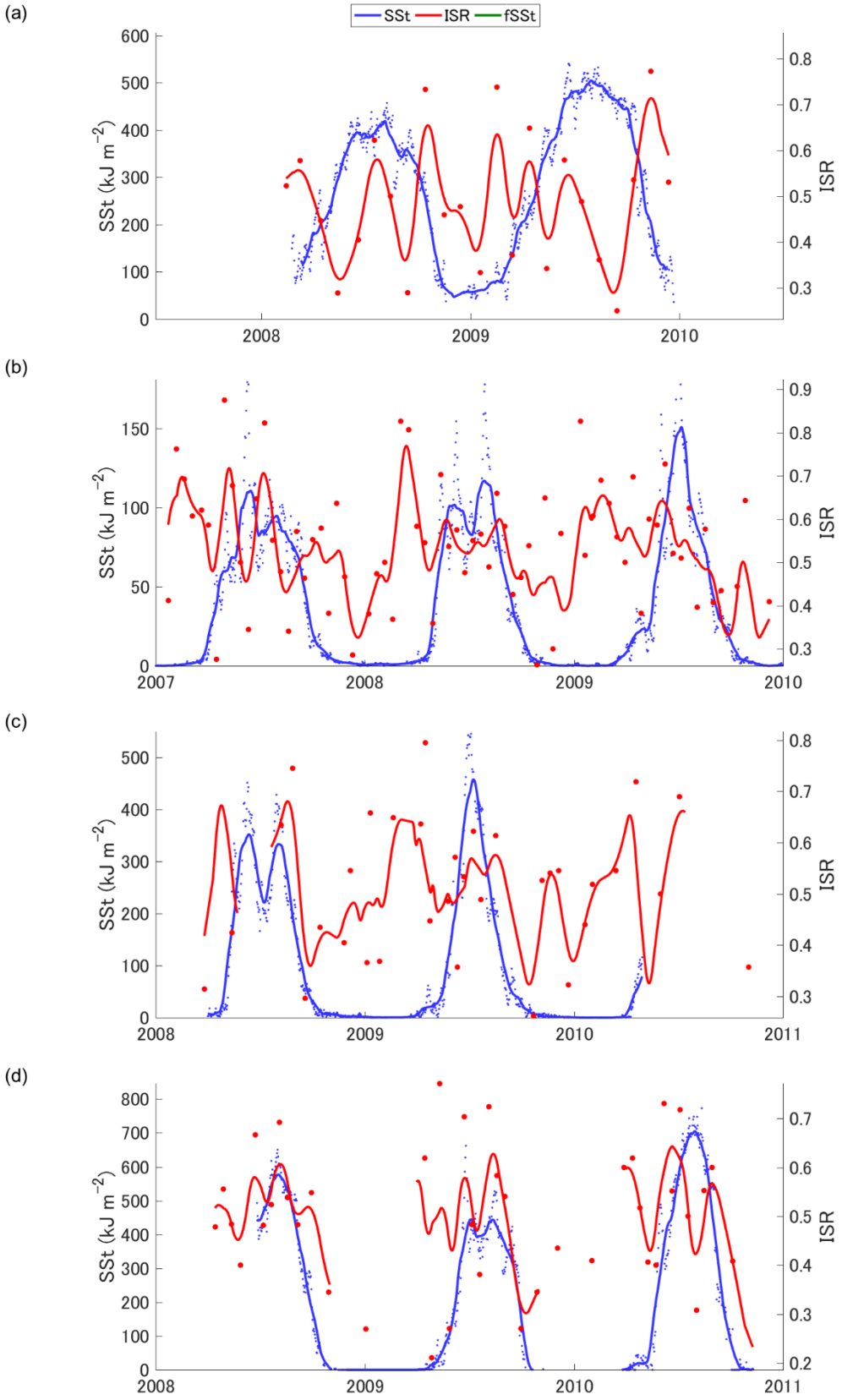


Figure 5.8. Species proportions that were increasing (ISR) and Schmidt stability (SSt) for (a) Lake Annie, (b) Lake Esthwaite, (c) Lake Feeagh and (d) Lake Mendota. Dots are raw values and lines are smoothed data (monthly linear interpolation).

Figure 5.9 summarises the frequency domain patterns (power spectrum density; PSD) of all indices, including Schmidt stability (*SSt*), species richness (S), cell density evenness (E), and proportion of species that was increasing (ISR). Annual patterns of *SSt* occurred in all lakes, and bi-annual patterns were observed in lakes Esthwaite, Feeagh and Mendota. Annual pattern of S was strong in lakes Esthwaite and Mendota, while more frequent cycles (bi-annual to tri-annual) were found in all lakes. ISR and E frequency patterns were reasonably similar to S in Lakes Esthwaite, Feeagh and Mendota.

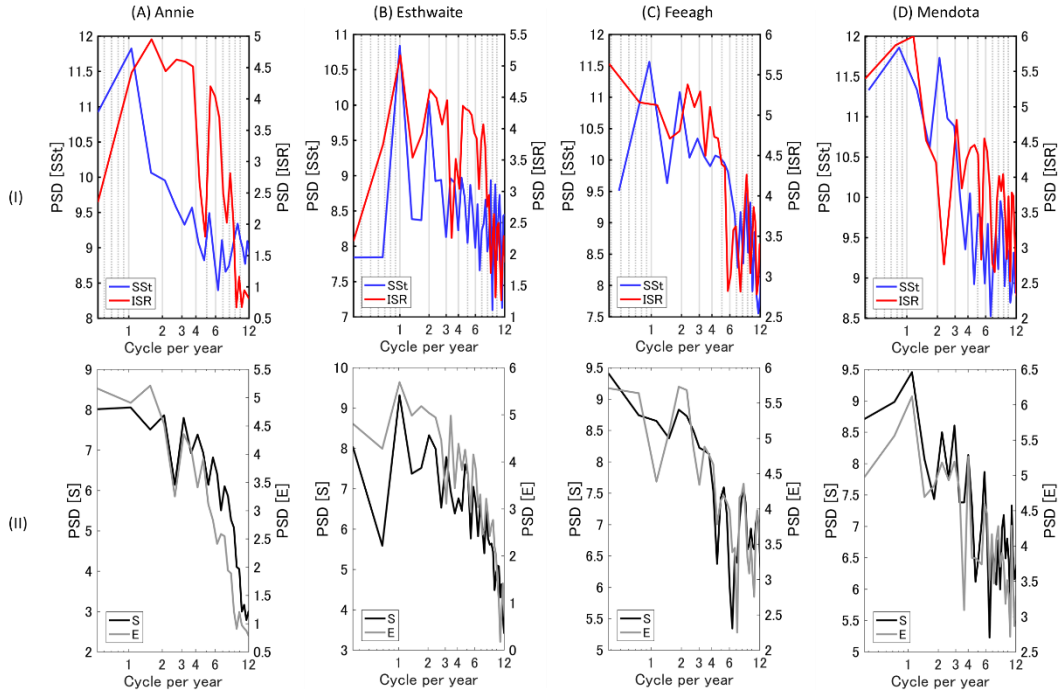


Figure 5.9. Power density spectra of (I) Schmidt stability (SSt) and proportion increasing (ISR) and (II) species richness (S) and evenness (E) for lakes (a) Annie, (b) Esthwaite, (c) Feeagh and (d) Mendota.

5.4.4 Growth rate distribution function across the community (β)

A value of β (the growth rate distribution function) was calculated from the slope of regressions that related the log transformed growth rate to the abundance rank (see Figure 5.1). The average regression models' goodness of fit is summarised in Table 5.7 for the four lakes. The overall goodness of fit performance (r^2) was high, suggesting that where distributions were skewed to just a few fast-growing species, the number of species increasing in abundance was also low.

Table 5.7. Assessment summaries of linear relationships between per capita rate of increase ranks vs per capita rate of increase in each observations in Lakes Annie, Esthwaite, Feeagh and Mendota. Each lakes' summary values (goodness of fit) were made by averaging the linear fit lines' goodness of fit in each observations (N = sample numbers used). Number of samples that which did not perform significant linear relationships are also indicated.

	Annie	Esthwaite	Feeagh	Mendota
Mean goodness of fit (r^2)	0.94	0.96	0.96	0.95
Number of samples with no significance (p >0.05)	1	0	0	1
Sample numbers used	19	67	33	29

The values of β is plotted against number of species increasing in abundance (Si) in Fig. 4.10 for each sample day for the four lakes. The results presented in Figure 5.10 suggest that if the number of species that are increasing is known, then it is possible to make predictions about the rate of growth of these species (Pearson coefficients of determination values of 0.3 to 0.7).

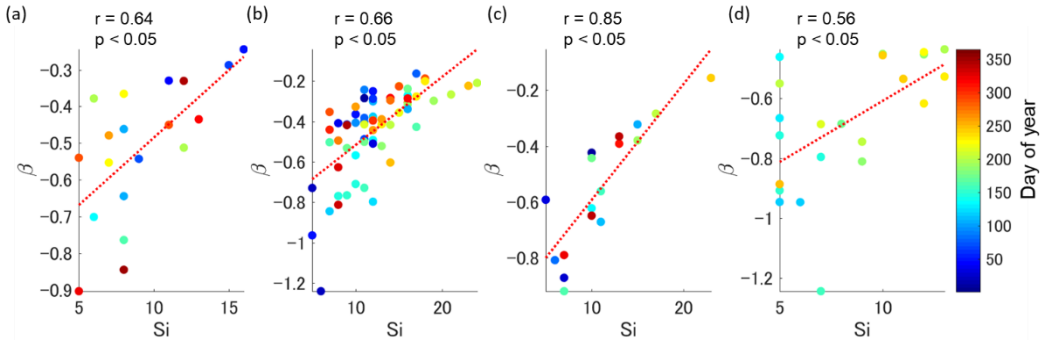


Figure 5.10. Scatter diagrams showing relationships between β (the growth rate distribution function for those species that were growing) and Si (number of species with increasing cell concentrations) for lakes (a) Annie, (b) Esthwaite, (c) Feeagh and (d) Mendota. Colours represent day of year.

5.5 Discussion

5.5.1 Constituents of richness – a hybrid analysis of diversity changes and population changes

Species richness is a widely accepted measure of diversity because of its simplicity, and many descriptive models have been developed based on this metric (e.g., the Intermediate Disturbance Hypothesis; Connell, 1978). However, due to continuous species turnover (Rosindell et al., 2011) and relative abundance reordering (Jones et al., 2018), changes measured by species richness are an order-of-magnitude smaller than species assemblage turnover rates (Hillebrand et al., 2018). This difference makes it difficult to construct a mechanistic description of community dynamics based on richness alone. Hence, a method is needed to describe trends in phytoplankton communities using both qualitative trends (richness changes) and quantitative trends (population and/or compositional change). A suitable hybrid approach to modelling species assemblage dynamics includes a mechanistic description. A focus on variability of the constituents of species dynamics (e.g. Hillebrand et al. 2018) provides the basis of a mechanistic approach and has been demonstrated in this study to be useful to better understand species dynamics.

One of the challenges for trend analysis based on conventional field observations is the relatively long time interval between samplings compared to the time scales of phytoplankton turnover (Reynolds, 1995). For example, per capita rate of increase may miss rapid oscillations and the peak rate of increase population when used for in situ observations (Huston, 1979). I limited species population dynamics expressions to binary terms (increase or decrease in abundance), instead of numerically quantifying population differences between the two samples. This approach was used to reduce the chance of misinterpreting the data with longer intervals. Binary expressions did not describe each species' precise growth dynamics, but provided simple information on individual species' changes between two consecutive samples. The time series results of richness constituents (Figure 5.2) showed evidence that positive and negative growth is not limited to individual species but can be a community level trend also. For example, there were times when a much greater number of species increased than decreased in a community, suggesting that the environment favoured a larger proportion of species growing without negatively interacting with each other.

5.5.2 Four constituents of population dynamics

Four constituents were used in this study to describe population dynamics. Increases or decreases in the population of a species implies that the environment associated with the time period between the two samples was either favourable or not favourable, respectively. Species recruitment (Sr) is unique compared to the other constituents, due to its uncertain origins and unpredictable nature. Recruitment is a key constituent in species richness models and my results confirm that it was a strong determinant of richness in all four lakes in this study.

Recruitment occurs from processes including resuspension, vertical or horizontal migration (Trimbee et al., 1984), germination from dormant forms (Wood et al., 2009), and immigration from outside the system (Wood et al., 2014). Observations at monthly or lesser frequency have a large chance of missing species that are recruited but do not establish viable populations, and are therefore not included as a number of newly recruited species (Sr). The factors driving Sr are therefore (1) how many species are actually recruited, and (2) what proportion of species establishes or sustains large enough populations to be sampled. The latter proportion may be examined by testing the relationship between Si/(Si + Sd + Se) and Sr, assuming that newly recruited species outnumber existing species. There was a weak but significant positive correlation between Si/(Si + Sd + Se) and Sr in three of the study lakes, suggesting that Sr was related to the demographics of the previous community. Therefore, recruitment can be partially explained from the existing community's trajectory of population changes.

5.5.3 Proportion of constituents and ISR

If species richness S(t-1) from a previous phytoplankton sample is large, naturally, a greater number of species may potentially increase (Si), decrease (Sd), or become extinct (Se) in the current sample. To normalize the values, proportional measures were proposed in this study, by dividing by the four constituents, i.e., Sc = Si + Sr + Sd + Se.

Relatively strong negative correlations between Sr/Sc and Se/Sc were found in all lakes, implying that extinction and recruitment tend to be inversely related. Similarly $Sr/S_{(t-1)}$ was negatively correlated with $Se/S_{(t)}$ in all lakes (Appendix, Table 1). Strong negative correlations were also found between Si/Sc and Sd/Sc (and also between $Si/S_{(t-1)}$ and $Sd/S_{(t-1)}$; Appendix, Table 1) in all lakes. There was

no relationship between Sd and Se, indicating that high proportions of decreasing population do not necessarily lead to an rapidly increased proportion of extinction.

5.5.4 Seasonal behaviour of ISR

ISR [= (Si + Sr) / Sc] was defined as the proportion of species that was increasing. This value is intended to predict whether a community will increase or decrease towards its equilibrium species richness. Species richness of lake phytoplankton fluctuates throughout a year. It is theorised that physical and biological interactions play major roles to supress and limit individual species numbers and biomass increases. One of the reasons that species can increase their population may be because of vacancies in niches, which alter carrying capacity. Where there are increases in carrying capacity, competitive exclusion may not be strong and species may increase their population. This situation can be theoretically seen when there is increased temperature and light availability in the spring, or when nutrients are released during a phase of cell lysis (Brussaard, 2004). Therefore, it is useful to consider the way in which the timing of environmental changes influences the carrying capacity.

The Phytoplankton Ecology Group (PEG) model (Somer et al., 1986; Somer et al., 2012) is a dominant paradigm used to explain seasonal dynamics of phytoplankton biomass in response to environmental changes, although it does not directly explain species diversity changes. In the PEG model, spring phytoplankton blooms are considered to be triggered by increases in light and to some extent temperature, and end with increased grazing pressure. Blooms increase productivity in the system, opening up more “spaces” to allow for an increase in the species diversity. The results followed the typical “spring pattern” of the PEG model that is associated with the presence of fast-growing species due to increased temperature and light. For Lake Esthwaite, for example, biomass increases in spring coincided with an increase in ISR and a decline in evenness (Figure 5.6). This condition likely allowed a greater proportion of species to proliferate rather than be excluded due to abundant resources. Increases in evenness were observed in late spring in this lake, coinciding with a decrease in ISR and S, which may have been associated with grazing pressure. During spring, species with higher abundance ranks continued to increase (Figure 5.4), likely leading to increases in primary productivity, until grazing pressure intensified.

After the initiation of stratification, species with higher abundance ranks decreased (Figure 5.4). The PEG model proposes that the main environmental pressure on phytoplankton biomass under stratified conditions shifts from grazing to nutrient limitation at this time. Unlike grazing, reduced access to nutrients may limit total ecosystem productivity, sometimes with an associated decline in total biomass. The results in Lakes Esthwaite and Feeagh (Figure 5.7) showed that ISR decreases dramatically during stratification even though richness remains higher than pre-stratification values. It is likely that competitive exclusion occurred during this period, allowing multiple species to oscillate in population size. Coexistence of phytoplankton under resource competition has been explained theoretically by biological chaos and continuous oscillation of biomass (Huisman and Weissing, 1999; Benincà et al., 2015). A recent study has observed that continuous environmental variation allows supersaturated coexistence using a multiple resource limitation model (Sarker et al., 2018). The observations of continual reductions in biomass of dominant species and increases in less abundant species in summer (Figure 5.4) fits with supersaturated coexistence.

The results indicate seasonal breakdown of stratification coincides with declines in ISR in lakes Esthwaite, Feeagh and Mendota. Extinction leading to a decline in richness tended to occur gradually rather than abruptly in lakes Esthwaite and Feeagh. This result suggests that physical stress due to destratification does not necessarily trigger immediate species extinctions, probably because resources increase and act as a buffer. Unlike the other study lakes, Lake Annie exhibited a seasonal low in species richness towards the end of stratification. In this lake, autumn mixing events were associated with increased ISR, followed by large increases in species richness. Resource limitation due to more prolonged vertical stratification in this lake may have resulted in intense competition during late summer.

5.5.5 Self-organization of growth rates

One of the findings of this study was that species tend to self-organize into linear distributions of per capita rates of population change (R). The slope of the line relating R to the rank of R indicated how many species were increasing in the community (Si; Figure 5.10). No two species showed identical responses (in terms of population change). This study only examined Si, the number species that continued to exist between samples, as it was not possible to represent Sr and Se in

terms of growth rate. However, it is possible that the distribution of rate of decline of species (i.e., plotting the negative R values against the rank of R) to follow distribution properties similar to rate of increase distribution, and could even be treated on the same axis.

Multiple limiting resources (e.g., light, nutrients and temperature) in competition models can be used to understand fluctuations in total biomass or functional group biomass (Page et al., 2018). These models are capable of capturing seasonal variations in community carrying capacities in lakes. Niche apportionment (Spatharis et al., 2009) could be developed into species level competition models similar to space patch occupancy competition models (e.g. Tilman 1994; Kondoh 2001) for evaluating carrying capacity. Species level space competition models may provide a basis to examine coexistence.

5.5.6 Future recommendations

Applied ecological modelling and forecasting rely on process-based approaches to produce projections based on environmental data inputs to a model. Phytoplankton community ecology often employs theoretical models, and adaptation of such theories to process based models tends to be limited to functional group level. Process based "descriptive" models assume resource competition and growth limitation based on carrying capacity (including light-shading limitation). Individual species are required to have tailored sets of parameters which function as "competition coefficients" between each pair of species in a given environment. The competition coefficients need to be prescribed under conditions of environmental change. Competition can be based on the current rank of the species abundance. This species independent neutral approach should be explored in more detail in phytoplankton ecological modelling. Robust and species independent metrics developed in this chapter should be of great help to assess phytoplankton model performance.

5.5.7 Limitations of the study

Species richness is a fundamental and common metric of community ecology studies but can contain observation errors related to detection limits. For example, a single cell will be counted equally towards richness as a much larger population (Hillebrand et al., 2018). Consideration of sample replication and rarefaction curve techniques (e.g. Gotelli et al., 2001) could allow for more accurate estimates of species richness based on using multiple samples.

The study was not intended to directly predict species diversity or why and what caused changes in populations. ISR shows directional changes in diversity. Use of additional information about resource availability could be utilized to further investigate the ISR values, in particular, whether the environment is favourable for the number of species to increase or decrease.

The study focused on methods with neutral species characteristics. However, as lumpy coexistence theory predicts, neutrality might be true only for species with similar characteristics (e.g., functional groups). Therefore, it may be useful to divide species into different functional characteristics before conducting the analysis (Burson et al., 2018). Because functional group modelling has been successfully used to predict the abundance of each group (e.g. Page et al., 2018), true assemblage dynamics based on purely neutral population dynamics may be shown.

5.5.8 Conclusions

The chapter explored conventional phytoplankton observation data from four lakes from different regions of the globe, Annie, Esthwaite, Feeagh and Mendota. The results provide a new perspective on changes in diversity and population. Use of binary terms (population increase or decrease) should provide less biased observations of changes in population than conventional non-binary population change assessments. The number of species that were increasing or decreasing was inversely correlated while the number of species that were decreasing was not correlated with the number of species that had become extinct. The proportion of increasing and recruited species to the total species number, ISR, showed seasonal features that were similar to diversity changes. ISR increased during spring time, likely reflecting the increase of productivity at this time. Decreases in ISR coincided with destratification in three of the lakes, indicating that mixing negatively affected the community. However, in Lake Annie, increases in ISR coincided with destratification and were followed by increases in richness, suggesting that mixing benefited the community. The analysis of rate of population increase to rank of rate of population increase showed growth rates that were linearly distributed from fastest growing to slowest growing species. The techniques developed in this study can be used for verification of population models.

5.6 Acknowledgments

I thank the data providers (Lake Annie: Evelyn Gaiser, Florida International University; Feeagh: Eleanor Jennings, Elizabeth Ryder, Elvira de Eyto, Karin Sparber and colleagues, Dundalk Institute of Technology and Marine Institute; Lake Esthwaite: Stephen Maberly and Centre for Ecology & Hydrology, University of Lancaster; Lake Mendota: Paul Hanson, the University of Wisconsin-Madison). The work was based on the dataset created through Global Lake Ecological Observatory Network (GLEON) The Theory Group (TTG) working group. Part of this chapter's work will be used to publication of the TTG working group, where co-authors Bas Ibelings, Cayelan Carey, David Hamilton Evelyn Gaiser, Elizabeth Ryder, Karin Sparber, Lisette Senerpont Domis, Paul Hanson, Sam Fey, and Stephen Maberly will be included in the publication.

5.7 References

Adrian, R., C. M. O'Reilly, H. Zagarese, and others. 2009. Lakes as sentinels of climate change. Limnol. Oceanogr. 54: 2283–2297. doi:10.4319/lo.2009.54.6_part_2.2283

Alroy, J. 2015. The shape of terrestrial abundance distributions. Sci. Adv. 1: e1500082. doi:10.1126/sciadv.1500082

Benincà, E., B. Ballantine, S. P. Ellner, and J. Huisman. 2015. Species fluctuations sustained by a cyclic succession at the edge of chaos. Proc. Natl. Acad. Sci. 112: 6389–6394. doi:10.1073/pnas.1421968112

Brussaard, C. P. D. 2004. Viral control of phytoplankton populations-a review. J. Eukaryot. Microbiol. 51: 125–138. doi:10.1111/j.1550-7408.2004.tb00537.x

Burson, A., M. Stomp, E. Greenwell, J. Grosse, and J. Huisman. 2018. Competition for nutrients and light: testing advances in resource competition with a natural phytoplankton community. Ecology 99: 1108–1118. doi:10.1002/ecy.2187

Connell, J. H., and R. O. Slatyer. 1977. Mechanisms of Succession in Natural Communities and Their Role in Community Stability and Organization. Am. Nat. 111: 1119–1144. doi:10.1086/283241

de Eyto, E., E. Jennings, E. Ryder, K. Sparber, M. Dillane, C. Dalton, and R. Poole. 2016. Response of a humic lake ecosystem to an extreme precipitation event: physical, chemical, and biological implications. Inl. Waters 6: 483–498. doi:10.1080/IW-6.4.875

Descamps-Julien, B., and A. Gonzalez. 2005. Stable coexistence in a fluctuating environment: an experimental demonstration. Ecology 86: 2815–2824. doi:10.1890/04-1700

Dudgeon, D., A. H. Arthington, M. O. Gessner, and others. 2006. Freshwater biodiversity: importance, threats, status and conservation challenges. Biol. Rev. 81: 163. doi:10.1017/S1464793105006950

Estrada, M., and E. Berdalet. 1997. Phytoplankton in a turbulent world. Sci. Mar. 61: 125–140.

Gallina, N., O. Anneville, and M. Beniston. 2011. Impacts of extreme air temperatures on cyanobacteria in five deep peri-alpine lakes. J. Limnol. 70: 186. doi:10.4081/jlimnol.2011.186

Gotelli, N. J., and R. K. Colwell. 2001. Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. Ecol. Lett. 4: 379–391. doi:10.1046/j.1461-0248.2001.00230.x

Grenney, W. J., D. A. Bella, and H. C. Curl. 1973. A theoretical approach to interspecific competition in phytoplankton communities. Am. Nat. 107: 405–425. doi:10.1086/282843

Haberman, K. L., L. B. Quetin, and R. M. Ross. 2003. Diet of the Antarctic krill (Euphausia superba Dana). J. Exp. Mar. Bio. Ecol. 283: 79–95. doi:10.1016/S0022-0981(02)00466-5

Hamilton, D., C. Carey, L. Arvola, and others. 2015. A Global Lake Ecological Observatory Network (GLEON) for synthesising high–frequency sensor data for validation of deterministic ecological models. Inl. Waters 5: 49–56. doi:10.5268/IW-5.1.566

Hillebrand, H., B. Blasius, E. T. Borer, and others. 2018. Biodiversity change is uncoupled from species richness trends: Consequences for conservation and monitoring M. Cadotte [ed.]. J. Appl. Ecol. 55: 169–184. doi:10.1111/1365-2664.12959

Huisman, J., and F. J. Weissing. 1999. Biodiversity of plankton by species oscillations and chaos. Nature 402: 407–410. doi:10.1038/46540

Huston, M. 1979. A general hypothesis of species diversity. Am. Nat. 113: 81–101. doi:10.1086/283366

Hutchinson, G. E. 1961. The paradox of the plankton. Am. Nat. 95: 137–145. doi:10.1086/282171

Imberger, J., and J. C. Patterson. 1989. Physical Limnology, p. 303–475. In Advances in Applied Mechanics. 27. 303-475. doi:10.1016/S0065-2156(08)70199-6

Jones, S. K., J. Ripplinger, and S. L. Collins. 2017. Species reordering, not changes in richness, drives long-term dynamics in grassland communities T. Coulson [ed.]. Ecol. Lett. 20: 1556–1565. doi:10.1111/ele.12864

Klug, J. L., D. C. Richardson, H. A. Ewing, and others. 2012. Ecosystem effects of a tropical cyclone on a network of lakes in northeastern North America. Environ. Sci. Technol. 46: 11693–11701. doi:10.1021/es302063v

Kondoh, M. 2001. Unifying the relationships of species richness to productivity and disturbance. Proc. R. Soc. B Biol. Sci. 268: 269–271. doi:10.1098/rspb.2000.1384

Levine, S. H. 1976. Competitive interactions in ecosystems. Am. Nat. 110: 903–910. doi:10.1086/283116

Page, T., P. J. Smith, K. J. Beven, and others. 2018. Adaptive forecasting of phytoplankton communities. Water Res. 134: 74–85. doi:10.1016/j.watres.2018.01.046

Pesant, S., L. Legendre, M. Gosselin, E. Bauerfeind, and G. Budéus. 2002. Wind-triggered events of phytoplankton downward flux in the Northeast Water Polynya. J. Mar. Syst. 31: 261–278. doi:10.1016/S0924-7963(01)00065-3

Read, J. S., D. P. Hamilton, I. D. Jones, K. Muraoka, L. A. Winslow, R. Kroiss, C. H. Wu, and E. Gaiser. 2011. Derivation of lake mixing and stratification

indices from high-resolution lake buoy data. Environ. Model. Softw. 26: 1325–1336. doi:10.1016/j.envsoft.2011.05.006

Reynolds, C. S. 1995. The intermediate disturbance hypothesis and its applicability to planktonic communities: Comments on the view of Padisak and Wilson. N. Z. J. Ecol. 19: 219–225.

Reynolds, C. S. 1993. Scales of disturbance and their role in plankton ecology. Hydrobiologia 249: 157–171. doi:10.1007/BF00008851

Richerson, P., R. Armstrong, and C. R. Goldman. 1970. Contemporaneous disequilibrium, a new hypothesis to explain the "paradox of the plankton". Proc. Natl. Acad. Sci. 67: 1710–1714. doi:10.1073/pnas.67.4.1710

Roelke, D. L., and P. M. Eldridge. 2008. Mixing of supersaturated assemblages and the precipitous loss of species. Am. Nat. 171: 162–175. doi:10.1086/524955

Rosindell, J., S. P. Hubbell, and R. S. Etienne. 2011. The unified neutral theory of biodiversity and biogeography at age ten. Trends Ecol. Evol. 26: 340–348. doi:10.1016/j.tree.2011.03.024

Sarker, S., U. Feudel, C. L. Meunier, P. Lemke, P. S. Dutta, and K. H. Wiltshire. 2018. To share or not to share? Phytoplankton species coexistence puzzle in a competition model incorporating multiple resource-limitation and synthesizing unit concepts. Ecol. Modell. 383: 150–159. doi:10.1016/j.ecolmodel.2018.05.021

Schippers, P., A. M. Verschoor, M. Vos, and W. M. Mooij. 2001. Does "supersaturated coexistence" resolve the "paradox of the plankton"? Ecol. Lett. 4: 404–407. doi:10.1046/j.1461-0248.2001.00239.x

Sommer, U., J. Padisák, C. S. Reynolds, and P. Juhász-Nagy. 1993. Hutchinson's heritage: the diversity-disturbance relationship in phytoplankton. Hydrobiologia 249: 1–7. doi:10.1007/BF00008837

Sommer, U., R. Adrian, L. De Senerpont Domis, and others. 2012. Beyond the plankton ecology group (peg) model: mechanisms driving plankton succession. Annu. Rev. Ecol. Evol. Syst. 43: 429–448. doi:10.1146/annurev-ecolsys-110411-160251

Spatharis, S., D. Mouillot, T. Do Chi, D. B. Danielidis, and G. Tsirtsis. 2009. A niche-based modeling approach to phytoplankton community assembly rules. Oecologia 159: 171–180. doi:10.1007/s00442-008-1178-8

Tilman, D. 1994. Competition and biodiversity in spatially structured habitats. Ecology 75: 2–16. doi:10.2307/1939377

Trimbee, A. M., and G. P. Harris. 1984. Phytoplankton population dynamics of a small reservoir: use of sedimentation traps to quantify the loss of diatoms and recruitment of summer bloom-forming blue-green algae. J. Plankton Res. 6: 897–918. doi:10.1093/plankt/6.5.897

Weithoff, G. 2001. The intermediate disturbance hypothesis--species diversity or functional diversity? J. Plankton Res. 23: 1147–1155. doi:10.1093/plankt/23.10.1147

Whittaker, R. H. 1972. Evolution and measurement of species diversity. taxon 21: 213. doi:10.2307/1218190

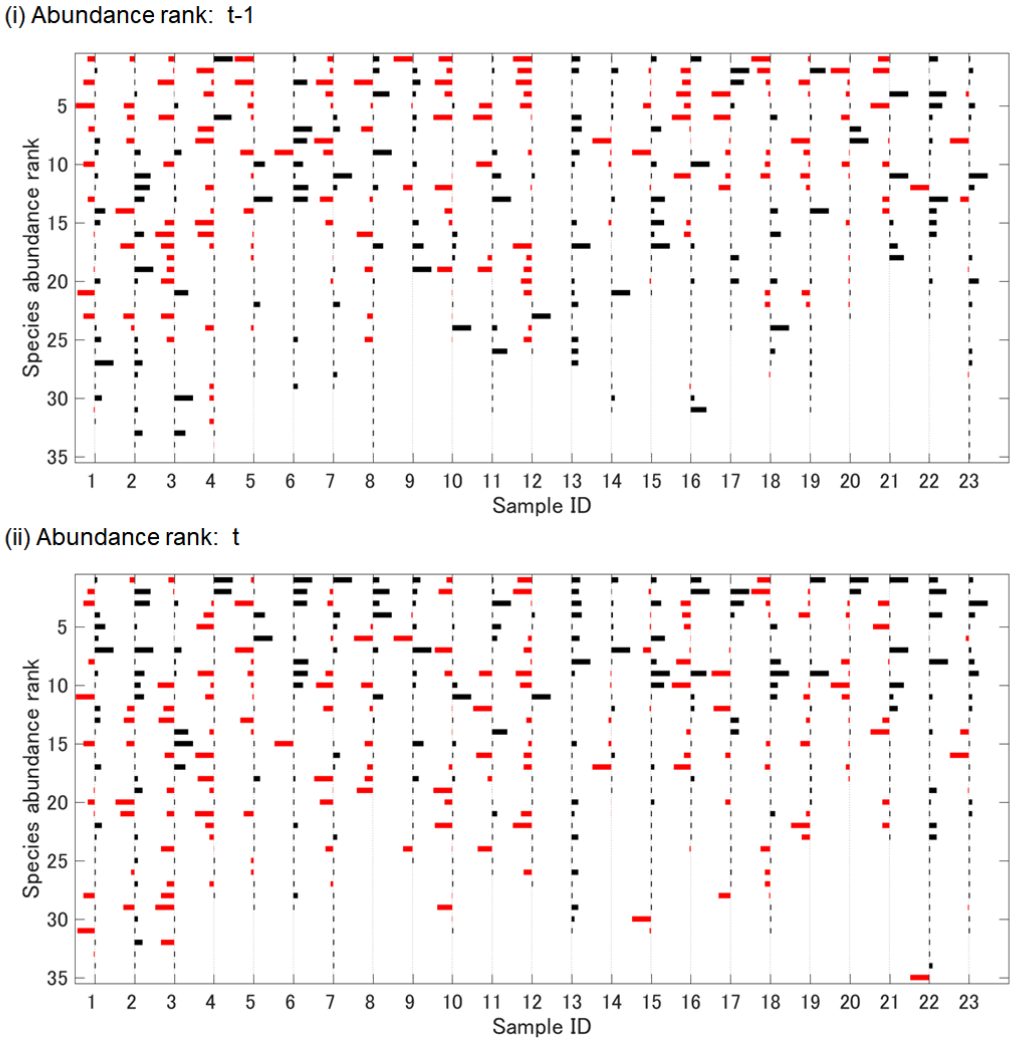
Williamson, C. E., J. E. Saros, W. F. Vincent, and J. P. Smol. 2009. Lakes and reservoirs as sentinels, integrators, and regulators of climate change. Limnol. Oceanogr. 54: 2273–2282. doi:10.4319/lo.2009.54.6_part_2.2273

Withrow, F. G., D. L. Roelke, R. M. W. Muhl, and J. Bhattacharyya. 2018. Water column processes differentially influence richness and diversity of neutral, lumpy and intransitive phytoplankton assemblages. Ecol. Modell. 370: 22–32. doi:10.1016/j.ecolmodel.2018.01.002

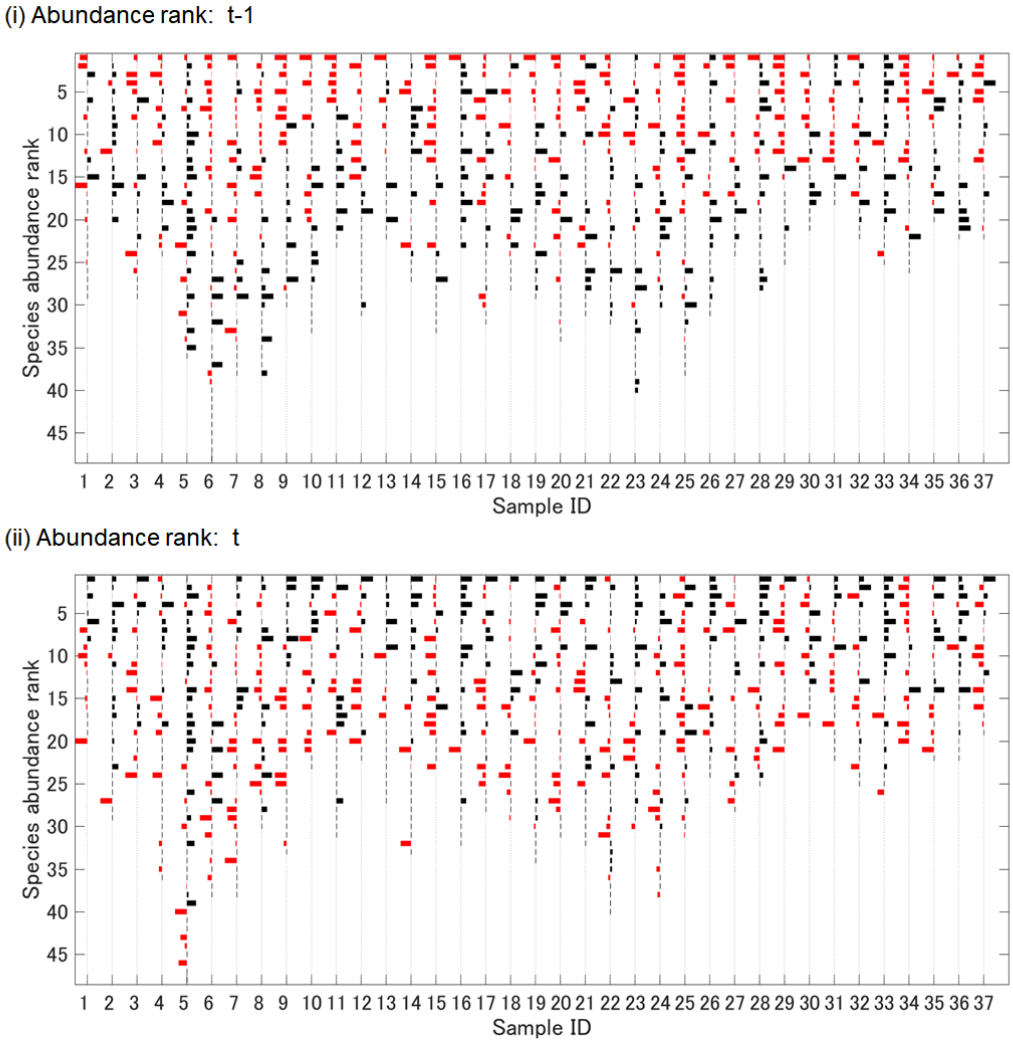
Wood, S. A., K. Jentzsch, A. Rueckert, D. P. Hamilton, and S. C. Cary. 2009. Hindcasting cyanobacterial communities in Lake Okaro with germination experiments and genetic analyses. FEMS Microbiol. Ecol. 67: 252–260. doi:10.1111/j.1574-6941.2008.00630.x

Wood, S. A., X. Pochon, L. Luttringer-Plu, B. N. Vant, and D. P. Hamilton. 2014. Recent invader or indicator of environmental change? A phylogenetic and ecological study of Cylindrospermopsis raciborskii in New Zealand. Harmful Algae 39: 64–74. doi:10.1016/j.hal.2014.06.013

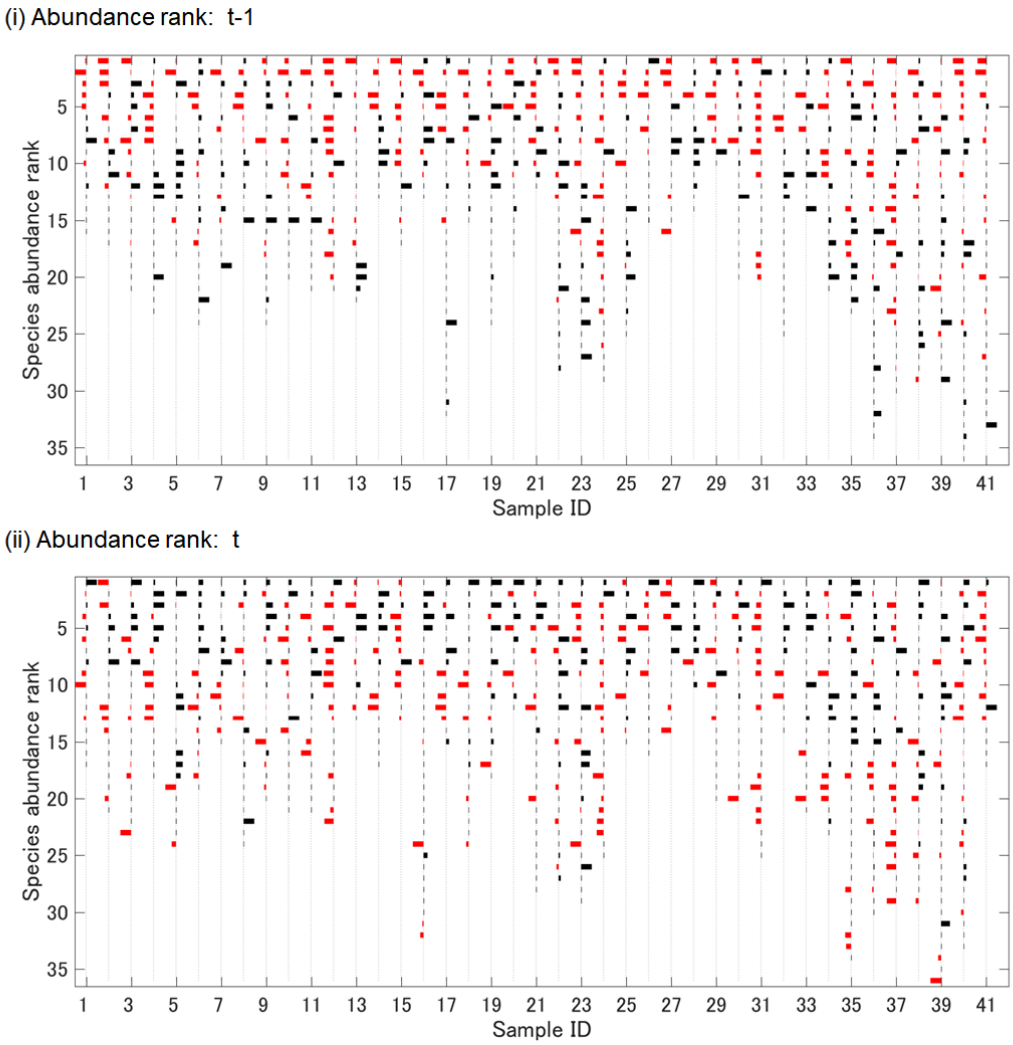
5.8 Supplementary materials



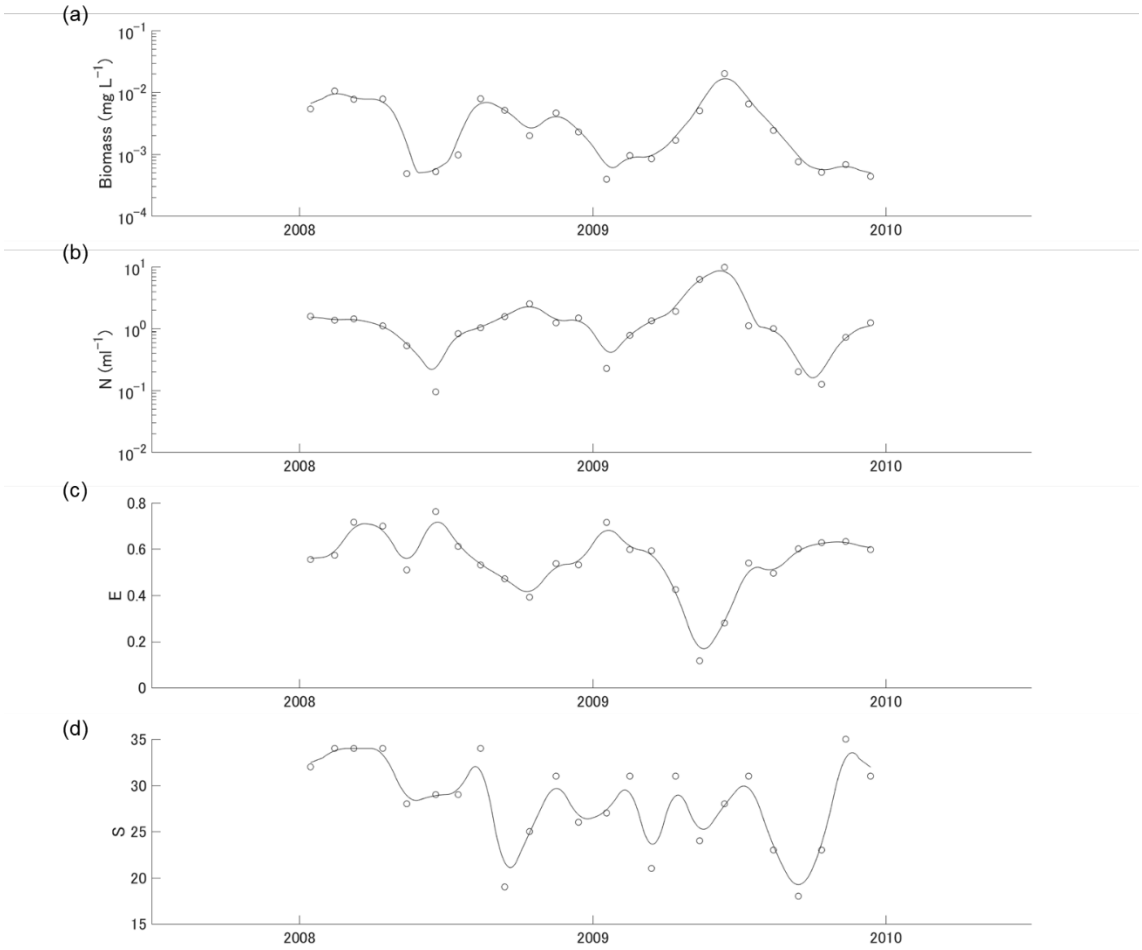
Appendix A-1. Lake Annie cell density growths (black) and decline (red) between two consecutive samples in relation to their previous (i; t-1) and current (ii; t, i.e. consequence of the growth/decline) population abundance ranks. The rate of increase and decrease were normalized by maximum values of increase and decrease of each sample and shown in logarithmic scale to visualise the relative trend of the dynamics.



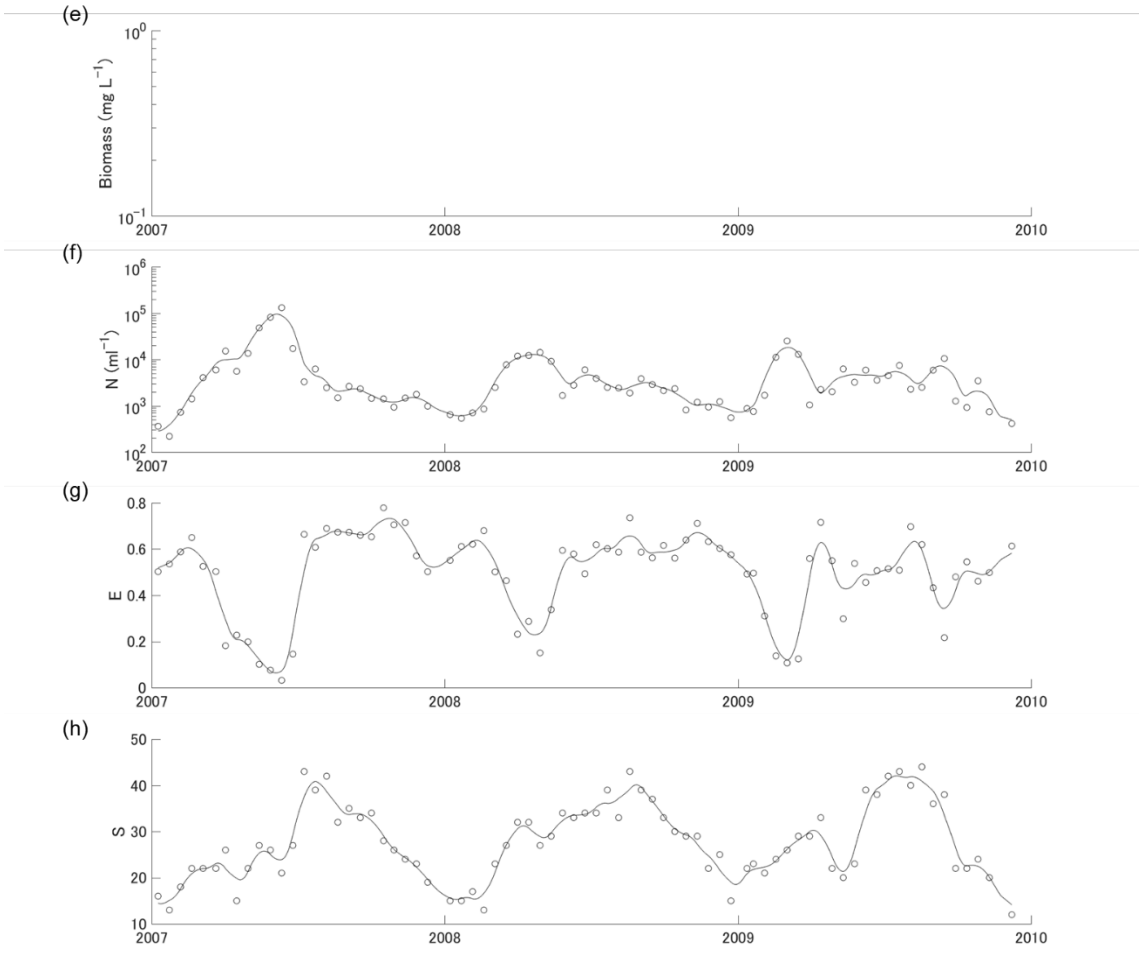
Appendix A-2. Lake Feeagh cell density growths (black) and decline (red) between two consecutive samples in relation to their current (t, i.e. consequence of the growth/decline) population abundance ranks.



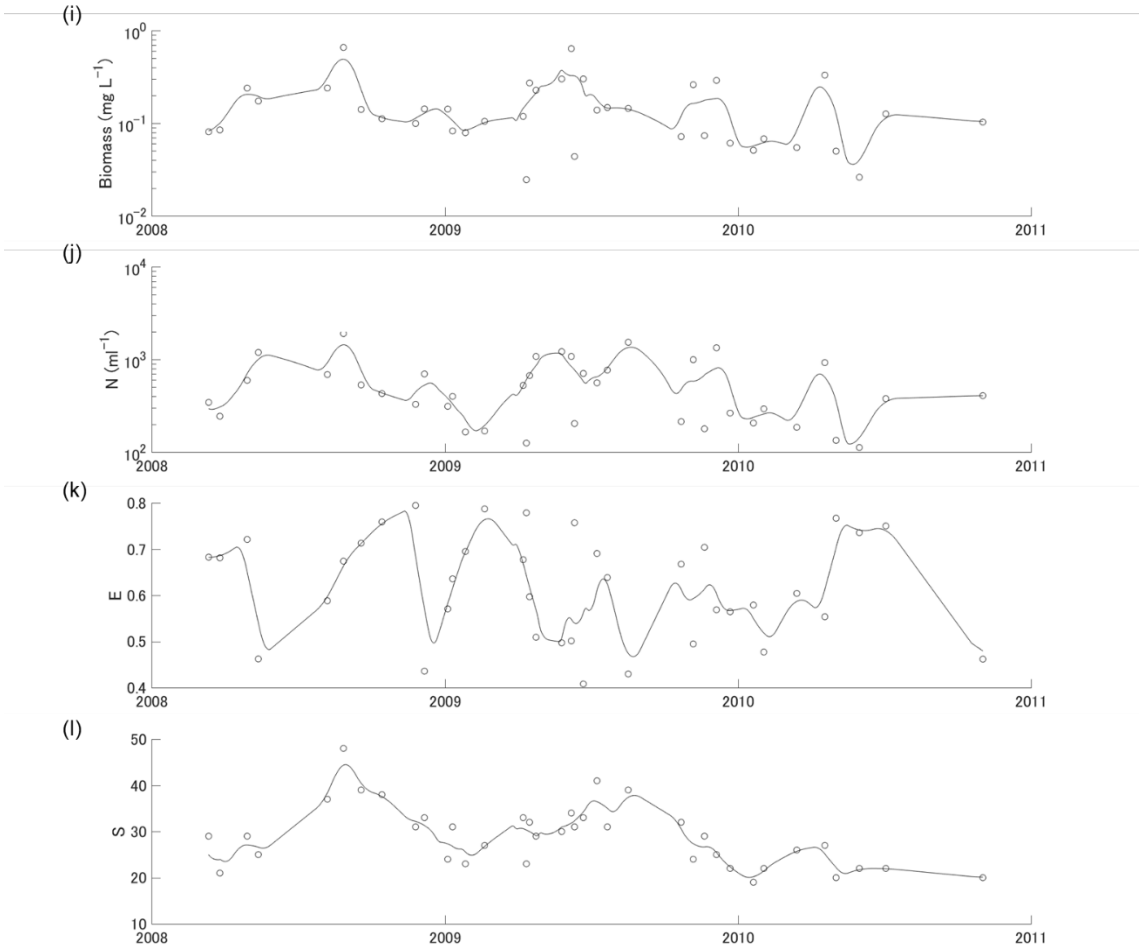
Appendix A-3. Lake Mendota cell density growths (black) and decline (red) between two consecutive samples in relation to their previous (i; t-1) and current (ii; t, i.e. consequence of the growth/decline) population abundance ranks.



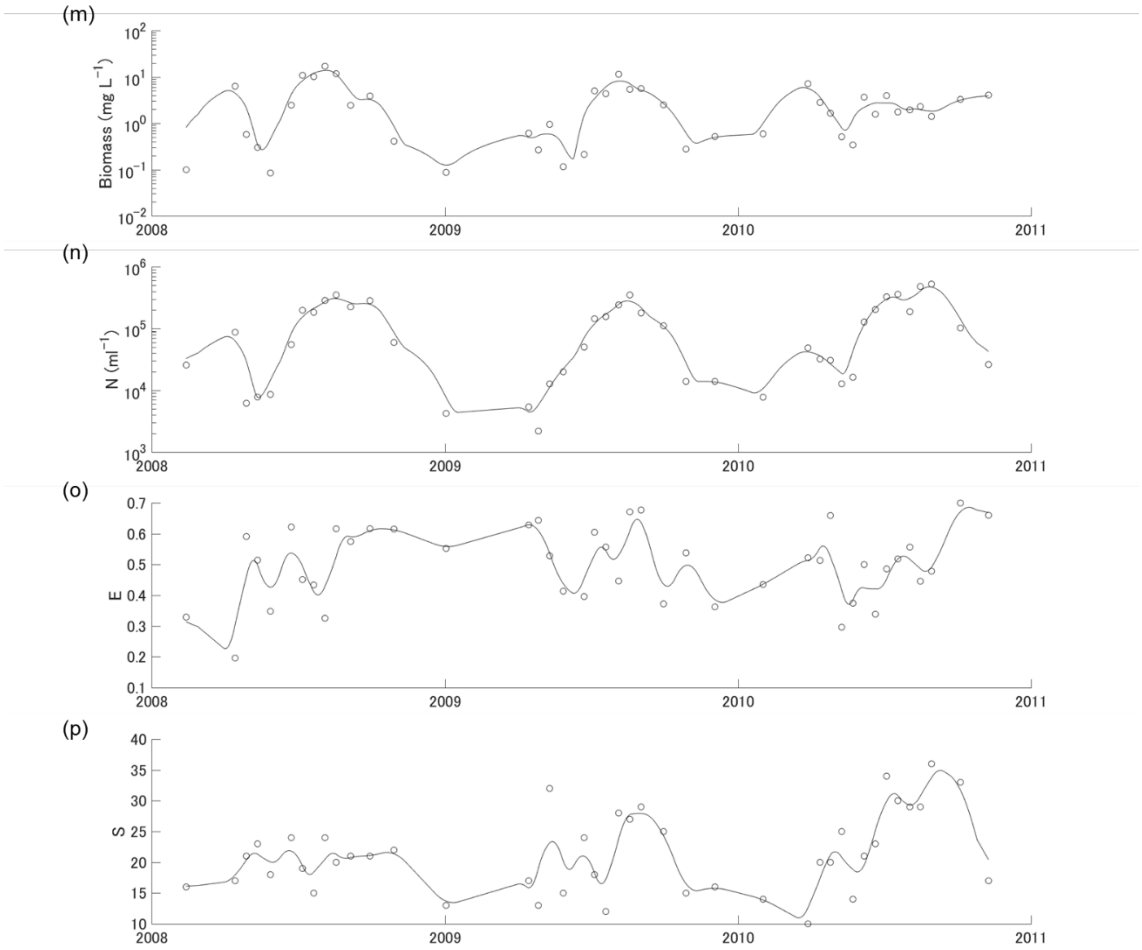
Appendix B-1. Time series of total biomass (a, e, i, m), abundance (b, f, j, n), evenness (c, g, k, o) and richness (d, h, l, p). The data were from Lakes Anni (a-d), Esthwaite (e-h), Feeagh (i-l), and Mendota (m-p). Biomass was estimated from average size and abundance when size observation was available.



Appendix B-2. cont'd. (Lake Esthwaite).



Appendix B-3. cont'd. (Loch Feeagh)



Appendix B-4. cont'd. (Lake Mendota)

Appendix Table 1a. Values of significant ($p < 0.05$) Pearson's correlation coefficient among all constituents of species richness (values not significant are represented by -). S(t-1): species richness at time t-1, S(t): species richness at time t, dS: species richness difference (S(t) – S(t-1)), Sr: number of newly recruited species from sample at time t, Si: number of species with increasing cell counts in sample at time t, Sd: number of species with decreasing cell counts in sample at time t, Se: number of newly extinct species at time t, Sc: total number of species at time t (Sr(t)+Si(t)+Sd(t)+Se(t)), and dt: days between times t-1 and t.

			$S_{(t-1)}$		**$S_{(t)}$**		**dS**		**Sr**		**Si**		**Sd**		**Se**		**Sr/Sc**		**Si/Sc**		**Sd/Sc**		**Se/Sc**		**ISR**	
Annie	Esthwaite	$S_{(t-1)}$	1.0	1.0	-	0.8	-0.7	-0.4	-	-	-	0.6	-	0.8	0.8	0.6	-0.6	-0.5	-	-	-	0.3	0.7	-	-0.4	-0.3
Feeagh	Mendota		1.0	1.0	-	-	-	-0.6	-	-	0.5	0.6	-	-	-	0.8	-0.5	-0.6	-	-	-	-	-	0.5	-	-0.4
Annie	Esthwaite	$S_{(t)}$	-	0.8	1.0	1.0	0.7	0.3	0.8	0.6	0.4	0.7	-	0.6	-	-	0.6	-	-	-	-	-	-0.6	-0.4	0.6	-
Feeagh	Mendota		-	-	1.0	1.0	0.7	0.6	0.7	0.8	0.8	-	-	0.5	-0.5	-	0.5	0.5	0.6	-	-	-	-0.8	-0.6	0.7	0.6
Annie	Esthwaite	dS	-0.7	-0.4	0.7	0.3	1.0	1.0	0.8	0.8	-	-	-	-0.3	-0.8	-0.8	0.9	0.9	-	0.3	-	-0.4	-0.9	-0.8	0.8	0.8
Feeagh	Mendota		-	-0.6	0.7	0.6	1.0	1.0	0.9	0.9	0.5	-	-	-	-0.8	-0.8	0.9	0.9	-	-	-	-	-0.9	-0.9	0.8	0.8
Annie	Esthwaite	Sr	-	-	0.8	0.6	0.8	0.8	1.0	1.0	-	0.3	-	-	-	-	0.9	0.8	-	-	-0.5	-0.4	-0.5	-0.5	0.7	0.6
Feeagh	Mendota		-	-	0.7	0.8	0.9	0.9	1.0	1.0	-	-	-	-	-0.5	-0.4	0.9	0.9	-	-	-0.6	-	-0.7	-0.7	0.8	0.7
Annie	Esthwaite	Si	-	0.6	0.4	0.7	-	-	-	0.3	1.0	1.0	-0.8	-	-	-	-	-	1.0	0.8	-0.8	-0.4	-	-0.4	0.8	0.5
Feeagh	Mendota		0.5	0.6	0.8	-	0.5	-	-	-	1.0	1.0	-	-	-	-	-	-	1.0	0.9	-0.6	-0.5	-0.5	-	0.8	-
Annie	Esthwaite	Sd	-	0.8	-	0.6	-	-0.3	-	-	-0.8	-	1.0	1.0	-	0.4	-	-0.5	-0.8	-0.5	0.9	0.8	-	-	-0.8	-0.6
Feeagh	Mendota		-	-	-	0.5	-	-	-	-	-	-	1.0	1.0	-	-	-0.5	-	-0.5	-0.4	0.9	0.8	-	-	-0.6	-
Annie	Esthwaite	Se	0.8	0.6	-	-	-0.8	-0.8	-	-	-	-	-	0.4	1.0	1.0	-0.5	-0.5	-	-0.4	-	-	0.9	0.8	-0.4	-0.6
Feeagh	Mendota		-	0.8	-0.5	-	-0.8	-0.8	-0.5	-0.4	-	-	-	-	1.0	1.0	-0.5	-0.6	-	-	-	-	0.9	0.9	-0.6	-0.7
Annie	Esthwaite	Sr/Sc	-0.6	-0.5	0.6	-	0.9	0.9	0.9	0.8	-	-	-	-0.5	-0.5	-0.5	1.0	1.0	-	-	-0.4	-0.6	-0.7	-0.6	0.8	0.8
Feeagh	Mendota		-0.5	-0.6	0.5	0.5	0.9	0.9	0.9	0.9	-	-	-0.5	-	-0.5	-0.6	1.0	1.0	-	-0.5	-0.5	-	-0.7	-0.7	0.8	0.7
Annie	Esthwaite	Si/Sc	-	-	-	-	-	0.3	-	-	1.0	0.8	-0.8	-0.5	-	-0.4	-	-	1.0	1.0	-0.8	-0.7	-	-0.5	0.8	0.8
Feeagh	Mendota		-	-	0.6	-	-	-	-	-	1.0	0.9	-0.5	-0.4	-	-	-	-0.5	1.0	1.0	-0.7	-0.5	-0.5	-	0.8	-
Annie	Esthwaite	Sd/Sc	-	0.3	-	-	-	-0.4	-0.5	-0.4	-0.8	-0.4	0.9	0.8	-	-	-0.4	-0.6	-0.8	-0.7	1.0	1.0	-	-	-0.8	-0.8
Feeagh	Mendota		-	-	-	-	-	-	-0.6	-	-0.6	-0.5	0.9	0.8	-	-	-0.5	-	-0.7	-0.5	1.0	1.0	-	-	-0.8	-0.4
Annie	Esthwaite	Se/Sc	0.7	-	-0.6	-0.4	-0.9	-0.8	-0.5	-0.5	-	-0.4	-	-	0.9	0.8	-0.7	-0.6	-	-0.5	-	-	1.0	1.0	-0.6	-0.7
Feeagh	Mendota		-	0.5	-0.8	-0.6	-0.9	-0.9	-0.7	-0.7	-0.5	-	-	-	0.9	0.9	-0.7	-0.7	-0.5	-	-	-	1.0	1.0	-0.8	-0.8
Annie	Esthwaite	ISR	-0.4	-0.3	0.6	-	0.8	0.8	0.7	0.6	0.8	0.5	-0.8	-0.6	-0.4	-0.6	0.8	0.8	0.8	0.8	-0.8	-0.8	-0.6	-0.7	1.0	1.0
Feeagh	Mendota		-	-0.4	0.7	0.6	0.8	0.8	0.8	0.7	0.8	-	-0.6	-	-0.6	-0.7	0.8	0.7	0.8	-	-0.8	-0.4	-0.8	-0.8	1.0	1.0
Annie	Esthwaite	$Sr/S_{(t-1)}$	-0.6	-0.5	0.6	-	0.9	0.9	0.9	0.8	-	-	-0.5	-0.5	-0.5	-0.5	1.0	1.0	-	-	-0.5	-0.6	-0.7	-0.6	0.8	0.7
Feeagh	Mendota		-0.5	-0.6	0.5	0.5	0.9	0.9	0.9	0.9	-	-	-0.5	-	-0.5	-0.6	1.0	1.0	-	-0.4	-0.5	-	-0.6	-0.7	0.8	0.7
Annie	Esthwaite	$Si/S_{(t-1)}$	-	-0.3	-	-	0.5	0.6	0.5	0.4	0.9	0.6	-0.9	-0.6	-	-0.6	0.5	0.5	1.0	0.9	-0.9	-0.8	-	-0.7	0.9	1.0
Feeagh	Mendota		-	-	0.7	-	0.7	-	0.6	-	0.9	0.8	-0.6	-0.5	-0.5	-	0.5	-	0.9	0.9	-0.8	-0.6	-0.7	-	0.9	0.7
Annie	Esthwaite	$Sd/S_{(t-1)}$	-	0.3	-	-	-	-	-	-	-0.8	-0.4	0.9	0.8	-	-	-	-0.4	-0.8	-0.7	1.0	1.0	-	-	-0.7	-0.7
Feeagh	Mendota		-	-0.4	-	-	-	0.5	-	-	-0.6	-0.6	0.9	0.7	-	-0.6	-	0.5	-0.7	-0.7	1.0	0.8	-	-0.6	-0.6	-
Annie	Esthwaite	$Se/S_{(t-1)}$	0.6	-	-	-0.4	-0.7	-0.7	-	-0.3	-	-0.4	-	-	0.9	0.8	-	-0.4	-	-0.5	-	-	0.9	1.0	-	-0.6
Feeagh	Mendota		-	0.5	-0.7	-0.5	-0.8	-0.8	-	-0.5	-0.6	-	-	-	0.9	0.9	-	-0.6	-0.6	-	-	-	1.0	1.0	-0.6	-0.7
Annie	Esthwaite	$Sr/S_{(t)}$	-	-0.5	0.6	-	0.8	0.8	0.9	0.8	-	-	-0.5	-0.5	-	-0.4	0.9	1.0	-	-	-0.5	-0.6	-0.5	-0.4	0.7	0.7
Feeagh	Mendota		-0.5	-0.5	-	0.4	0.7	0.8	0.9	0.9	-	-	-0.5	-	-	-	1.0	1.0	-	-0.6	-0.6	-	-	-0.5	0.6	0.6
Annie	Esthwaite	$Si/S_{(t)}$	-	-	-	-	-	-	-	-	0.9	0.7	-0.8	-0.5	-	-	-	-	1.0	0.9	-0.8	-0.7	-	-	0.6	0.6
Feeagh	Mendota		-	0.4	-	-	-	-0.5	-	-0.6	0.9	0.8	-0.6	-0.5	-	-	-	-0.7	0.9	0.9	-0.7	-0.6	-	0.4	0.6	-
Annie	Esthwaite	$Sd/S_{(t)}$	-	0.3	-0.4	-	-0.5	-0.6	-0.7	-0.5	-0.9	-0.5	0.9	0.8	-	0.4	-0.6	-0.7	-0.8	-0.8	1.0	1.0	-	0.5	-0.9	-0.9
Feeagh	Mendota		-	-	-0.5	-	-0.6	-	-0.7	-	-0.7	-0.5	0.8	0.7	-	-	-0.6	-	-0.8	-0.5	1.0	0.9	-	-	-0.9	-0.7
Annie	Esthwaite	$Se/S_{(t)}$	0.6	-	-0.6	-0.4	-0.9	-0.8	-0.6	-0.5	-	-0.4	-	-	0.9	0.8	-0.7	-0.6	-	-0.5	-	-	1.0	1.0	-0.6	-0.7
Feeagh	Mendota		-	0.5	-0.7	-0.6	-0.9	-0.9	-0.6	-0.6	-0.5	-	-	-	0.9	0.9	-0.6	-0.6	-0.5	-	-	-	1.0	1.0	-0.7	-0.7
Annie	Esthwaite	dt	-	-	-	-	-	-	-	-	-0.5	-	0.4	-	-	-	-	-	-0.5	0.2	0.4	-	-	-	-	-
Feeagh	**Mendota**		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

Appendix Table 1b.

			$Sr/S_{(t-1)}$		**$Si/S_{(t-1)}$**		**$Sd/S_{(t-1)}$**		**$Se/S_{(t-1)}$**		**$Sr/S_{(t)}$**		**$Si/S_{(t)}$**		**$Sd/S_{(t)}$**		**$Se/S_{(t)}$**		**dt**	
Annie	Esthwaite	$S_{(t-1)}$	-0.6	-0.5	-	-0.3	-	0.3	0.6	-	-	-0.5	-	-	-	0.3	0.6	-	-	-
Feeagh	Mendota		-0.5	-0.6	-	-	-	-0.4	-	0.5	-0.5	-0.5	-	0.4	-	-	-	0.5	-	-
Annie	Esthwaite	$S_{(t)}$	0.6	-	-	-	-	-	-	-0.4	0.6	-	-	-	-0.4	-	-0.6	-0.4	-	-
Feeagh	Mendota		0.5	0.5	0.7	-	-	-	-0.7	-0.5	-	0.4	-	-	-0.5	-	-0.7	-0.6	-	-
Annie	Esthwaite	dS	0.9	0.9	0.5	0.6	-	-	-0.7	-0.7	0.8	0.8	-	-	-0.5	-0.6	-0.9	-0.8	-	-
Feeagh	Mendota		0.9	0.9	0.7	-	-	0.5	-0.8	-0.8	0.7	0.8	-	-0.5	-0.6	-	-0.9	-0.9	-	-
Annie	Esthwaite	Sr	0.9	0.8	0.5	0.4	-	-	-	-0.3	0.9	0.8	-	-	-0.7	-0.5	-0.6	-0.5	-	-
Feeagh	Mendota		0.9	0.9	0.6	-	-	-	-	-0.5	0.9	0.9	-	-0.6	-0.7	-	-0.6	-0.6	-	-
Annie	Esthwaite	Si	-	-	0.9	0.6	-0.8	-0.4	-	-0.4	-	-	0.9	0.7	-0.9	-0.5	-	-0.4	-0.5	-
Feeagh	Mendota		-	-	0.9	0.8	-0.6	-0.6	-0.6	-	-	-	0.9	0.8	-0.7	-0.5	-0.5	-	-	-
Annie	Esthwaite	Sd	-0.5	-0.5	-0.9	-0.6	0.9	0.8	-	-	-0.5	-0.5	-0.8	-0.5	0.9	0.8	-	-	0.4	-
Feeagh	Mendota		-0.5	-	-0.6	-0.5	0.9	0.7	-	-	-0.5	-	-0.6	-0.5	0.8	0.7	-	-	-	-
Annie	Esthwaite	Se	-0.5	-0.5	-	-0.6	-	-	0.9	0.8	-	-0.4	-	-	-	0.4	0.9	0.8	-	-
Feeagh	Mendota		-0.5	-0.6	-0.5	-	-	-0.6	0.9	0.9	-	-	-	-	-	-	0.9	0.9	-	-
Annie	Esthwaite	Sr/Sc	1.0	1.0	0.5	0.5	-	-0.4	-	-0.4	0.9	1.0	-	-	-0.6	-0.7	-0.7	-0.6	-	-
Feeagh	Mendota		1.0	1.0	0.5	-	-	0.5	-	-0.6	1.0	1.0	-	-0.7	-0.6	-	-0.6	-0.6	-	-
Annie	Esthwaite	Si/Sc	-	-	1.0	0.9	-0.8	-0.7	-	-0.5	-	-	1.0	0.9	-0.8	-0.8	-	-0.5	-0.5	0.2
Feeagh	Mendota		-	-0.4	0.9	0.9	-0.7	-0.7	-0.6	-	-	-0.6	0.9	0.9	-0.8	-0.5	-0.5	-	-	-
Annie	Esthwaite	Sd/Sc	-0.5	-0.6	-0.9	-0.8	1.0	1.0	-	-	-0.5	-0.6	-0.8	-0.7	1.0	1.0	-	-	0.4	-
Feeagh	Mendota		-0.5	-	-0.8	-0.6	1.0	0.8	-	-	-0.6	-	-0.7	-0.6	1.0	0.9	-	-	-	-
Annie	Esthwaite	Se/Sc	-0.7	-0.6	-	-0.7	-	-	0.9	1.0	-0.5	-0.4	-	-	-	0.5	1.0	1.0	-	-
Feeagh	Mendota		-0.6	-0.7	-0.7	-	-	-0.6	1.0	1.0	-	-0.5	-	0.4	-	-	1.0	1.0	-	-
Annie	Esthwaite	ISR	0.8	0.7	0.9	1.0	-0.7	-0.7	-	-0.6	0.7	0.7	0.6	0.6	-0.9	-0.9	-0.6	-0.7	-	-
Feeagh	Mendota		0.8	0.7	0.9	0.7	-0.6	-	-0.6	-0.7	0.6	0.6	0.6	-	-0.9	-0.7	-0.7	-0.7	-	-
Annie	Esthwaite	$Sr/S_{(t-1)}$	1.0	1.0	0.5	0.5	-	-0.4	-	-0.4	0.9	0.9	-	-	-0.6	-0.7	-0.7	-0.5	-	-
Feeagh	Mendota		1.0	1.0	0.5	-	-	0.5	-	-0.6	1.0	0.9	-	-0.6	-0.7	-	-0.6	-0.6	-	-
Annie	Esthwaite	$Si/S_{(t-1)}$	0.5	0.5	1.0	1.0	-0.8	-0.8	-	-0.6	-	0.4	0.8	0.8	-0.9	-0.9	-	-0.6	-0.4	0.3
Feeagh	Mendota		0.5	-	1.0	1.0	-0.7	-0.5	-0.6	-	-	-	0.8	0.7	-0.9	-0.7	-0.6	-	-	-
Annie	Esthwaite	$Sd/S_{(t-1)}$	-	-0.4	-0.8	-0.8	1.0	1.0	-	-	-	-0.4	-0.9	-0.8	0.9	0.9	-	-	0.5	-
Feeagh	Mendota		-	0.5	-0.7	-0.5	1.0	1.0	-	-0.6	-	-	-0.8	-0.8	0.9	0.6	-	-0.6	-	-
Annie	Esthwaite	$Se/S_{(t-1)}$	-	-0.4	-	-0.6	-	-	1.0	1.0	-	-	-	-0.2	-	0.3	0.9	0.9	-	-
Feeagh	Mendota		-	-0.6	-0.6	-	-	-0.6	1.0	1.0	-	-	-	-	-	-	0.9	0.9	-	-
Annie	Esthwaite	$Sr/S_{(t)}$	0.9	0.9	-	0.4	-	-0.4	-	-	1.0	1.0	-	-	-0.7	-0.7	-0.5	-0.4	-	-
Feeagh	Mendota		1.0	0.9	-	-	-	-	-	-	1.0	1.0	-	-0.7	-0.6	-	-	-	-	-
Annie	Esthwaite	$Si/S_{(t)}$	-	-	0.8	0.8	-0.9	-0.8	-	-0.2	-	-	1.0	1.0	-0.8	-0.7	-	-	-0.6	-
Feeagh	Mendota		-	-0.6	0.8	0.7	-0.8	-0.8	-	-	-	-0.7	1.0	1.0	-0.7	-0.5	-	-	-	-
Annie	Esthwaite	$Sd/S_{(t)}$	-0.6	-0.7	-0.9	-0.9	0.9	0.9	-	0.3	-0.7	-0.7	-0.8	-0.7	1.0	1.0	-	0.5	-	-
Feeagh	Mendota		-0.7	-	-0.9	-0.7	0.9	0.6	-	-	-0.6	-	-0.7	-0.5	1.0	1.0	-	-	-	-
Annie	Esthwaite	$Se/S_{(t)}$	-0.7	-0.5	-	-0.6	-	-	0.9	0.9	-0.5	-0.4	-	-	-	0.5	1.0	1.0	-	-
Feeagh	Mendota		-0.6	-0.6	-0.6	-	-	-0.6	0.9	0.9	-	-	-	-	-	-	1.0	1.0	-	-
Annie	Esthwaite	dt	-	-	-0.4	0.3	0.5	-	-	-	-	-	-0.6	-	-	-	-	-	1.0	1.0
Feeagh	**Mendota**		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1.0	1.0

Appendix Table 2. Spearman's rank correlation of previous ($S_{(t-1)}$) and current ($S_{(t)}$) species richness values, change in S ($S_{(t)}$ - $S_{(t-1)}$), constituents of richness (Sr, Si, Sd & Se) as well as proportion of these constituents (Sr/Sc, Si/Sc, Sd/Sc, Se/S & (Si+Sr)/Sc = ISR). For lakes Feeagh and Mendota, observation intervals greater than three weeks were removed from the analysis. Only values with $p < 0.05$) are shown.

Lakes		Variables	$S_{(t)}$		$S_{(t)}$ - $S_{(t-1)}$		Sr		Si		Sd		Se		Sr/Sc		Si/Sc		Sd/Sc		Se/Sc		ISR	
Annie	Esthwaite	$S_{(t-1)}$	-	0.76	-0.71	-0.38	-	-	-	0.57	-	0.76	0.81	0.62	-0.67	-0.48	-	-	-	0.33	0.63	-	-	-0.36
Feeagh	Mendota		-	-	-	-0.52	-	-	-	0.55	-	-	-	0.79	-	-0.52	-	-	-	-	-	0.43	-	-0.39
Annie	Esthwaite	$S_{(t)}$			0.55	0.26	0.74	0.55	0.43	0.72	-	0.54	-	-	0.43	-	-	-	-	-	-0.52	-0.38	0.54	-
Feeagh	Mendota				0.71	0.68	0.74	0.74	0.73	-	-	0.54	-	-	0.53	0.58	0.58	-	-0.47	-	-0.67	-0.68	0.81	0.59
Annie	Esthwaite	$S_{(t)}$ - $S_{(t-1)}$					0.64	0.71	-	-	-	-0.32	-0.82	-0.73	0.88	0.86	-	0.32	-	-0.41	-0.92	-0.84	0.72	0.75
Feeagh	Mendota						0.81	0.91	-	-	-	-	-0.85	-0.76	0.80	0.91	-	-	-	-	-0.93	-0.93	0.84	0.81
Annie	Esthwaite	Sr							-	0.29	-	-	-	-	0.79	0.81	-	-	-0.53	-0.42	-	-0.41	0.70	0.53
Feeagh	Mendota								-	-	-	-	-	-0.47	0.92	0.95	-	-	-0.54	-	-0.64	-0.79	0.78	0.70
Annie	Esthwaite	Si									-0.76	-	-	-	-	-	0.95	0.71	-0.87	-0.40	-	-0.36	0.74	0.48
Feeagh	Mendota										-	-	-	-	-	-	0.95	0.91	-0.68	-0.43	-0.45	-	0.75	-
Annie	Esthwaite	Sd											-	0.39	-0.48	-0.50	-0.82	-0.52	0.91	0.81	-	-	-0.74	-0.68
Feeagh	Mendota												-	-	-	-	-0.51	-	0.86	0.81	-	-	-0.52	-
Annie	Esthwaite	Se													-0.56	-0.48	-	-0.41	-	-	0.92	0.81	-0.42	-0.56
Feeagh	Mendota														-0.48	-0.57	-0.44	-	-	-	0.93	0.79	-0.65	-0.67
Annie	Esthwaite	Sr/Sc															-	-	-0.48	-0.60	-0.66	-0.52	0.82	0.73
Feeagh	Mendota																-	-0.40	-	-	-0.63	-0.79	0.69	0.71
Annie	Esthwaite	Si/Sc																	-0.85	-0.66	-	-0.48	0.73	0.77
Feeagh	Mendota																		-0.67	-	-0.50	-	0.76	-
Annie	Esthwaite	Sd/Sc																			-	-	-0.78	-0.84
Feeagh	Mendota																				-	-	-0.69	-
Annie	Esthwaite	Se/Sc																					-0.61	-0.64
Feeagh	Mendota																						-0.79	-0.86

Appendix Table 3. Spearman's rank correlation of previous ($S_{(t-1)}$) and current ($S_{(t)}$) species richness values, change in S ($S_{(t)}$ - $S_{(t-1)}$), constituents of richness (Sr, Si, Sd & Se) as well as proportion of these constituents (Sr/Sc, Si/Sc, Sd/Sc, Se/S & (Si+Sr)/Sc = ISR). Values are only shown for $p < 0.05$. All results are from Lake Esthwaite, but with various intervals including two (actual sampling frequency), four and eight weeks. Pearson's correlation coefficient values are given for randomly selected intervals (from two, four, six and eight weeks), and averages of ten iteration results are shown (denoted as "Random").

Intervals (weeks)		Variables	$S_{(t)}$		$S_{(t)}$ - $S_{(t-1)}$		Sr		Si		Sd		Se		Sr/Sc		Si/Sc		Sd/Sc		Se/Sc		ISR	
Two	Four	$S_{(t-1)}$	0.76	-	-0.38	-0.52	-	-	0.57	0.55	0.76	-	0.62	0.79	-0.48	-0.52	-	-	0.33	-	-	0.43	-0.36	-0.39
Eight	Random		-	-	-0.52	-0.69	-	-	0.55	-	-	0.64	0.79	0.52	-0.52	-0.54	-	-	-	0.51	0.43	0.37	-0.39	-0.53
Two	Four	$S_{(t)}$			0.26	0.68	0.55	0.74	0.72	-	0.54	0.54	-	-	-	0.58	-	-	-	-	-0.38	-0.68	-	0.59
Eight	Random				0.68	0.50	0.74	0.62	-	0.68	0.54	-	-	-	0.58	-	-	-	-	-	-0.68	-0.48	0.59	0.35
Two	Four	$S_{(t)}$ - $S_{(t-1)}$					0.71	0.91	-	-	-0.32	-	-0.73	-0.76	0.86	0.91	0.32	-	-0.41	-	-0.84	-0.93	0.75	0.81
Eight	Random						0.91	0.69	-	0.44	-	-	-0.76	-0.58	0.91	0.69	-	0.42	-	-0.50	-0.93	-0.71	0.81	0.74
Two	Four	Sr							0.29	-	-	-	-	-0.47	0.81	0.95	-	-	-0.42	-	-0.41	-0.79	0.53	0.70
Eight	Random								-	-	-	-	-0.47	-0.36	0.95	0.89	-	-	-	-0.58	-0.79	-0.56	0.70	0.70
Two	Four	Si									-	-	-	-	-	-	0.71	0.91	-0.40	-0.43	-0.36	-	0.48	-
Eight	Random										-	-	-	-	-	-	0.91	0.79	-0.43	-	-	-0.48	-	0.52
Two	Four	Sd											0.39	-	-0.50	-	-0.52	-	0.81	0.81	-	-	-0.68	-
Eight	Random												-	0.45	-	-0.60	-	-0.43	0.81	0.89	-	-	-	-0.70
Two	Four	Se													-0.48	-0.57	-0.41	-	-	-	0.81	0.79	-0.56	-0.67
Eight	Random														-0.57	-0.60	-	-0.56	-	-	0.79	0.90	-0.67	-0.75
Two	Four	Sr/Sc															-	-0.40	-0.60	-	-0.52	-0.79	0.73	0.71
Eight	Random																-0.40	-	-	-0.69	-0.79	-0.65	0.71	0.82
Two	Four	Si/Sc																	-0.66	-	-0.48	-	0.77	-
Eight	Random																		-	-0.48	-	-0.59	-	0.68
Two	Four	Sd/Sc																			-	-	-0.84	-
Eight	Random																				-	-	-	-0.79
Two	Four	Se/Sc																					-0.64	-0.86
Eight	Random																						-0.86	-0.81

Chapter Five
Synthesis and future perspectives

5.9 Overview

It is relatively new that functioning of an ecosystem is associated with a changing biological community (Jax, 2005). However, numerical expression of ecosystem function and changing diversity is a challenging task. Quantifying how ecosystem functioning fluctuates requires reference to normative status of the system. Establishment of normative status requires elimination of strong physical interferences or internal biogeochemical activities. Species assemblage perturbations may be a major driver of species richness changes (Hillebrand et al., 2018), therefore, it is important to consider how population or assemblage dynamics change richness. This thesis uses new approaches to express functioning of lake ecosystems and the biological community changes. Chapter two identified the times when external forcing in a lake most influenced lake ecosystem simulation outputs, resulting in suppression of the influence of internal biogeochemical processes. This was achieved by examining the simulation result spreads using modifications of the biogeochemical model parameters. Chapter three identified times when surface dissolved oxygen (DO) signals were predominantly influenced by biological processes using in-situ high frequency observations. During these periods, biological processes dominated the signal and there were minimal evidences of physical interferences in the data. Chapter four used conventional lake phytoplankton observations, where population changes and species counts were collectively used to inform species richness constituents as the population changed dynamically. The primary findings of the research chapters are provided below, followed by future research directions.

5.10 Research summary

Chapter two showed how parameter uncertainty analysis is often overlooked in lake ecological numerical modelling practice. Communication of uncertainty is highly important for developing confidence in the model output, and the resultant

uncertainty distribution can help decision makers determine suitable management practices for implementation. This chapter applied a Monte-Carlo parameter perturbation to a calibrated model and generated an ensemble of model simulations. The one-dimensional (1-D) lake ecological model DYRESM-CAEDYM was manually calibrated to Lake Waahi, and multiple parameters were perturbed from their original calibrated values (±5, 10, 25 and 50% perturbations). The ensemble results stabilised within 2000 model iterations. This number of iterations is easily within the capability of current computing hardware, and this sensitivity analysis could therefore be implemented as standard practice. Visualization of the ensemble results, including the spread and density of ensemble results, showed that biogeochemical parameters that primarily influenced in-lake processes had lesser sensitivity as inflow volume increased, corresponding to longer residence time. The parameter sensitivity changes across seasons and events imply that calibration could also change with season and hydrological state. Observations of changes in the dominant processes in the lake model are interesting and not well explored.

Chapter three explored interdisciplinary methods for evaluation of high-frequency dissolved oxygen observations for which there has tended to be strong emphasis conventionally on biological rather than physical parameters. The motivation for chapter three was assessment of data quality in lake metabolism assessments. Ecosystem metabolism models have been used to study the productivity from continuous dissolved oxygen observations. The method suffers from "noisy" environmental observations that neither post-model QA/QC processing or manual visual filtering can necessarily fully correct. The former method is affected by the parameterization method and model choice, whilst the latter method is simply labour intensive and not repeatable. Therefore, this chapter took advantage of both expert decision and automation methods by applying expert panel and data mining methods. The expert panel system was used to classify time series observations into categories of data quality. The Symbolic Aggregate approXimation (SAX) method was used to transform the original time series of dissolved oxygen to produce variables for a model that simulated expert panel decisions, using a logistic method. The model performance adequately reproduced the collective experts' decisions on the suitability of raw dissolved oxygen observations for a metabolism model. As a result of SAX approximation, the estimates of metabolism became robust and applicable to wider range of data. The results illustrated that the dominant DO diurnal patterns of 18 global lakes are not homogeneous but have different shapes.

SAX has been applied in information technology (IT) studies but has had limited application in the field of ecology, however, it appears to have outstanding potential to be an excellent tool for knowledge discovery in ecological settings.

Chapter four used conventional phytoplankton monitoring data from four lakes (Lakes Annie, Feeagh, Esthwaite and Mendota) to analyse in detail the constituents of species richness. The four constituents, namely the number of species that were recruited, went extinct, increased in abundance and decreased in abundance. Simplicity in expression of population dynamics (in binary term: increased or decreased) of each species should have provided less biased information regarding the well-being of species than the use of numeric expression of population changes. The constituents were expressed in relative proportion to total species richness. Negative correlation was found between proportion of number of species increasing and decreasing, but no significant correlation was found between number of species went extinct and decreasing. It was found that number of recruited species significantly correlated with the ratio of increased species to previous samples' richness in three of the study lakes. Another analysis found that fastest growing species to slowest growing species (in discrete samples) linearly distributes their rate of population change, while the slope of the line was strongly influenced by number of species increasing. The techniques developed in this study can be used for verification of population models. The proportion of increasing and recruited species to the total number of species, ISR, showed similar and different features to richness. Both ISR and richness increased during spring time, reflecting the increase of ecosystem productivity. ISR decreased dramatically during thermal stratification, likely due to limited resources and intense competition, while richness only gradually decreased. Destratification in three study lakes coincided with further decrease in ISR, while it increased significantly in Lake Annie, probably because of long stratification duration and reduced nutrient concentration of the lake largely benefited from nutrient injection from mixing.

Throughout the thesis, I created tools to express functioning of ecosystems (Chapters 2-3),and changing biological communities (Chapter 4). Functioning of the system was shown in Chapter 2 as model ensemble spreads, by interpolating wide-spread ensemble results that relate to when biogeochemical model parameters are dominating the results as opposed to when external forces limited the influence

of in-lake processes and the model ensemble spread is narrow. In Chapter 3, functioning of the biology was defined in a binary form related to when a system is dominated by biological activities. Expression of changes in the biological community is challenging, mainly due to relatively long intervals between samples. However, by expressing each species' population dynamics in binary format and counting increases/decreases in species, Chapter 4 provided dynamic metrics for diversity, using individual species population dynamics status.

5.11 Implications and future research directions

The methodological tools introduced in this thesis are used to express functioning of ecosystems and changes in biological communities. The underlying motivation was to understand the processes that interfere with, and disrupt, biological activities in lakes. Without a good knowledge of what constitutes a disturbance that works to interrupt the normative state of functioning of the system, it is difficult to formulate succession-disturbance based models. For example, every ecological community exists as a result of adaptation and succession under a unique set of conditions, and in phytoplankton, these conditions can change over very short time scales (e.g., days). The magnitude and frequency of a disturbance varies, and no single factor can describe the full extent of a disturbance. Instead, one can look to how a community behaves in response to a disturbance.

Relating parameter uncertainty to inflow volume in Chapter 2 helped to explain the dynamic balance between internal lake dynamics and external forces. This chapter has provided a means to understand the temporally changing dominance of internal versus external processes, which has not been well studied in the past. Where sensitivity of model simulation output to biogeochemical parameters increases, it is likely that internal processes are operating with relatively little interference from external forces, and vice-versa. The perturbed model ensemble spreads can be employed in ecological modelling projects to examine the collective sensitivity to parameters. Together with individual parameter sensitivity analysis, the ensemble output spread should express not only uncertainty of the model but also assist with calibration processes and ultimately improve model formulation.

In relation to Chapter 3, changes in dissolved oxygen can be used to indicate the net effect of a number of ecological processes operating in lakes. Being able to

detect when oxygen signals are atypical can be used to indicate ecosystem processes that are not encompassed within conventional lake metabolism models. These periods can be used to examine the magnitude and frequency of physical interferences that may critically influence ecosystem behaviour. These physical interferences may relate to heterogeneity arising from seasonal changes, various stratification regimes or interannual variations. Automatic identification allows metabolism models to be run in an automated way but also opens up new physics-ecology integrated research opportunities. The SAX method used in this thesis appears to be a promising tool for exploring interactions of physics and biology, and is applicable to different variables. For example, recent autonomous high-frequency observations from lake buoys (Hamilton et al., 2015) would be highly amenable to SAX analysis, to allow for pattern detection and carry out unique inter-or-intra lake feature comparisons.

The use of population abundance constituents in Chapter 4 provides additional new analysis methods for conventional phytoplankton observations. Retrospective analysis of increases and decreases in ISR (proportion of species that were increasing) through time may be used to describe the net effect of disturbance on a phytoplankton community. A descriptive model of ISR may be created using functioning of the system, for instance, using ecosystem metabolism as well as knowledge from Chapter 3 on whether biological processes were dominating the system. ISR, together with the four species richness constituents, should be compared with metrics that signal biodiversity changes, such as species turnover rate or proportional abundance change metrics (Hillebrand et al., 2018). By using more frequent and regular observations, it is possible to analyse detailed time series of assemblage changes without losing too many of the processes that may be important as drivers of diversity changes.

5.12 References

Hamilton, D., C. Carey, L. Arvola, and others. 2015. A Global Lake Ecological Observatory Network (GLEON) for synthesising high–frequency sensor data for validation of deterministic ecological models. Inl. Waters 5: 49–56. doi:10.5268/IW-5.1.566

Hillebrand, H., B. Blasius, E. T. Borer, and others. 2018. Biodiversity change is uncoupled from species richness trends: Consequences for conservation and monitoring M. Cadotte [ed.]. J. Appl. Ecol. 55: 169–184. doi:10.1111/1365-2664.12959

Jax, K. 2005. Function and "functioning" in ecology: what does it mean? Oikos 111: 641–648. doi:10.1111/j.1600-0706.2005.13851.x

Chapter Six

Appendix

Ethics approval (Chapter 3)