A Contribution Towards Real-Time Forecasting of Algal Blooms in Drinking Water Reservoirs by means of Artificial Neural Networks and Evolutionary Algorithms

By Amber Lee Welk
Bachelor of Natural Resource Management (Hons.)

A Thesis Submitted for the Degree of
Doctor of Philosophy

Discipline of Ecology and Evolutionary Biology
School of Earth and Environmental Science
The University of Adelaide
December 2007

TABLE OF CONTENTS

ABSTRACT

Historical water quality databases from two South Australian drinking water reservoirs were used, in conjunction with various computational modelling methods for the ordination, clustering and forecasting of complex ecological data. Techniques used throughout the study were: Kohonen artificial neural networks (KANN) for data categorisation and the discovery of patterns and relationships, recurrent supervised artificial neural networks (RANN) for knowledge discovery and forecasting of algal dynamics and hybrid evolutionary algorithms (HEA) for rule-set discovery and optimisation for forecasting algal dynamics. These methods were combined to provide an integrated approach to the analysis of algal populations including interactions within the algal community and with other water quality factors, which results in improved understanding and forecasting of algal dynamics.

The project initially focussed on KANN for the patternising and classification of the historical data to reveal links between the physical, chemical and biological components of the reservoirs. This offered some understanding of the system and relationships being considered for the construction of the forecasting models. Specific investigations were performed to examine past conditions and the impacts of different management regimes, as well as to discover sets of conditions that correspond with specific algal functional groups.

RANN was then used to build models for forecasting both Chl-a and the main nuisance species, *Anabaena*, up to 7 days in advance. This method also provided sensitivity analyses to demonstrate the relationship between input and output variables by plotting the reaction of the output to variations in the inputs. Initially one year from the data set was selected for the testing of a model, as per the split-sample technique. To further test the models, it was later decided to select several years for testing to ensure the models were useful under changed conditions, and that test results were not misleading regarding the models true capabilities. RANN were firstly used to create reservoir specific or ad-hoc models. Later, the models were trained with the merged data sets of both reservoirs to create one model that could be applied to either reservoir.

Another method of forecasting was trialled and compared to RANN. HEA was found to be equal or superior to RANN in predictive power, also allowed sensitivity analysis and provided an explicit, portable rule set. The HEA rule sets were initially tested on selected years of data,

however to fully demonstrate the models potential, a process for *k*-fold cross-validation was developed to test the rule-set on all years of data. To further extend the applicability of the HEA rule-set; the idea of rule-based agents for specific lake ecosystem categories was examined. The generality of a rule-based agent means that, after successful validation on several lakes from one category, the agent could then be applied to other water bodies from within that category that had not been involved in the training process. The ultimate test of the rule-based agent for the warm monomictic and eutrophic lake ecosystem category was to be applied to a real-time monitoring and forecasting situation. The agent was fed with online, real-time data from a reservoir that belonged to the same ecosystem category but was not used in the training process. These preliminary experiments showed promising results. It can be concluded that the concept of rule-based agents will facilitate real-time forecasting of algal blooms in drinking water reservoirs provided on-line monitoring of relevant variables has been implemented.

Contributions of this research include: (1) to offer insight into the capabilities of 3 kinds of computational modelling techniques applied to complex water quality data, (2) novel applications of KANN including the division of data into separate management periods for comparison of management efficiency, (3) to both qualitatively and quantitatively elucidate relationships between water quality parameters, (4) research toward the development of a forecasting tool for algal abundance 7 days in advance that could be generic for a particular lake ecosystem category and implemented in real-time, and (5) to suggest a thorough testing method for such models (*k*-fold cross validation).

STATEMENT OF ORIGINALITY

This work contains no material that has been accepted for the award of any other degree or diploma in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text.

I give consent to this copy of my thesis, when deposited in the University Library, being available for loan and photocopying, subject to the provisions of the Copyright Act 1968.

Signed:……. Date:……..

Amber Lee Welk

ACKNOWLEDGMENTS

I would like to thank my supervisors Associate Professor Friedrich Recknagel, Associate Professor Holger Maier and Mike Burch. Fred, my principal supervisor, has been invaluable with his enthusiasm and guidance – and his knack of pushing me when I need it!

This work was made possible by the financial support of the CRC for Water Quality and Treatment and I'd like to thank them for furthering my learning and development by allowing me to attend and present at numerous conferences throughout my candidature.

I must acknowledge Greg Ingleton, not just for the provision of the Happy Valley data, but also for patiently responding to all my pesky queries.

A big thank you goes to Dr Hongqing Cao, for developing the HEA that was used during the project.

Thanks to all in the Ecoinformatics and Watershed Ecology Lab for providing a friendly working environment. And special thanks must go to my fellow students Lydia Cetin, Anita Talib, Grace Chan and Bridget McDowell for sharing in my achievements and laughing at the disasters that have occurred throughout the project!

Most importantly, I'd like to express my appreciation for my family and friends. Firstly, thanks go to Dr Grewal for what can only be described politely as 'perseverance' in convincing me to go down this road. My friends have been a great help in ensuring I don't take things too seriously all the time, for their support and distractions I thank them very much. Thanks to the world's greatest grandparents for letting me retreat to Wilmington to hide from my project when it got too much. As in everything I do, my exceptionally patient mum has been my number one supporter, and her pride in me and what I accomplish makes me think I must be doing something right. My mum, my sister Naomi (who offered countless alternatives to work and a vast supply of chocolate) and partner Anthony have borne the brunt of my frustrations throughout the project but have been unwavering in their support and have offered endless encouragement. I hope you all know that you are part of anything and everything I achieve in life.

LIST OF PUBLICATIONS

Conference Presentations

Recknagel F, Welk A, Kim B, Takamura N. Explanation of processes determining phytoplankton abundances and succession in two lakes with different trophic states by means of long-term data patterns and non-supervised Artificial Neural Networks. *4th Conference of the International Society for Ecological Informatics (ISEI4), BEXCO, Busan, Korea, 24-28 October 2004.*

Welk A, Recknagel F, Burch M. Ordination, clustering and forecasting of phytoplankton dynamics in the Myponga drinking water reservoir by means of supervised and non-supervised artificial neural networks. *MODSIM05, International Congress on Modelling and Simulation: Advances and Applications for Management and Decision Making, Melbourne, Australia, 12-15 December 2005.*

Welk A. Towards A Rule-based Agent for Forecasting Chlorophyll-a Concentrations in South Australian Drinking Water Reservoirs. *Cooperative Research Centre (CRC) for Water Quality and Treatment, 5th Postgraduate Student Conference, Melbourne, Victoria, 10-13 July 2006.*

Welk A, Recknagel F, Cao H, Chan W-S, Talib A. Rule-based agents for forecasting algal population dynamics in freshwater lakes discovered by hybrid evolutionary algorithms. *5th Conference of the International Society for Ecological Informatics (ISEI5), Santa Barbara, USA, 4-6 December 2006.*

Peer Reviewed Papers

Welk A, Recknagel F, Burch M. (2005) Ordination, clustering and forecasting of phytoplankton dynamics in the Myponga drinking water reservoir by means of supervised and non-supervised artificial neural networks. Proceedings of the International Congress on Modelling and Simulation: Advances and Applications for Management and Decision Making, MODSIM05, Melbourne, Australia, 12-15 December 2005.

Recknagel F. Welk A, Kim B, Takamura N. (2006) Artificial Neural Network Approach to Unravel and Forecast Algal Population Dynamics of Two Lakes Different in Morphometry and Eutrophication. In: Ecological Informatics – Scope, Techniques and Applications (Ed: F. Recknagel) 2nd Edition. Springer-Verlag. Berlin, p325-345.

Recknagel F, Kim B, Welk A. (2006) Unravelling and prediction of ecosystem behaviours of Lake Soyang (South Korea) in response to changing seasons and management by means of artificial neural networks. *Verh.Internat.Verein.Limnol.* 29 p 1497-1502

Recknagel F, Cao H, Kim B, Takamura N, Welk A. (2006) Unravelling and forecasting algal population dynamics in two lakes different in morphometry and eutrophication by neural and evolutionary computation. *Ecological Informatics,* 1 p 133-151.

Welk A, Recknagel F, Cao H, Chan W-S, Talib A. (2008) Rule-based agents for forecasting algal population dynamics in freshwater lakes discovered by hybrid evolutionary algorithms. *Ecological Informatics,* 3 p 46-54

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

It is well known that the Australian environment, particularly freshwater resources, is currently under extreme pressure. Not the least of which, are the drinking water reservoirs, which provide potable water to households and industry. In many regions, the current drought is considered one of the worst on record and consequently low storage levels are being experienced in many Australian reservoirs (National Climate Centre 2007).
The impacts of drought on water quality in reservoirs can be many and varied. Low storage water levels in South Australian reservoirs have been known to create or compound the following: increased turbidity, increased metal levels (particularly Manganese and Iron), increased phosphorous concentrations, increased dissolved organic carbon, and most relevant to this study, increased cyanobacteria presence (CRC – Drought and Water Quality Report 2005).

Under normal circumstances, algal blooms are considered a serious water quality issue but during drought conditions they are even more frequent, intense and problematic. Depending on the nature of the bloom and the species involved, blooms can greatly increase water treatment efforts and costs, by drastically degrading water quality. Cyanobacteria blooms are of particular concern in drinking water reservoirs due to the nature and competitive advantages of this algal group that can enable a bloom to rapidly get out of control resulting in water quality degradation from taste and odour compounds, filaments clogging pipes and filters and toxic products among other things.

Considering the restricted volumes of raw drinking water currently available in South Australia, it cannot be afforded to have reservoirs become temporarily unusable and be taken offline, as there is already too much pressure on reservoirs to keep up with demand in their normal service area. Consequently, at this time it is more important than ever to be able to forecast potential water quality issues to a reasonable level of accuracy.

With some research suggesting that climate change will lead to 25% less precipitation in southern Australia by the year 2050, compounded by increased demand due to population growth, water resources will be under unprecedented stress (Ragab & Prudhomme 2002). Not only will there be less rainfall, but also increased solar radiation and higher temperatures, which stimulate algal

growth. It is clear that water management must consider new approaches to combat these water quality issues.

Current South Australian practice is to wait for evidence of substantial algal population levels before responding with undesirable and controversial CuSO4 dosing – and at this point, it is for damage control rather than prevention of a bloom event. Ideally, optimal management of algal populations would be proactive rather than reactive – and aim to avert such water quality degradation. However, reliable short-term forecasts are required to achieve such a management style. Computational and ecological modelling is a way to provide such forecasts.

The value of modelling in environmental management has been recognised for some time. Realistic and well validated models are able to predict future states or behaviour of dynamic systems, elicit interactions between components of the system and provide data and information that traditional field or laboratory techniques could not produce (Howard 1997; Whigham & Fogel 2003). The understanding of a system and its stressors that can be gained from a model, provides essential input to management and political decisions regarding actions to be taken in response to observed or expected conditions (Jorgensen 1994). Modelling is considered essential for any operational enterprise (Steel 1997) and Ferguson (1997) highlights the need for scenario analysis in lake management, stating that when a range of control options exist, it is necessary to have a predictive capability to enable the options being presented to be explored, especially for demonstrating the likely outcome to those who invest in the chosen option. Modelling is able to provide this service, as well as analysing past conditions and management.

In this study, complex time-series data from two South Australian drinking water reservoirs will be utilised in conjunction with various computational modelling methods for the ordination, clustering and forecasting of complex ecological data. Computational modelling is used in a wide range of fields from biology to finance, and basically involves modelling real world occurrences or problems on a computer, to gain better understanding and develop solutions.

More specifically, this study belongs to the emerging and rapidly developing field of Ecological Informatics, which aims to analyse, interpret, forecast and manage complex ecological data through the design and application of computational techniques (Recknagel 2003). Data driven techniques, such as those to be used in this study, are quite valuable in the field of ecology where there may be large and complex data sets available for analysis, but the system or occurrence

under study is not completely understood. Such techniques can utilise and justify long-term data collection and monitoring e.g. of water bodies.

To best manage algal dynamics and maximise bloom prevention, the answer cannot be prediction and early warning alone. A thorough understanding of causal relationships between physical, chemical and biological factors is imperative so that conditions promoting blooms can be avoided. Therefore the study uses several techniques with different focuses and outcomes to achieve not only forecasting of algal dynamics but to also offer insight into factors driving and controlling them.

Techniques to be used throughout this study are: Kohonen artificial neural networks (KANN) for data categorisation and the discovery of patterns and relationships, recurrent supervised artificial neural networks (RANN) for knowledge discovery and forecasting of algal dynamics and hybrid evolutionary algorithms (HEA) for rule-set discovery and optimisation for forecasting algal dynamics. These methods are combined to provide an integrated approach to the analysis of algal populations, including interactions within the algal community and with other water quality factors, which results in improved understanding and forecasting of algal dynamics.

Contributions of the study

The ultimate aim of this research is to provide forecasting tools for algal dynamics in reservoirs that are compatible with online real-time monitoring data, thus enabling real-time forecasting. To achieve this several general hypotheses must be investigated:

- KANN can reveal important relationships between physical and chemical water quality factors and algal dynamics, thus explaining seasonality and succession, by classification and clustering of temporal patterns.
- RANN can be used to forecast algal dynamics, whilst also providing sensitivity analysis to quantitatively describe underlying relationships.
- HEA can discover and extract rule-sets from time series water quality data that can be used for explanation and reasonably accurate prediction of algal dynamics.

During the investigation of these hypotheses, several novel approaches will be explored including:
KANN for the analysis of management regime impacts, KANN for the analysis of conditions corresponding with specific algal groups, integrated use of KANN and RANN for qualitative and quantitative description of relationships, and HEA rule-set extraction for same category lakes and linked stringent testing procedure.
Ultimately, the questions of most importance which are addressed by this research project are:

Is it possible to train a model to forecast algal dynamics using only electronically measurable data, so it could be compatible with a real-time monitoring/forecasting situation?

What inputs are used to train a model is dependent upon many things including availability, purpose of the model etc. For many models during this project, as many inputs as possible will be used, as the purpose is to look at relationships between input and output variables, not solely forecasting. However, if a model is to be used in a real-time situation it must only rely on inputs that can be monitored in real time. At the time of the study, accurate nutrient probes are not available for data loggers. Nutrient concentrations have traditionally been considered important inputs into algal population models, as algal abundance is usually very related to nutrient levels in the water – so can an accurate model be created that did not rely on nutrient concentration data but only electronically measurable input variables?

Is it possible to develop forecasting rule-sets for algal populations that can be generic agents for particular lake ecosystem categories?

The suitability of an ad-hoc or generic forecasting model is dependent upon the intended use and application, and each has pros and cons. Generic models reduce the labour required to develop and maintain models to be applied at several different sites and can be used for lakes with insufficient historical data to develop an ad-hoc model. They also allow comparative assessments between same category lakes to investigate specific algal population dynamics at varying habitat and water quality conditions within the same lake category. Ad-hoc or lake specific models often provide improved predictive results when compared to generic models applied to the same test site. Generality often compromises predictive accuracy; having been trained with many very different situations and conditions, clear patterns and relationships can be very difficult to observe and learn. The nature and interactions between factors driving algal

dynamics are extremely complex and the significance of each parameter is often catchment dependent. Therefore, similar conditions can have somewhat contradictory effects in different locations (Creagh 1992). To reduce the confusion, increase predictive ability and yet still provide a model that can be applied to a large cross-section of water bodies, could a model be developed that was generic for a particular set of lake ecosystem characteristics? Lake ecosystem categories defined by circulation type and trophic state, as suggested by Recknagel et al. (2006), will be used for this purpose.

Can these forecasting tools be implemented in a real-time environment to provide algal population forecasts up to a week in advance?

Throughout the study, the forecasting models will be tested on historical data that is kept separate from the training data, thus the testing data is unknown to the model and the model cannot simply replicate the patterns. This is a simulated real-time situation. Whilst there is merit in the development and testing of models in such a manner for research into model applications and underlying relationships driving algal dynamics, ultimately forecasting models are designed for practical use in a real-life environment, not a simulated one, and therefore must be tested in a real-life situation.

2. LITERATURE REVIEW

2.1 Modelling of ecological systems and processes

Models have long been used as simplified surrogates of real systems or problems, to provide improved understanding or solutions. Models can be used to address questions that often cannot be answered solely by experiments or observations, or to conduct experiments that are too expensive or impractical in the field. Basically, ecological modelling involves systems analysis and simulation in ecology and natural resource management. Models of virtually every possible type of ecological interaction have been developed (competition, parasitism, disease, mutualism, plant-herbivore interactions etc). Ecological models have two major aims: to provide insight into the structure and functioning of ecological systems; and to provide predictions about future prospects of particular populations, communities, or ecosystems under changing environmental, climate and management conditions. Some models only simulate the density of organisms, treating all organisms of any species as identical, also called mass action models. At the other extreme, the movement and fate of each individual organism may be followed in a complicated computer simulation by means of individual based models (McGraw & Hill 2005). The application of models to freshwater systems also has a long history with classical applications of the Lotka-Volterra predator-prey model (Parker 1968) for phytoplankton and zooplankton interactions and the lake eutrophication model by Vollenweider (1968).

As ecological research has grown in sophistication, models are increasingly used as decision support tools for management and policy-makers. The value of modelling in environmental management has been recognised for some time. Accurate and well validated models are able to predict future states or behaviour of dynamic systems, elicit interactions between components of the system and provide data and information that traditional field or laboratory techniques could not produce (Howard 1997; Whigham & Fogel 2003). The understanding of a system and its stressors that is gained from a model, provides essential input to management and political decisions regarding actions to be taken in response to observed or expected conditions (Jorgensen 1994). Modelling is considered essential for any operational enterprise (Steel 1997) and Ferguson (1997) highlights the need for scenario analysis in lake management, stating that when a range of control options exist, it is necessary to have a predictive capability to enable the

options being presented to be explored, especially for demonstrating the likely outcome to those who invest in the chosen option. Modelling is able to provide this service, as well as analysing past conditions and management.

Most ecological modelling applications can be broadly classified into one of 3 categories: the statistical approach, the differential equation approach and the computational approach (Recknagel 2003). This research focuses on only the computational approach that belongs to the emerging and rapidly developing field of Ecological Informatics.

2.1.1 Ecological Informatics

Recknagel (2003) defines Ecological Informatics as an interdisciplinary framework for the processing, archival, analysis and synthesis of ecological data by advanced computational technology. It aims to analyse, interpret, forecast and manage complex ecological data and systems through the design and application of computational techniques. Biologically inspired computation techniques, such as fuzzy logic, adaptive agents, artificial neural networks and evolutionary algorithms are a focus of Ecological Informatics. The two latter methods are used throughout this research and such techniques utilise and justify long-term data collection and monitoring of ecological systems.

Artificial Neural Networks (ANNs) and Evolutionary Algorithms (EAs) are both data driven and machine learning techniques, meaning that patterns and relationships are extracted from within the ecological data sets, where *a priori* knowledge is not required, and are able to be memorised and recalled. Data driven techniques, such as those to be used in this study, are quite valuable in the field of ecology where there may be large and complex data sets available for analysis, but the system or occurrence under study is not completely understood. Machine learning techniques have demonstrated their capacity to overcome many of the restrictions faced by conventional modelling approaches in terms of generality, realism and accuracy (Recknagel 2001). Compared with statistical models they handle non-linearity and high complexity well, and also provide valuable insights into data clustering and input–output relationships. Compared with deterministic models they can predict and elucidate short-term events and be combined with deterministic models to evolve ecosystem structures (Whigham & Recknagel 2001). Sometimes machine learning methods offer modelling and analysis options that cannot be provided by any other paradigm (Fielding 1999b). The methods used within the study can also be classified as

data mining techniques, which search for valuable information in large volumes of data, aiming to find patterns that can give accurate answers to future states (Weiss & Indurkhya 1998).

2.2 Artificial Neural Networks (ANNs)

Artificial Neural Networks (ANNs) are computer programs designed for information processing. The original inspiration for the technique was the central nervous system, particularly the biological neurons in the human brain. ANNs mimic the connectivity and function of the human brain and, similarly, have the ability to learn and be trained from the data presented to them (Freeman 2000; Lek *et al.* 2000). In a neural network model, simple neurons (alternatively called nodes, processing elements, PEs or units) are connected together to form a network, hence the term "neural network." The neurons are tightly interconnected and organised into different layers. The input layer receives the input; the output layer produces the final output and usually one or more hidden layers are in between the two.

Many studies have discovered the advantages and superiority of ANNs compared to conventional methods and examples include studies by Brey (1996), Lek et al. (1996), Paruelo and Tomasel (1997), Scardi and Harding Jr (1999), Karul and Soyupak (2006) and Jeong et al. (2007). Advantages of ANNs over conventional methods include:

- Use of ANNs requires less expert knowledge of the method and of the underlying problem being investigated, and are thus suitable for use by more people (Boddy & Morris 1999).
- Unlike statistical methods, the functional form of the model does not have to be specified *a priori* and model complexity can be altered easily by changing the architecture (Bowden 2003; French 1996).
- They have the ability to process large amounts of data at high speed (Bowden 2003).
- They are capable of dealing with incomplete, noisy or limited data and are robust (Burke 1991; Howard 1997; Kartam *et al.* 1997; Silvert & Baptist 2000).
- Unlike conventional statistical methods that are mainly limited to linear relationships and are inflexible, ANNs are parallel and distributed information extraction processors that have adaptive and self-organising properties that can handle non-linear data and high complexity (Chon et al. 2006).
- Non-linear models, such as ANNs, are best for analysis of temporal data (Chakraborty et al. 1992).

- ANNs can be used in situations where problem definition is difficult and the primary components governing the dynamics, and the interactions between them, are either unknown or poorly understood (Fielding 1999b; French 1996).
- ANNs do not require drastic simplification of a system in order to model it, as traditional techniques do (Harris 1996; Wilson & Recknagel 2003). Some modelling methods such as multiple regression assume linear relationships between variables, and in an effort to deal with non-linear systems, some variables are transformed. Despite these manipulations the results for use of these modelling methods in ecology remain disappointing.
- They can be quickly constructed using available data at a very low cost compared to developing conventional expert systems (Sung 1998).

Criticism of ANNs includes the lack of explicit model representation but mainly concerns their nature and categorisation as black box models. Black box models are models where a portion of the system is not seen or known by the user and much of the inner workings of the model are unknown. Several authors have commented that the lack of transparency and complicated internal workings of ANNs are disadvantageous and mean that one cannot know the exact flow of data and can only estimate the form of relations between variables and the output as they are given no indication of the processes involved in the modelled system (Fielding 1999b; Schleiter *et al.* 2006; Whitehead *et al.* 1997). However, this criticism is waning as new techniques are developed which allow the analysis of relationships captured by the trained model (Ball et al. 2000). Analysis of network weights has been used to gain some information but in an indirect fashion, and sensitivity analyses have been used successfully to gain further information and understanding from ANN models (Boddy & Morris 1999).

In ecological modelling applications, revealing the relationship between input and output variables is a very important aspect of the research. In the case of aquatic systems, knowledge of the influence of physical, chemical and biological water quality parameters on algal dynamics is desired so the system can be understood and managed as best as possible, and many authors have successfully contributed to this using different forms of sensitivity analysis (Goethals *et al.* 2006; Jeong *et al.* 2001; Recknagel *et al.* 2006c; Scardi 2000). Further explanation of the particular type of sensitivity analysis used throughout this study is in the methods section.

ANNs typically start out with randomised weights for all the connections between their nodes. This means that they do not 'know' anything and must be trained. Basically, there are two methods for training an ANN, giving rise to two broad categories: Supervised (SNN) and Non-supervised (NSNN).

With SNN, the training of the network or model is guided by many examples and patterns from cross-sectional or time-series data for a particular problem. The patterns are presented to the SNN with an associated observed output that the network can use to guide its predicted output (thus the learning is supervised), and after numerous repetitions of this during training, the network learns to recognise the pattern. Furthermore, the networks are usually able to generalise and will give suitable outputs for patterns that are similar, but not necessarily the same as the patterns the network was trained with (Boddy & Morris 1999). SNN are frequently used for classification and prediction of ecological variables.

NSNN use unsupervised learning and therefore do not focus on measured output data during training but classify the data by recognising different patterns. On most occasions NSNNs have been used for ordination and visualisation of complex ecological data. SNN and NSNN are discussed in more detail in the following sections.

2.2.1 Supervised Artificial Neural Networks (SNN)

The typical structure of a supervised neural network consists of an input layer, hidden layer/s and an output layer. The input layer contains measured data from external inputs such as nutrient concentration, density and composition of phytoplankton. The output layer also consists of measured data such as Chl-*a* (Recknagel 1997). The neural network then determines weighted connections between input and output layers utilising interconnected neurons, located in the hidden layer/s. The neurons feed a non-linear function, such as the sigmoid function, with the sum of their inputs in either feed forward or feedback (recurrent) mode. After this, the resulting value of the neuron is multiplied by a weighting factor (Rogers & Vemuri 1994). Therefore, each neuron has an individual weight parameter for each input-output connection. SNNs are typically initialised with randomised weights for all the connections between their nodes. During training, where the aim is to minimise the output error with regard to the known desired output, the neural network adjusts the interconnecting weights by means of a training algorithm. Once the interconnecting weights are optimised in training, they may remain fixed and the SNN can then be used for predictions (Recknagel et al. 1998).

The category of SNN can be separated into two further groups: feedforward or feedback SNNs, where classification is reliant upon how data is processed through the network, specifically whether only external inputs are used or feedback inputs (or looped connections) are considered also. Fig. 1 shows the basic structure of the two types of SNNs.

Figure 1. Basic conceptual structure of the two types of supervised artificial neural network (SNN): a) supervised feedforward; b) supervised feedback (from Recknagel and Cao, 2007)

2.2.1.1 Supervised Feedforward Artificial Neural Network

Supervised Feedforward Artificial Neural Networks consist of connections that move from the input layer, through the hidden layer/s to the output layer only. Neurons from one layer of the Feedforward Network are connected to all neurons in the next layer, but do not include looped connections or feedbacks, thus making the supervised feedforward approach faster than the feedback approach (Lek et al. 2000). Some of the most popular SNNs belong to this group including the multi-layer perceptron (MLP) (Minski & Pappert 1969), often applied with the back-propagation training algorithm. Back-propagation, described and developed by Werbos (1974) and Rumelhardt (1986) amongst others, compares the SNN's output to the desired result and determines the error. It then adjusts the input weights appropriately to reduce the approximation error and repeats this until convergence is reached.

This technique has been employed prolifically in the past in many different fields as it can be used for both cross-sectional and time-series data. Useful applications in aquatic ecology include studies on macroinvertebrate and fish communities in streams (Hoang *et al*. 2003: Lek *et al*.

1996), phytoplankton production in estuaries (Scardi 1996; Scardi & Harding Jr 1999), and phytoplankton dynamics in lakes and rivers (Maier *et al.* 1998; Recknagel *et al.* 1997).
Of particular relevance to this study is research by Wilson (2004) aiming to determine whether a standardised, generic ANN model representation can be developed to achieve short-term forecasts of algal blooms in lakes and rivers. Mainly multi-layer perceptrons in conjunction with the back-propagation algorithm were used, and models were trained with bootstrap samples of data to improve prediction error and reduce the likelihood of overfitting. A sensitivity analysis through time approach was used where an input variable is swept over a range of values, while the values of the remaining inputs are blocked at each data set value in turn (Wilson 2004). It was found that the generic ANN model was largely driven by the lagged output variable, suggesting that previous algal numbers are greatly influential in determining current algal population levels. Results and contributions of this work are discussed further in this chapter, specifically section 2.4.2.

2.2.1.2 Supervised Feedback or Recurrent Artificial Neural Networks (RANN)

Supervised Feedback or Recurrent Artificial Neural Networks (RANN) (Pineda 1987) consider not only external inputs but also looped connections feeding back activations from the previous step, providing extra information and training. RANNs are modifications of the typical ANN in that the structure is similar, except that the system state for time *t* is calculated by considering the system state at time *t-1*. If the activations of neurons in the hidden layer embody the 'hidden' state of the system, the copied activations of the time *t-1* are considered as feedback inputs for the determination of weights of neurons at time *t*. Therefore the feature and strength of RANN lies in the fact that, when calculating the output for a given time, the network considers not only the environmental driving variables as input for an algal dynamics forecast, but also the activation weights from the time step before. Fig. 2 demonstrates the structure of RANN for the prediction of blue-green algae abundance and considers typical input variables.

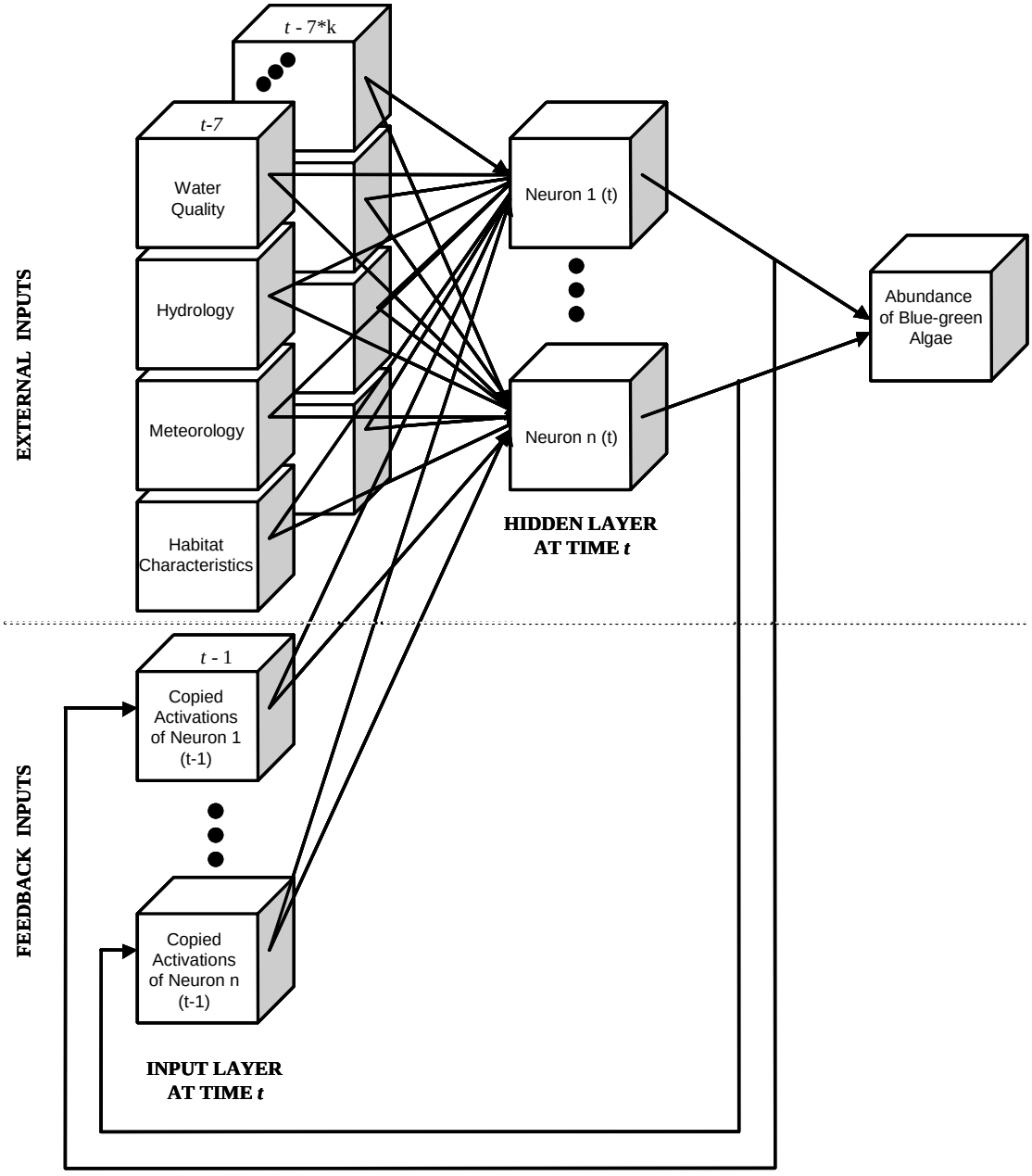


Figure 2. Architecture of a recurrent artificial neural network (RANN) for predicting abundance of blue green algae, using typical inputs

Using RANN, Jeong et al. (2001) predicted the timing and magnitude of Chl-*a* in the Nakdong River 3 days ahead with high accuracy. Four years of interpolated daily data (1995-1998) from numerous meteorological, physico-chemical and biological inputs was used to train the model, which was tested on one independent test year (1994). Only one hidden layer was used, with the hyperbolic tangent function selected for both hidden and output layers, and momentum set at 0.7; however training time and number of neurons was varied to find optimum results (see 3.5.2.1 for definitions of network architecture components). Visual assessment and linear regression

between observed and predicted values was used to evaluate model performance and the best RANN was subject to sensitivity analysis where input parameters were disturbed by +/- 1 to 2 standard deviations to examine interactions between Chl-*a* and the input variables. Results (see Fig. 3) showed that observed and predicted Chl-*a* values fitted well giving a r^2 value of 0.82, and importantly, the major summer peak was well forecast.

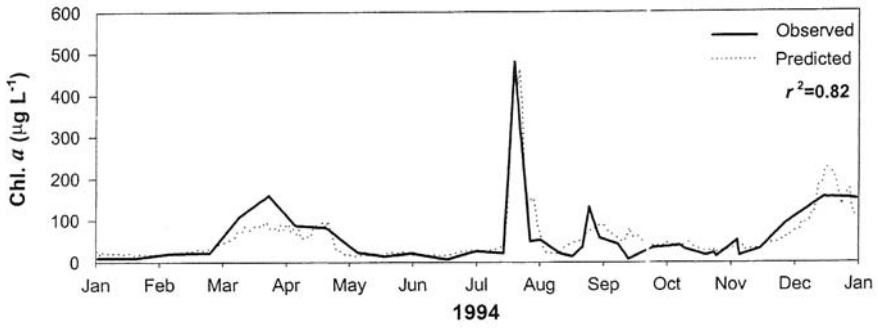


Figure 3. Results of RANN forecasting of Chl-*a* concentration 3 days in advance (Jeong et al, 2001)

Sensitivity analyses demonstrated that Chl-*a* was particularly responsive to chemical inputs such as pH and nutrients (Fig. 4), with further analysis of the impact of disturbed inputs over a range of values. For example, Fig. 5a shows that only changes in solar radiation in the lower range of observed values elicited an increase in Chl-*a* concentrations, whereas higher levels of evaporation were obviously very stimulating for algal biomass. Fig. 5b demonstrates that both increases in either pH and DO correspond with increasing Chl-*a* concentrations.

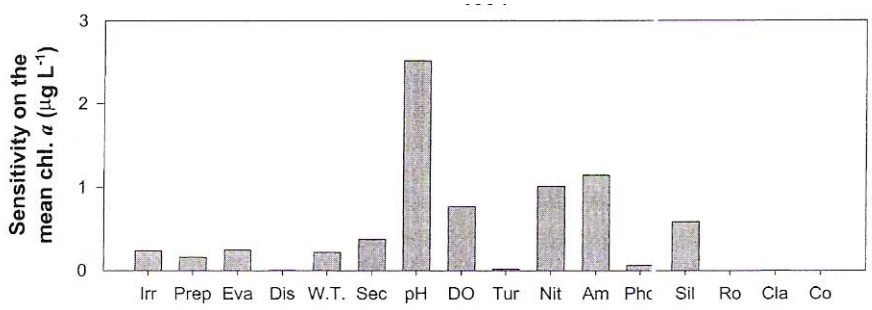


Figure 4. Results of sensitivity analysis with disturbance of +/- 1 standard deviation to the input data (Jeong et al, 2001)

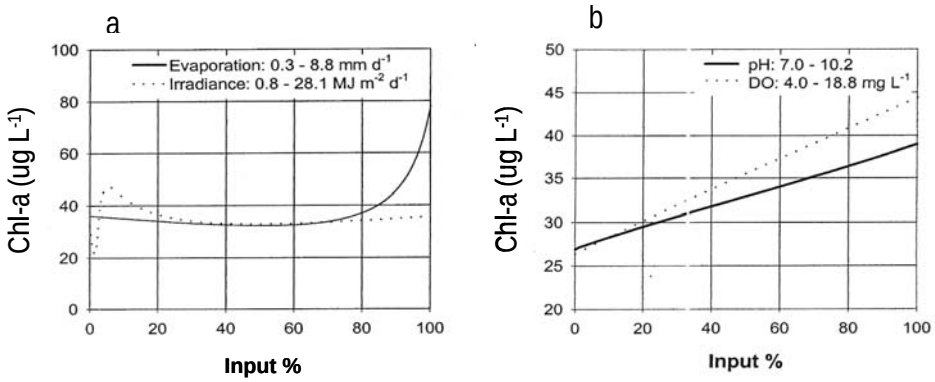


Figure 5. Results of sensitivity analysis with wide-ranged disturbance of +/- 2 standard deviations to the input data: a) evaporation and irradiance; b) pH and DO (Jeong et al. 2001)

In another study by Jeong et al. (2003) the target output was refined to species level, with RANN successfully used to predict the seasonal succession between the nuisance blue-green algae species *Microcystis aeruginosa* and the diatom species *Stephanodiscus hantzschii* in the Nakdong River. Applying the same approach as the previous study again yielded good results, with both *Microcystis aeruginosa* and *Stephanodiscus hantzschii* being accurately forecast 4 days in advance, giving r^2 values of 0.68 and 0.73 respectively. Again sensitivity analysis was used to examine input-output relationships and allowed comparisons between the species. These studies were successful in not only forecasting algal growth events in advance, but also revealing information that influenced the algal dynamics. Accurate algal population predictions from a useful forecasting horizon were achieved and sensitivity analyses were demonstrated to be valuable for improving understanding of the system and testing hypotheses.

Another relevant study is by Walter et al. (2001) who used RANN to predict eutrophication effects, including algal blooms, in the Burrinjuck Reservoir. The model was reasonably accurate in the prediction of timing and magnitudes of Chl-*a* for seven days ahead, indicating its potential to provide early warning for tactical control of algal blooms in freshwater lakes. A sensitivity analysis during this study interestingly revealed that algal abundance in the Burrinjuck Reservoir was not only driven by inorganic nutrients, water transparency and temperature but also, to a large extent, by hydrological characteristics such as water depth that is subject to high seasonal variations in this water body. Talib et al., (2005) used RANN to successfully forecast the succession of the algal groups *Oscillatoria* and *Scenedesmus* in two Dutch lakes, 5 days in advance. Finally, this author used RANN to successfully forecast Chl-*a* and specific algal functional groups in Lake Soyang, a temperate stratified lake (Welk 2003). In summary, RANNs have proved to be very powerful for time-series modelling of Chl-*a* and specific algal abundance.

The success of these studies, particularly Jeong et al. 2001 and 2003, were influential in determining the ANN approach used for this research project. Although many researchers have produced useful results with feedforward applications, RANN has proven to be particularly powerful for time-series modelling. Successful RANN studies have been able to achieve the right balance between accurate forecasting and knowledge discovery. However, depending on the intended use of the models, the lack of explicit model representation could be seen as a limitation of RANN, making the model not easily shared or applicable to other situations.

As with the successful studies described above, initial RANN experiments for this project will use daily-interpolated data to train the RANN and independent year/s will be used for testing, by means of the leave one out method. The network architecture will be as described above, with optimal training time and number of nodes to be determined for each model. Sensitivity analyses looking at sensitivity about the mean and specific input-output sensitivity will be carried out as these techniques have yielded much useful information in previous studies. Differences in approach will include a reduction in input variables used to create the models, as both studies discussed used a large number of inputs that are not often consistently available for many water bodies; as well as the inclusion of RMSE as a component of model assessment along with visual examination and r^2 values.

2.2.2 Non-Supervised Artificial Neural Networks (NSNN)

NSNN employ unsupervised learning to discover and visualise patterns in data. NSNN do not focus on a desired output but classify complex data by recognising specific patterns or similarities, therefore the internal organisation of the network is solely dependent on the input stimulus. NSNN are more often a two-layered network with the first layer being the input layer, with one node for each parameter. The input data then pass to every node in the second layer, the two-dimensional grid of nodes. Each node has a set of weights on its inputs, which are modified during the learning phase (Boddy & Morris 1999). The outcome of NSNN is the ordination and clustering of the input data, and this method is capable of discovering significant patterns or features in the input data in a way that is comparable to the traditional Principal Component Analysis (PCA) or hierachial clustering analysis (Jongman et al. 1987), whilst also having the ability to cope with non-linearities (Boddy & Morris 1999). The features or patterns found in the input data are expressed by Euclidean distances, which are calculated between inputs and can be visualised as a unified distance matrix (U-matrix) or as a partitioned map (*K*-

means). The U-matrix visualises the relative distances between neighbouring data in the input data. The *k*-means algorithm simply partitions the input data into a specified number of clusters based on the U-matrix (Recknagel et al. 2006b).

The most popular type of NSNN is the Kohonen Artificial Neural Network (KANN), which has been used during this study.

2.2.2.1 Kohonen Artificial Neural Network (KANN)

KANN (Kohonen 1982, 1984), also know as self-organising maps (SOMs), can be used to give a comprehensive view of the patterns within a data set. KANN allows great insight into the way data are distributed and organised in n-dimensional space by mapping the incoming patterns. This information extracted from multi-dimensional data is mapped into reduced dimensional (usually 2 dimensional) space and visualised in a simplistic fashion, as patterns or clusters arranged on a hexagonal lattice or grid (Giraudel & Lek 2006). KANN use competitive or self-organising learning, where neighbouring cells within the network interact and adaptively develop into detectors of a specific input pattern (Bowden et al. 2006). This adaptive nature is what gives KANN its similarity to processes within the brain and, hence, its place as an ANN.

KANN have two layers, the input layer and the Kohonen layer (see Fig. 6). The input and Kohonen layers are fully connected. Neurons in the Kohonen layer are not connected, and measure the distance of their weights to the input pattern. Competitive learning allows the neurons to compete with each other in Euclidean map space (Jeong & Joo 2003). Using this principle of competitive learning, the elements compete to respond to an input stimulus, based on distance from weights to the input pattern, and the winner, which has the smallest distance, adapts itself to respond more strongly to that input stimulus (Maier 1995). After training, data is clustered referring to certain criteria such as seasons of the year or nutrient ranges according to a calculated U-matrix, and the result of this is the relational patterning of the data set, which can be graphically visualised and compared using component planes.

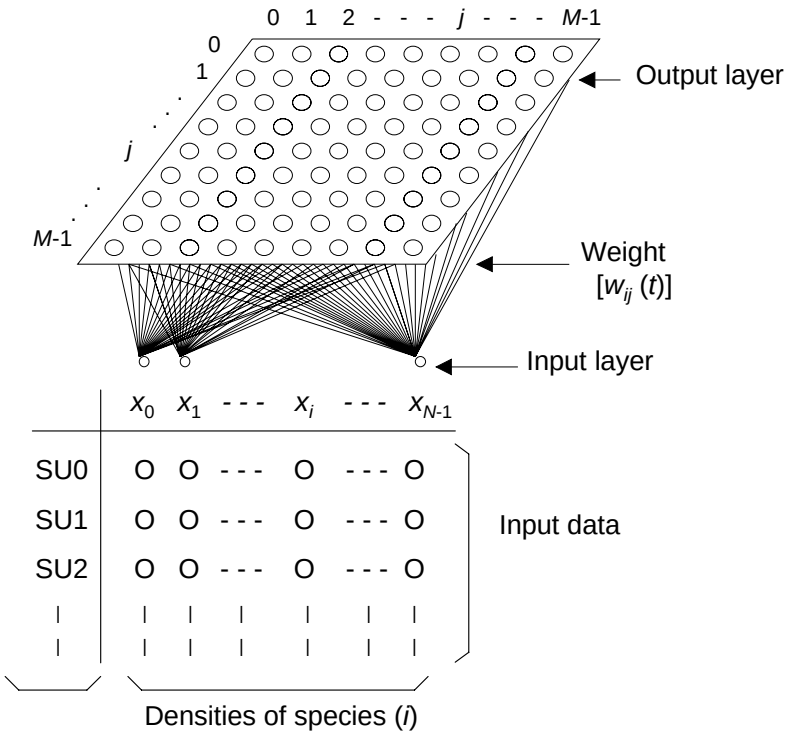


Figure 6 - Kohonen Artficial Neural Network for non-linear cluster analysis of ecological data. (From Chon et al. 1996)

KANN are a very adaptable tool and can be extremely useful for studying ecological communities. KANN have shown through numerous studies that they are capable of overcoming the restrictions of conventional multivariate statistics and have been proven to be a superior tool for successful non-linear ordination and clustering of ecological data (Giraudel & Lek 2006). Analysis and patternising of ecological community data has been successfully carried out in numerous studies (Chon *et al.* 1996; Foody 1999; Lee *et al.* 2007). KANN has been productively applied to many different areas of aquatic ecology including fish assemblages (Brosse *et al.* 2001; Kruk *et al.* 2007), benthic macroinvertebrate communities (Chon *et al.* 2000; Park *et al.* 2006; Song *et al.* 2006), and more relevantly, algal dynamics (Oh *et al.* 2007; Recknagel *et al.* 2006b). KANN can be used to make comparative assessment of periods of data, for example periods of different management regimes within a water body. Welk (2003) analysed chemical and biological differences coinciding with varying intensities of fish farming in Lake Soyang over a 10 year period, and later Recknagel et al. (2006b) used this method to examine three distinct management periods and facilitate a comparative analysis of two adjacent Dutch lakes regarding

short and long tern dynamics in response to bottom-up and top-down eutrophication control over 18 years. These two studies utilised whole year data interpolated to give daily values, whereas Chan et al. (2007) used 12 years of daily-interpolated data from only June to October to gain greater understanding of microcystin occurrence in Lake Suwa (Japan). The study was able to improve the understanding of the relationships between microcystin concentrations, *Microcystis* species abundance and rainfall intensity by revealing that total *Microcystis* abundance and extra-cellular microcystin concentrations are much higher in typical dry years than typical wet years. It also showed that high microcystin concentration in dry years was linked to the dominance of *Microcystis viridis*, whereas wet years were dominated by *Microcystis ichthyoblabe*. Other researchers have used KANN in conjunction with other modelling methods. Whigham (2005) combined KANN clustering and linear regression for a local modelling approach to algal dynamics forecasting, which resulted in improved forecasting of *Microcystis aeruginosa*. Recknagel et al. (2006c) combined KANN clustering results with input sensitivity graphs from RANN for improved understanding of habitat preferences and differences between the dominant genera in two lakes. This approach demonstrates relationships qualitatively by using KANN clustering and quantitatively by the graphs of the input sensitivity over a range of values obtained by RANN. The many successful studies using KANN have shown it to be a very useful tool with a wide range of possible applications.

2.3 Evolutionary Algorithms (EA)

Evolutionary computation is a bio-inspired machine learning method that employs principles from natural biological evolution, such as cross-over and mutation, to develop solutions to complex computational problems. It searches for suitable models or solutions to a problem by means of genetic operators and the principle of survival of the fittest. The premise behind this is that, as natural evolution has created complex systems and successfully discovered novel solutions for difficult and extremely complex problems, it should be an effective and appropriate approach to modelling of ecological systems (Whigham & Fogel 2006).

Evolutionary methods use an algorithmic statement of evolution to evolve solutions to the target problem. The general process of evolutionary algorithm use is detailed below and demonstrated in Fig. 7. It begins with a randomly generated collection of individuals (a population) where each individual is a possible solution to the problem (Bobbin & Recknagel 2003). If the data refers to a

well-understood phenomenon, it can be helpful to begin the search for rules from statements that have been proven by past research. These starting solutions are the basis for the search for effective predictors (Jeffers 1999). The algorithm, after testing the efficacy of the starting solutions (in this study the solutions are in the form of rules), will then modify the rules to improve their ability to predict the defined target. Variability is introduced to the individuals by cross-over and mutation, and evolution is achieved by these information exchanges by solutions in a population, which creates new solutions or modifications of existing solutions (Gibbs 2004). The performance of the individuals (their fitness) is evaluated based on available information (in this study RMSE is used), then ranked and the best individuals can be used to form a new population (generation) for further testing. This selection pressure generally means that individuals in the new generation will produce better solutions. This process is repeated many times until a satisfactory rule or rule set has been found from the search space (Bobbin & Recknagel 2003; Fielding 1999b; Flood & Kartam 1997).

NOTE: This figure is included on page 20 of the print copy of the thesis held in the University of Adelaide Library.

Figure 7- Summary of the process of evolutionary computing (from Morral, 2006)

Evolutionary Algorithms (Holland 1975) (EAs) can be applied to a broad range of problems and are especially valuable in finding solutions to problems that are difficult to solve using well-defined deterministic strategies or those that can be posed as a search for a set of particular values, conditions or structures (Whigham & Fogel 2006). Most often, EAs are used when there is an incomplete knowledge of the problem or how to solve it, and when the data set is very large (a

large search space), and solution searching using traditional optimisation methods is inefficient and ineffective (Morrall 2003). EAs have been established as a productive method for searching complex non-linear adaptive topographies to provide near-optimal solutions in real time (Whigham & Fogel 2006). They have the advantage of being able to incorporate quantitative variables and qualitative attributes into the same set of rules (Jeffers 1999).

EAs can be used for optimisation, equation and rule discovery, pattern searching and discovery, and construction of artificial systems among other things (Morrall 2003; Whigham & Fogel 2006). Bäck (1997) and Fogel (1998 and 2000) provide useful summaries of the history and applications of evolutionary computation and EAs. Of the many possible EA applications, of particular relevance here is discovery of models, in the form of equations or rule-sets, used to describe a natural system. The discovery of equations or rules by EA, unlike ANN, provides an explicit representation of the model. Whilst some methods and studies focus on the pre-processing of the data and the training of the models to get good predictive results, this ignores the opportunity to obtain knowledge from the learning algorithm. The model must have learnt some knowledge of patterns within the data in order to make predictions, but the representation of the knowledge is usually unavailable for inspection, as with ANNs (Bobbin & Recknagel 1999). However, evolutionary methods allow the development of models with more transparent knowledge representations and provide some explanation of how the model understands the system functionality. This can result in an improved understanding of the model predictions and behaviour. Further, it could potentially lead to greater comprehension of patterns and relationships existing within the data, which control particular occurrences such as algal blooms. Equation discovery by EAs aims to find an equation, usually a differential or difference equation, that describes the relationships between variables within a data set and enables modelling of the system. Predictive equation discovery has been applied to aquatic ecology quite successfully in numerous studies. Whigham and Recknagel (1999) used an evolutionary approach based on genetic programming to find several time series mathematical equations able to forecast the concentration of Chl-*a* in Lake Kasumigaura, Japan. Muttil and Lee (2005) also used a genetic programming to evolve an equation to forecast chlorophyll fluorescence in real-time, with the aim of preventing red tides in Hong Kong.

Jeong et al. (2003) modelled *Microcystis aeruginosa* bloom dynamics in the Nakdong River, comparing the equation discovery capabilities of multivariate linear regression (MLR) and the time-series optimisation genetic programming (TSOGP) system (Whigham & Keukelaar 2001).

Daily-interpolated data from 1995 to 1998 was used for equation discovery whilst 1994 data was used for model validation. Meteorological, hydrological, physico-chemical and biological data (a total of 19 variables) were available for inclusion in the predictive equation. The best performing equation was then subject to two types of sensitivity analysis, to help understand the dynamics of *Microcystis aeruginosa*. Results showed that the genetic programming approach produced a simpler equation (Eq.1), using only 4 of the available inputs that gave higher predictive accuracy than the equation discovered using MLR (Eq.2), which was complex and used all 19 input variables.

Eq.1. *Microcystis aeruginosa* (t+1)
= { 0.49293 x secchi depth(t) + 0.50707 x *Anabaena flos-aquae* (t) }
x { 0.83014 x evaporation(t) + 0.16986 x turbidity (t) }

Eq.2. *Microcystis aeruginosa*
= 47223.339 x rain – 0.009 x wind
+ 585821.152 x evaporation – 3.057 x 10^{-5} x discharge
- 0.0111 x secchi depth + 13054.426 x turbidity + 85020.205
x water temperature + 1102600.058 x pH
+ 167271.631 x dissolved oxygen – 0.0285 x nitrate + 2124.290
x ammonia – 0.001 x phosphate
+ 62798.045 x dissolved silica + 84.270 x Rotifera – 8.766 x Cladocera
- 0.0002 x Copepoda + 2.377 x *Anabaena flos-aquae*
- 1.326 x *Oscillatoria limosa* – 1.585 x *Stephanodiscus hantzschii*

The equations produced results giving r^2 values of 0.77 and 0.08 respectively. Clearly Eq.1, the evolutionary approach, provides the best equation which regard to both simplicity and predictive accuracy and was therefore chosen to be explored further by sensitivity analyses. The 'most influencing parameter (MIP)' sensitivity analysis on showed that *Anabaena flos-aquae* biovolume was the most influential variable, followed by turbidity and evaporation, with secchi depth shown to be unimportant. A 'sensitivity on wide-ranged disturbance (SWD)' analysis showed that *Anabaena flos-aquae* biovolume, turbidity and evaporation had positive linear relationships with the output *Microcystis aeruginosa*, whilst secchi depth appeared to have no relationship. The study demonstrated the superiority of evolutionary equation discovery over MLR and that the evolutionary approach is very suitable for modelling of this nature.

A recent study by Kim et al. (2007a) also used TSOGP to develop on equation to predict the time-series dynamics of the diatom *Stephanodiscus hantzschii* in the lower Nakdong River 3 days in advance. Not only was the equation successful in achieving useful forecasts with a test r^2 value of 0.78, but method also allowed the examination of the relationships between the 8 variables included in the equation (Andong dam discharge, Namkang dam storage, evaporation, water temperature, dissolved oxygen, pH, secchi transparency and silica). Four types of sensitivity analysis were used; MIP, SWD, sensitivity on time-series (STS) and sensitivity for simultaneous movement of parameters (SSMP). This extremely thorough analysis concluded that water temperature was the most sensitive parameter, followed by dissolved oxygen and secchi transparency. The study clearly demonstrated evolutionary computation as a useful approach for developing models that are capable of interpreting the complexity of ecosystem behaviour and providing accurate forecasts.

Although the equation discovery evolutionary approach has given positive results in many studies, rule discovery is a particularly promising branch of evolutionary algorithms for ecological modelling. Genetic Algorithm For Finding Existing Relationships (GAFFER) (South 1994), as described and demonstrated in Jeffers (1999), and Genetic Algorithm for Rule-set Production (GARP) (Stockwell 1992) are examples of discovering relationships in data sets and finding rules for numerical prediction and classification.

Whigham and Recknagel (1999) noted that a rule-based approach to forecasting algal dynamics may be more suitable than equation based forecasting, as rules allow different phases to be recognised and for different equations to be used for each phase. Jeffers (1999) states that they have the advantage of being able to incorporate quantitative variables and qualitative attributes into the same set of rules. Kim et al. (2007) compared two algorithms of evolutionary computation, an algebraic function model and a rule-based model, both designed to forecast *Microcystis aeruginosa* dynamics in the lower Nakdong River. Results showed that the rule-based model outperformed the algebraic function model with regard to predictive performance as it correctly forecast both timing and magnitude of the algal blooms, whereas the function model could only accurately forecast the timing of the events. Bobbin (2002) proposed a new self-adaptive, symbiotic model evolution framework (SASME) for evolving rule sets for learning problems, with much emphasis on knowledge discovery and being able to gain information from the model representation itself. The SASME method was unique in two main ways: 1) method – the evolutionary learning algorithm combined the real-valued optimisation power of self-adaptive evolutionary algorithms with a novel self-adaptive strategy for the evolution of discrete structures,

and 2) representation – the method evolved entire solutions to problems as rule sets with exception lists allowing for the explicit representation of default hierarchies (Bobbin 2002). The SASME framework was applied to produce predictive rules for algal abundance in lakes from measured data, and the resulting rule sets were interpreted to discover what they reveal about the causes of algal species succession in lakes. One example discovered a predictive rule for Chl-*a* in Lake Kasumigaura (see Fig. 8), using water quality data from 1984 to 1993. Results show that the method could forecast Chl-*a* levels to a reasonably accurate level with a RMSE of 35.76. Further, the model representation (Fig. 9) gave explanation of the relationship between water quality conditions and Chl-*a* levels.

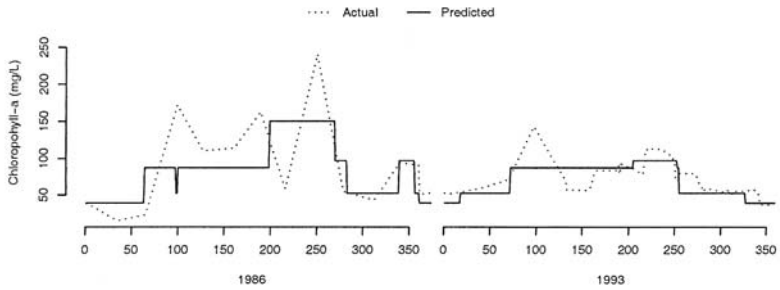
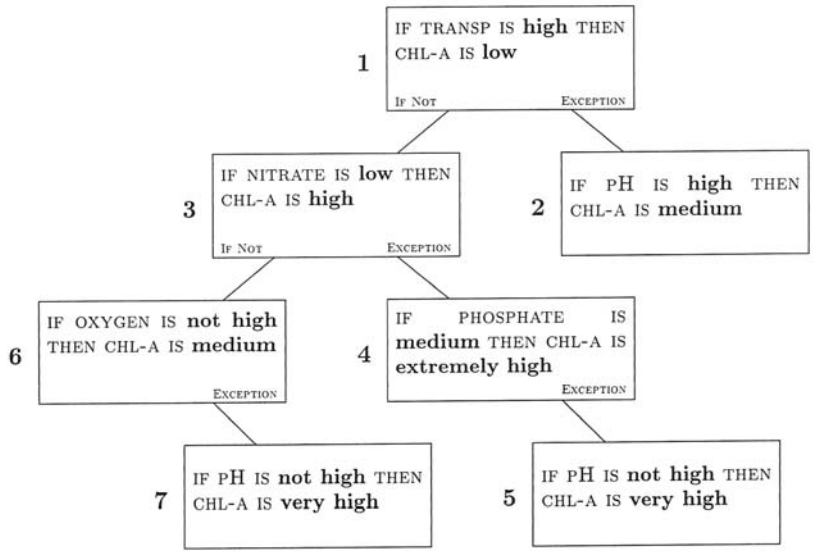


Figure 8. Results of Chl-*a* forecasting using a predictive rule-set obtained using the SASME framework (Bobbin, 2002)



Description	Value
Transparency is high	Transp > 93cm
pH is high	pH > 9.2
pH is not high	pH < 9.2
Nitrate is low	NO3 < 560µg/l
Phosphate is medium	25 < PO4 <= 168µg/l
Oxygen is not high	DO < 18mg/l
Chlorophyll-a is low	Chl-a = 44mg/l
Chlorophyll-a is medium	Chl-a = 60mg/l
Chlorophyll-a is high	Chl-a = 101mg/l
Chlorophyll-a is very high	Chl-a = 112mg/l
Chlorophyll-a is extremely high	Chl-a = 174mg/l

Figure 9. An evolved rule set for predicting algae (Bobbin, 2002)

Bobbin and Recknagel (2003) also discovered rules for the prediction of filamentous blue-green algae and *Microcystis* in Lake Kasumigaura, and found that information gained from the rule-sets was consistent with literature on conditions preferred for algal growth.

Due to the achievements of the previously discussed studies applying the rule discovery aspect of evolutionary algorithms to ecological data sets, rule-sets were chosen as the form of model representation for this project. Rules are a good form of model representation as they are easy to interpret and can be used independently of the system in which they were created (Whigham & Fogel 2006).

Whilst the examples above did discover predictive equations and rules for population dynamics, they did have some limitations. Only constant parameters were considered as part of simple rules rather than complex functions with multiple attributes, and parameters influencing the output

values were generated randomly rather than being simultaneously optimised during evolution. Although Whigham and Recknagel (2001) performed the hill climbing mutation for the optimisation of the random real numbers and Bobbin and Recknagel (2003) used the self-adapting algorithm to tune the parameters, both struggle with increasing rule complexity (Cao et al. 2006).

2.3.1 Hybrid Evolutionary Algorithms (HEA)

The Hybrid Evolutionary Algorithm (HEA) (Cao et al. 2004) used in this project differs from and improves on the previous EA approaches by using genetic programming to generate and optimise the structure of the rule set, and then using a general genetic algorithm to optimise the random parameters in the rule set (see Fig. 10). Rules discovered by HEA are in the form of IF-THEN-ELSE and their complexity can be controlled. Fitness of the solutions is assessed using the Root Mean Square Error (RMSE) (Cao et al. 2006). Sensitivity analyses, in a similar manner to that carried out in conjunction with RANN, are used to assess the influence of changes in input on the output variable. However, the sensitivity analysis is improved by allowing the examination of both the THEN and ELSE sections of the rule separately to give more specific information about different phases of algal growth.

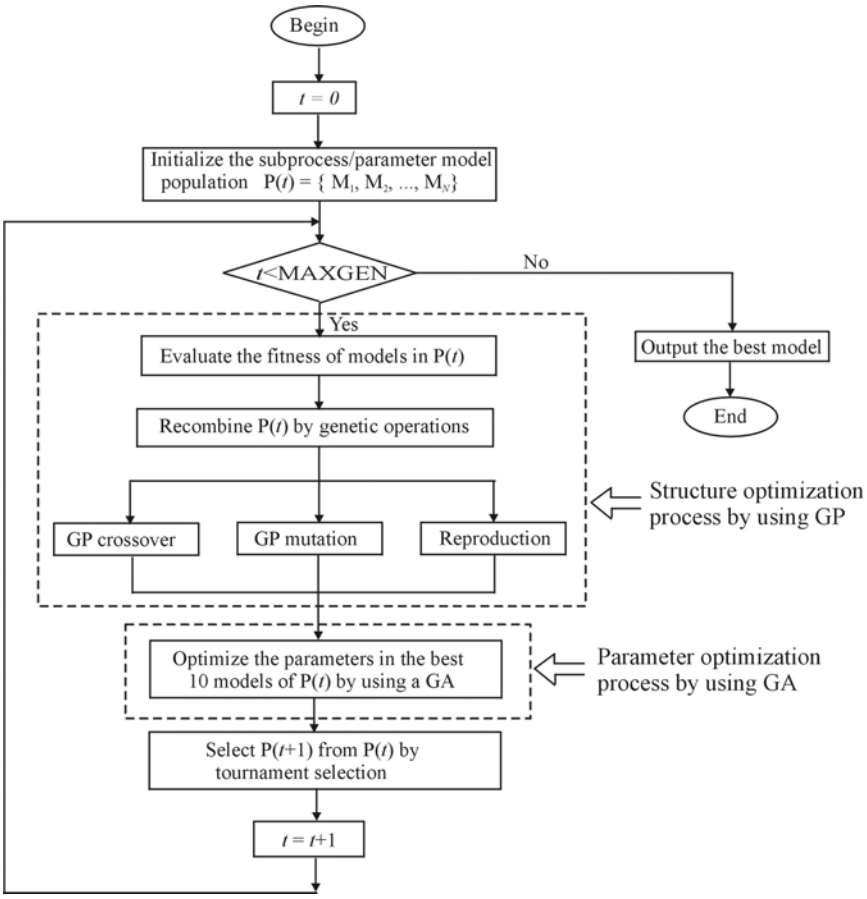


Figure 10- Flowchart of the hybrid evolutionary algorithm (from Cao et al. 2006)

A study by Cao et al. (2006b) successfully applied HEA to 7-day-ahead forecasting of seasonal abundances of blue-green algae and diatom populations in Lake Kasumigaura (Japan) and Lake Soyang (Korea), using 9 and 10 years of water quality data respectively, and each tested on one independent year of data. Resulting rule sets provided forecasts of dominant algal genera (*Microcystis* and *Cyclotella* in Lake Kasumigaura, and *Anabaena* and *Asterionella* in Lake Soyang) with relatively high accuracy (see Fig. 11), whilst also being explanatory with regard to relationships between physical and chemical input variables and the algal abundance output (Cao et al. 2006).

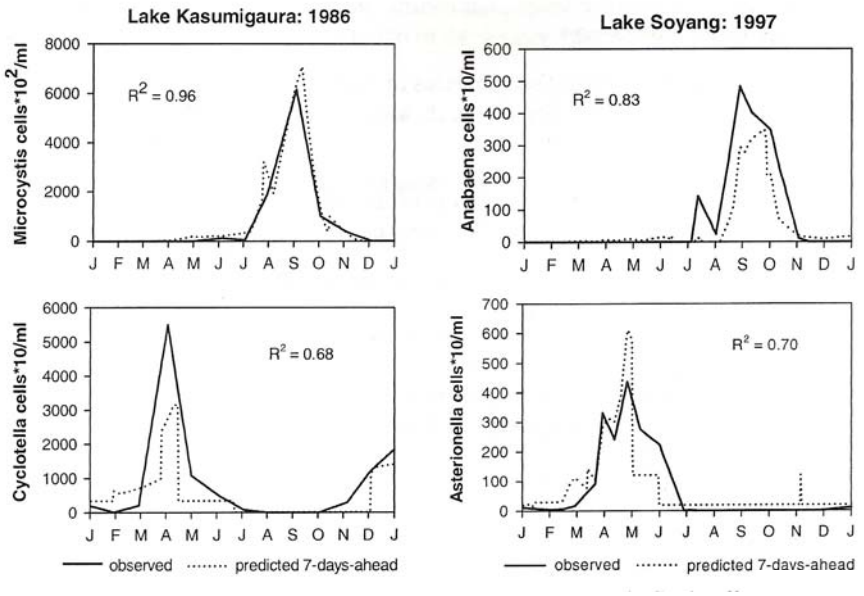


Figure 11. 7 day ahead forecasting of *Microcystis* and *Cyclotella* in Lake Kasumigaura (left column), and of *Anabaena* and *Asterionella* in Lake Soyang (right column) using a predictive rule set obtained by HEA (from Cao et al, 2006)

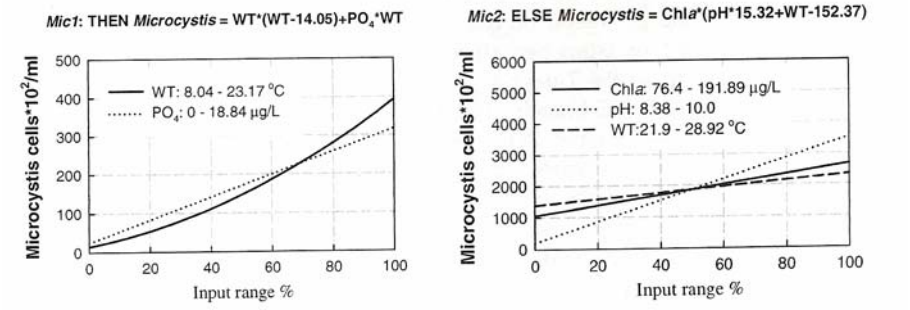


Figure 12. Sensitivity analysis with disturbance +/- standard deviation of input data for THEN (left) and ELSE (right) branches of predictive rule set for *Microcystis* (from (Cao et al, 2006)

The sensitivity analysis was able to show the relationships between the algal output and the particular variables in the THEN and ELSE branches separately. For example, Fig. 12 shows sensitivity analysis with disturbance (SWD) for both branches of the predictive rule set for *Microcystis*. Results show that the THEN branch is utilised for smaller *Microcystis* populations and shows positive relationships between the algae and increasing water temperature and PO_4 concentrations. The ELSE branch is used for larger Microcystis populations. The larger populations are shown to correlate with high water temperatures above 20ºC and pH levels over 8, which is supported by findings from both field observations and laboratory experiments (Reynolds 1984; Shapiro 1990).

Talib et al. (2005) applied HEA and RANN for the forecasting of phytoplankton abundance and succession in response to eutrophication control in two shallow Dutch lakes. This approach has also successfully been applied to algal dynamics in Nakdong River system (Cao *et al.* 2006a; Kim *et al.* 2007).

2.4 Summary of past modelling efforts at study sites

Various different types of models have been used over the past two decades to perform numerous tasks at either Myponga or Happy Valley reservoirs. A cross section of these applications is summarised below.

2.4.1 Differential equation based modelling of the Myponga Reservoir

In 1991 a report was written presenting the results of using the mathematical model DYRESM (Dynamic Reservoir Simulation Model) (Imberger et al. 1978) to simulate the performance of the three submersible mixers in operation at Myponga reservoir (Velzeboer et al. 1991). The model DYRESM is a one-dimensional numerical model for the prediction of temperature and salinity in small to medium sized lakes and reservoirs. It incorporates the layer concept, in which the reservoir is modelled as a system of numerous horizontal layers (Velzeboer et al. 1991). The aim of this application of the model was to determine the optimum position and orientation of the mixers for reservoir destratification by scenario analysis. The results from the use of the model determined the best timing and position for use of mechanical mixers, however, it was concluded that, even at these optimal conditions, although the water quality would improve slightly, the mixers only result in partial destratification of the reservoir and could not effectively control the growth of *Anabaena.*

The lake model SALMO (Simulation by means of an Analytical Lake MOdel) (Recknagel & Benndorf 1982) was applied to Myponga reservoir. SALMO is a process-based model that is designed to simulate food-web dynamics and nutrient cycles in lakes. It allows the simulation of both the epilimnion and hypolimnion during stratification due to its two-layered structure (Recknagel 1989). In this case it was used to simulate the annual dynamics of limnological variables including phyto- and zooplankton, orthophosphate, nitrate, detritus and dissolved oxygen. Manipulation of the input data allowed scenario analysis for the application of a range of

different management options to the reservoir. The scenarios examined were: water colour reduction, 50% phosphate elimination from the in-flowing water, biomanipulation, artificial destratification and a combination of biomanipulation and destratification. Besides a drastic reduction of external phosphorus loading, the combination of biomanipulation and destratification was found to be the best management option for the reservoir. The biomanipulation successfully kept algal populations under control in spring, while the destratification greatly reduced the summer peak to the lowest simulated for spring and summer under any management regime (Smalley 1996). However, in Australia biomanipulation does not always fulfil its potential as the common native zooplankton are not big consumers of cyanobacteria (Herath 1997). In the case of Myponga reservoir, the study concluded that destratification by itself would still prove to be of great advantage (Smalley 1996).

2.4.2 Artificial neural network based modelling of the Myponga Reservoir

Wilson (2004) focused on supervised feedforward artificial neural networks trained to make short term forecasts of algal blooms in lakes and reservoirs, to aid decision making for the operational management of eutrophication. The study aimed to address the question of whether a standardised, generic ANN model representation can be developed to achieve the above goal. Interestingly, the ANN model inputs were represented as summary statistics of sliding time windows, which increased the compatibility of typical time-series ANN model structures with data sets containing uneven sampling frequency and missing data. Models were trained with bootstrap samples of data to improve prediction error and reduce the likelihood of overfitting, with training data coming from six different sites around the world, one being Myponga reservoir. The model validation approach was a blocked leave-*k*-out cross validation, where the data is divided equally into *k* smaller time series or subsamples which are each used in turn as a validation set whilst the remainder are combined for a training set. Blocked leave-*k*-out was combined with bagging by taking bootstrap samples of the training data and running the entire leave-*k*-out procedure a number of times to get a reasonable bootstrap sample of model predictions on validation data (Wilson 2004). A sensitivity analysis through time approach was used where an input variable is swept over a range of values, while the values of the remaining inputs are blocked at each data set value in turn (Wilson 2004). Models were developed for 7 and 14 day ahead forecasts, with best results found to be achieved from 14 day ahead models. Lake specific

models were also created, using all available inputs, to compare generic and lake specific model performance.

The generic ANN model structure was applied to forecast Chl-*a* concentration and the three most abundant species for the lake. The input variables used in the generic structure (Fig. 13) were water temperature, secchi depth, phosphorus, nitrogen and a lagged input of the output variable (note: secchi depth was not available for the application to Myponga reservoir). The generic model predicting Chl-*a* achieved a reasonable performance with an r^2 value of 0.436 and visual inspection showed close correspondence between observed and predicted values (Wilson 2004). Every major bloom event was predicted, however there was generally a slight delay and underestimation of peak events. Models of *Ankistrodesmus spp.* and *Dictyosphaerium spp.* performed poorly (r^2 values of 0.118 and 0.029 respectively), with the *Scenedesmus spp.* model performing slightly better (r^2 =0.199). The main finding from the sensitivity analyses was that the generic ANN model was largely driven by the lagged output variable, suggesting that previous algal numbers are greatly influential in determining current algal population levels. Overall, results from the study showed that site and output specific input layers gave better results than generic models and that ANNs generalise better over short term as opposed to long term time scales.

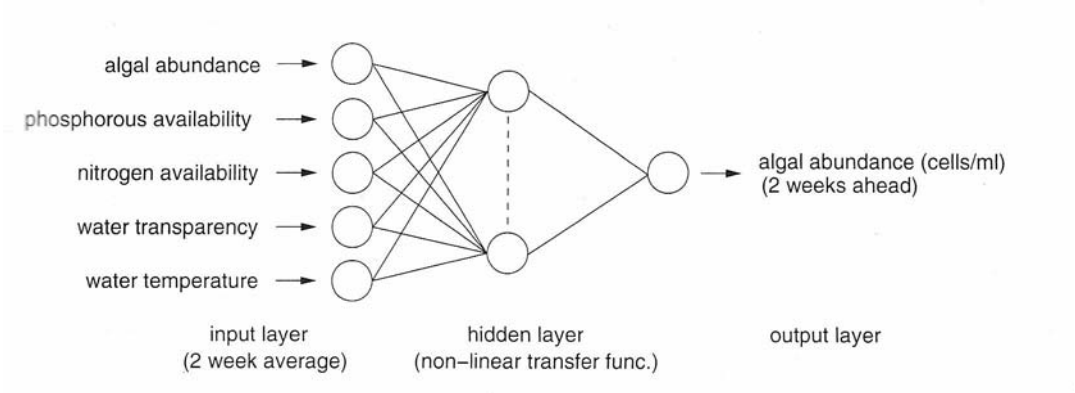


Figure 13. Generic supervised feedforward artificial neural network for 14-days ahead forecasts of algal abundance (from Wilson, 2004)

Some comparisons can be drawn between Wilson's work and the approach intended for this PhD project. Both look at developing models for algal population forecasting, with a focus on generic models. Relatively few inputs are used, in the hopes of the models being widely applicable to water quality databases. However Wilson used solely SNN for this purpose, whereas this project will use a range of machine learning methods, including SNN, in a combined approach for knowledge discovery and forecasting. With respect to the application of SNN, Wilson used supervised feedforward ANNs based on Multi Layer Perceptrons trained with the back

propagation method in contrast to RANN to be used in this project, belonging to the feedback category of SNNs. He also used non-interpolated data as opposed to the daily-interpolated data used for this project. Although Wilson's generic model was able to provide reasonable forecasts of Chl-*a* 2 weeks in advance, this project will focus on 7-day ahead forecasts (determined the best forecasting horizon through consultation with water management authorities and preliminary modelling experiments) and aims to increase accuracy levels. SNN model representation has traditionally been considered 'black box' and whilst a lot of this can be overcome by sensitivity analyses, HEA derived rule sets will be used to test for more instructive and explanatory model representation. The resulting rule-sets have the additional benefits of being relatively portable and easy to apply compared to SNN developed model structures. Wilson's models used a lagged input of the output variable, however in this work that will not be considered, in an effort to test the models' capabilities and to ensure autocorrelation is not supporting good results, and it is acknowledged that the inclusion of this feature would likely improve results significantly. Finally, the sensitivity analyses performed on best performing models are different. Though both studies include sensitivity analyses that suggest the most influential variable and further examine sensitivity by varying one input variable over a range of values and hold the others constant, there are differences in the method of the analysis and representation of the results.

2.4.3 Empirical based modelling of the Happy Valley Reservoir

A 2003 SA Water report (Ingleton 2003b) was written to describe and demonstrate a quantitative evaluation of the potential production of odours and toxins caused by *Anabaena circinalis* and *Microcystis aeruginosa* in the reservoir, based on a number of scenarios regarding source water quality, specifically nutrient concentrations. The method involved using a series of empirical relationships based on published and actual historical data to model the growth potential of cyanobacteria and the generation of water quality hazards in the Happy Valley Reservoir. The goal of this report was to allow assessment of current and future risks to the water treatment plant caused by increased nutrient pollution in the source water and associated increased algal blooms and toxins. It also allowed the examination of impacts of different reservoir management strategies and nutrient reduction programs.

2.5 Management history of the study sites

2.5.1 Myponga Reservoir

Due to its purpose as a drinking water supply storage; Myponga reservoir is highly managed. The reservoir has a history of problems such as large numbers of phantom midge larvae (*Chaoborus Sp.*) in the 1970s, thermal stratification and nuisance growths of cyanobacteria resulting in taste and odour problems in the reticulated drinking water since the 1960s (Kelly 1998; Velzeboer *et al.* 1991). Various water quality management techniques have been employed throughout the reservoir's existence in an attempt to manage or prevent such problems. The algicide copper sulphate ($CuSO_4$) was first used in Myponga reservoir in 1963 to control the growth of *Synura* and *Microcystis*, and later *Ceratium* and *Anabaena*. It was, and still is, used specifically to rapidly reduce the biomass of cyanobacteria, and therefore reduce the risk of the associated production of geosmin, the cause of taste and odour problems; and prevent the release of large amounts of potentially toxic substances (Lewis et al. 2003). The reservoir has been dosed, up to 5 times per year, most years since 1963 and, additionally, the inlet to the reservoir was dosed continuously from 1963 to 1977. In 1980 the Engineering and Water Supply Department began using a 30m long aerator, which bubbled compressed air through diffusers, to combat these problems. Phantom midge was successfully controlled using this method but the aeration did not destratify the reservoir or control excessive cyanobacterial growth. Consequently the reservoir was regularly dosed with $CuSO_4$ to control algal growth throughout the 1970s and 1980s (Velzeboer et al. 1991). In 1988, three submersible Flygt mixers were installed on the dam wall to replace the ineffective aeration system (ITT-Industries 1990). The mixers were to increase the dissolved oxygen content of the water and destratify the reservoir, which would prevent the development of algal blooms. The water quality of the reservoir was then monitored over the 1988 to 1989 summer period, to investigate the effectiveness of the mixers. From this data, it was concluded that the mixers were not successful in totally destratifying the reservoir, with both surface heating and weak sub surface structure evident for substantial periods of time, but were successful in reducing total algal biomass. However, $CuSO_4$ dosing was still required on three instances throughout the summer to control cyanobacterial blooms, specifically *Anabaena*. The use of the mixers also led to a layer of anoxic water below them, which contained significant concentrations of iron, manganese and nutrients. The Flygt mixers were also deemed to be not economically efficient (Kelly 1998; Smalley 1998; Velzeboer *et al.* 1991).

The Flygt mixers were replaced in 1990 with a new aerator, approximately 400m long, which extends out from the dam wall along the sediments of the reservoir and was run continuously throughout summer. However, cyanobacterial blooms still occurred and $CuSO_4$ was still required to be used as an additional method to control algal growth, with multiple treatments occurring each year. The aerator only achieved partial destratification and a small thermal gradient was found to persist near the surface of the water. Kelly (1998) suggested that this microstratification at the surface could be a contributing factor to the reoccurring algal blooms at Myponga reservoir. The microstratification generally causes a decrease in the mixed depth of algae, rather than an increase. This results in a greater gross phytosynthetic:respiration ratio and, consequently, more algal biomass.

Surface mixers (SMDI-5, Water Engineering Research Solutions, QLD Australia) were installed in 1999 primarily to limit the growth of buoyant cyanobacteria and, additionally, to destratify the reservoir. In order to circulate large quantities of water, physically large surface mixers were manufactured for Myponga reservoir. The blades are 4.9m in diameter and pump the water down through a draft tube (diameter 4.9m, length 13m) at a flow rate above $3.5m^3/s$ to a depth of about 14m. The surface mixers are powered by 4kW motors (Lewis *et al.* 2003; Lewis *et al.* 2002). The aerator and mixers are operated between October and March each year (Brookes et al. 2002a).

For the remainder of the period considered in the project (1999-2003), water quality management at Myponga reservoir consists of artificial mixing, achieved by the use of a multi-diffuser aerator and two raft-mounted mechanical surface mixers, as well as $CuSO_4$ application to manage the threat of cyanobacterial blooms in particular. The aerator is located adjacent to the dam wall at 30m depth. Its diffuser is over 200m in length, with 160 outlets delivering air at 120L/s via a 100KW compressor (Lewis et al. 2003). It was concluded from historical data that the aerator is effective in controlling the release of some metals from the sediments, minimally reduces the total annual algal biomass and has some destratification abilities.

Even with a management regime involving the use of an aerator and surface mixers, it is still necessary to use $CuSO_4$ dosing in mid-summer to control cyanobacterial growth, particularly *Anabaena*, in Myponga reservoir. In South Australia, $CuSO_4$ is applied regularly before the algae reaches bloom levels, usually at relatively low cell counts often between 1000-2000 cell/mL (Kelly 1998), to prevent taste and odour problems. The $CuSO_4$ is applied at a dose rate of 2mg/L and is active for a limited time (<15min) before forming inert complexes that descend to the sediment and out of the water column. During the period of algicide treatment, no artificial mixing

is carried out. Although phytoplankton biomass recovery is fast during the summer period, the successional development of the phytoplankton community can be altered by the algicide treatment.

2.5.2 Happy Valley Reservoir

Happy Valley reservoir has a history of persistent water quality problems related to the excessive growth of both phytoplankton and zooplankton. Historically, the major problem genera have been *Cyclotella*, *Dictyosphaerium* and *Oocystis*, which regularly dominated the reservoir with very large cell numbers or biomass. The phytoplankton blooms also help give rise to excessive zooplankton abundances, particularly *Calamoecia* and *Ceriodaphnia* (Burch 1987). With the completion of the adjoining water treatment plant (WTP) in 1991, problems often associated with blooms of the above listed plankton became less of a concern. Although diatom and green algae blooms can affect drinking water quality and block filters, the increasing presence of cyanobacteria species became most important. Cyanobacteria blooms negatively affect water quality through the production of potentially toxic substances as well as taste and odour compounds. Two potentially toxic species are present in the Happy Valley reservoir, *Microcystis aeruginosa* and *Anabaena circinalis*. *Microcystis aeruginosa*, although occurring infrequently, has long been considered a problem species in the reservoir (Burch 1987). *Anabaena circinalis* is much more prevalent and, consequently, is considered the main nuisance species at Happy Valley reservoir (Ingleton 2003b).

The Happy Valley WTP is not equipped to remove high concentrations of algal metabolites (>40g/L) such as odours or toxins which is why bloom events, particularly cyanobacterial, must be constrained or prevented (Ingleton 2003a).

Throughout the history of Happy Valley reservoir, two techniques have been used in an effort to reduce the frequency and severity of algal blooms. $CuSo_4$ dosing has been used regularly since the 1960s to rapidly reduce specific algae and zooplankton biomass when abundance reaches an unacceptable level for the target organism. Whilst $CuSo_4$ treatment is generally very effective at controlling bloom events, on occasions in Happy Valley reservoir it has had little effect on some of the problem species (Burch 1987).

Aeration was employed in 1981 for a number of reasons, namely to attempt to destratify the reservoir, thereby disturbing the favourable conditions prompting algal and zooplankton blooms in summer, increasing the dissolved oxygen concentration and preventing nutrient release from the sediment. Numerous problems were experienced throughout the reservoir due to the aerator

failing to achieve continuous or complete destratification. Although it did elevate dissolved oxygen concentrations, at times it also resulted in greater nutrient levels in the surface waters and, consequently, encouraged even more plankton growth (Burch 1987).

Over the summer of 1998-1999 a surface mounted mixer, similar to that placed in Myponga reservoir, was implemented in Happy Valley reservoir (CRCWQT 2003). In 2003 a new bubble plume aerator was installed, designed to destratify the deeper area of the reservoir (CRCWQT 2005) (Ingleton 2005). The $CuSo_4$ dosing in this reservoir is carried out when *Anabaena* densities reach 500 cells/ml, as the blooms tend to get out of control very quickly and then require even more algicide if not caught early (Daley & Ingleton 2006).

2.5.3 Key water management issues

Key management issues in Myponga and Happy Valley reservoirs are largely related to thermal stratification during the warmer months of the year (October - March). Despite management in place to destratify the reservoirs, this is not always achieved with weak or micro stratification occurring and occasional strong stratification still taking place during periods of high insolation and stable conditions (Lewis 2004).

Stratification is characterised by a warm, mixed upper stratum, called an epilmnion, where dissolved and particulate substances are considered to be homogenously distributed. In this layer, oxygen is provided by atmospheric aeration as well as photosynthesis and is therefore at a high level. The epilimnion sits over a cool, dense and undisturbed layer of water, known as the hypolimnion, where biological oxygen demand (BOD) is very high but there is no natural source of oxygen, thus the hypolimnion gradually becomes anoxic (Recknagel 2002). The stratum between the epilmnion and the hypolimnion is termed the metalimnion and is characterised by the thermocline, the plane where the greatest gradient in temperature occurs (Boulton & Brock 1999). There are many implications of stratification; most are detrimental to the quality of the water. The main impacts of concern are increased algal growth and release of substances from the sediment (McAuliffe & Rosich 1990). Stratification tends to promote the growth of cyanobacteria and the conditions are particularly encouraging for those species with buoyancy regulation abilities. The stability of the water column allows these algae to exploit their ability to gain sufficient light for photosynthesis in the epilimnion, and then lower themselves in the water column to gain access to nutrient rich water if needed (Brookes & Burch 2006). Both Myponga and Happy Valley reservoirs experience cyanobacteria blooms, predominantly *Anabaena circinalis*, in the summer months and this is a major concern. This species is problematic due to the potential for release

of geosmin, causing taste and odour issues; or saxitoxins that disrupt nerve function and transmission by blocking sodium channels in nerve cells, which can result in death by inhibiting the muscles essential for respiration (Humpage & Froscio 2006). Once these substances have become dissolved in the water, additional treatment is required for their removal, which increases costs and potentially imposes a withholding period on the water.

Stratification also promotes the release of undesirable substances from the sediment. The lack of oxygen input into the hypolimnion and the consumption of dissolved oxygen (DO) by biological processes, including bacterial respiration, steadily depletes the DO levels until the hypolimnion becomes anoxic. Under anoxic conditions, contaminants such as iron (Fe), manganese (Mn), phosphorus (P) and ammonia (NH_3) are released from the sediment and into the overlying water (Brookes & Antenucci 2006; McAuliffe & Rosich 1990). The release of such substances may lead to a variety of water quality issues including clogging of pipes, aesthetic issues and in extreme cases, health problems (Kirke 2000). It can also exacerbate problems with algal growth when the nutrient and metal rich water eventually becomes available to the surface layer. In Myponga reservoir, the release of Fe and Mn are of concern because if not removed by an oxidation-precipitation process, they cause water quality issues when oxidised by chlorine or air in the distribution system and cause dirty water at the tap. And again, increased treatment costs are associated with this issue. In this reservoir, it has been shown that the concentrations of these substances in the hypolimnion are correlated to the duration of thermal stratification (Brookes et al. 2000).

Artificial destratification is used to prevent the onset or disrupt existing stratification, thereby easing the biological and chemical water quality issues associated with it. Artificial mixing to destratify water bodies has two main functions: the control of phytoplankton (particularly cyanobacteria) by increasing the mixing depth of the cells, thereby decreasing their time in the euphotic zone; and reduction of internal loading caused by an anoxic hypolimnion, by increasing DO concentrations throughout the water column.

Currently, artificial mixing can be achieved through the use of one or both of bubble plume aerators and mechanical mixers. Bubble plume aerators are placed near the bottom of the storage, and release air bubbles that rise and entrain water from the lower layer of water. If energy input is sufficient, stratification should be weakened by this turbulence as the bottom water reaches the surface and the circulation moves the warm water from the surface layer downward. Mechanical mixers or impellers are a method of circulating and exchanging water

between the surface and lower stratums and there is a range of different designs (Kirke 2000; Lewis 2004).

Although artificial destratification has been widely adopted, the success of technique for the control of phytoplankton and sediment release is varied. Generally, much more success has been achieved in the reduction of metals and nutrients from the sediment than with the control of algae. In the case of total algal biomass, Imteaz and Asaeda (2000) demonstrate that while some authors have reported a decrease (Bernhardt 1967; Robinson *et al.* 1969), numerous authors have reported increases caused by the mixing (Barker 1976; Drury *et al.* 1975; Knoppert *et al.* 1970). In some cases where a water body experiences both sediment release and algae problems, significant control over sediments has been achieved but algal abundances have remained high (McAuliffe & Rosich 1990). Myponga appears to be one such reservoir. Several studies have shown that the aerator successfully reduced internal loading during 1986 to 1996, particularly the release of Fe and Mn (Brookes *et al.* 2002a; Brookes *et al.* 2000), yet cyanobacteria blooms are still an issue. Herath (1997) reports of the successful adoption of artificial mixing to control algal blooms throughout Europe but states that the Australian experience has been that approximately 70% of the reservoirs with destratification systems have shown no significant reduction of algal biomass. In some cases it can even exacerbate water quality issues by re-suspending substances like Fe and Mn, making them available in the surface layer. The varied success of artificial destratification for algal population control in Australia is not well understood. Where algal blooms are mainly controlled by light, artificial mixing should be able to reduce the problem by increasing the mixing depth of the algae, thereby reducing their time in the eupohtic zone (McAuliffe & Rosich 1990). Light limitation is considered to be the controlling factor of algal biomass in Myponga reservoir (Lewis et al. 2003), and therefore artificial mixing should be able to reduce phytoplankton biomass. Considering that, the study sites are theoretically prime candidates for algal control by means of artificial mixing and it may be that deficiencies in its implementation, rather than the concept as a whole, are the reason for the limited success thus far.

2.5.4 Limitations of current management regimes

The management practices used in Myponga and Happy Valley reservoirs to this point in time have not been sufficiently effective in decreasing the frequency and intensity of cyanobacterial blooms. In Myponga reservoir, the primary intention of the artificial mixing was to limit the growth of cyanobacteria (Lewis et al. 2003), which appears not to have been successful, although there

has been some indication of reduction in sediment release. Although the mixing is said to positively influence the succession of algal species, the reservoirs still experience problematic cyanobacteria blooms in the warmer seasons and there has been no significant impact on algal biomass. Thus far, artificial destratification in Happy Valley reservoir has also been unsuccessful in reducing the frequency and severity of cyanobacterial blooms.

Often complete destratification fails to be achieved by the aerators and mixers and the environment is still conducive to algal population growth and maintenance, resulting in continued reliance on $CuSO_4$ dosing to combat developing blooms.

The use of $CuSO_4$ is not desirable for a number of reasons and, ideally, other less harmful management methods should prevent the occurrence of a situation requiring the use of this algicide. While intracellular toxins in intact cells can be readily removed during the water treatment process, once the toxins are released into the water additional treatment may be necessary such as adsorption by activated carbon (Burch et al. 2002). Treating blooms with algicide causes cell lysis that releases intracellular toxins or taste and odour compounds into the water, potentially enforcing a withholding period onto the reservoir (House & Burch 2002). The toxins and odours degrade with time and, whilst there is little information on required withholding times for affected water bodies, it has been suggested that toxin breakdown in lakes could surpass 14 days (Jones & Orr 1994).

$CuSO_4$ is considered the algicide of choice in South Australia and is largely regarded as effective, economical and safe for operators to use, however its potential for environmental damage is a contentious issue (Burch et al. 2002). Copper tends to accumulate in the sediments and is largely considered to remain permanently bound to the bottom sediments. However, a number of studies (Prepas & Murphy 1988; Van Hullebusch *et al.* 2003) have demonstrated that a significant amount of sediment-borne copper is associated with the organic fraction, which may release back into the open water under particular conditions such as low DO concentrations. Hanson et al. (1984) proposed that copper accumulation in sediments may impact negatively on the benthic macroinvertebrate community. The sulphate component of $CuSO_4$ may change the pH of the water body and affect any nearby organisms during the initial reaction (Ingleton 2003a). Pitois (2000) concluded that $CuSO_4$ dosing could not be used as a lake management tool without degrading the water in terms of dissolved organics. Most concerning is the fact that $CuSO_4$ is non specific, consequently having an extreme ecological impact by also being toxic to potentially beneficial non-target organisms such as zooplankton and fish (Burch *et al.* 2002; Hrudey *et al.* 1999; Johnstone 1994). Impacts of copper on human health are not well established but Sparks

and Schreurs (2003) have reported some evidence linking it to Alzheimer's disease (Ingleton 2003a).

2.5.5 Potential improvements to management regime

In order to minimise water quality problems caused by cyanobacteria blooms, it is desirable to either operationally control or prevent them. Computational modelling is a way to contribute to both. It enables both the forecasting of growing algal population abundance several days in advance and a better understanding of processes and environmental conditions that accelerate algal growth. Appropriate measures can then be implemented to control the growth of algal populations before they reach bloom proportions.

The Murray-Darling Basin Commission (1993) suggested that to achieve a better understanding of the mechanisms that activate algal blooms, real-time in situ water quality monitoring might be required. Burch (1987) states that effective management of nuisance plankton populations in reservoirs is based on the ability to predict the timing and extent of their growth. The logical extension of the prediction of growth is artificial control of the parameters that regulate growth. This project ultimately combines both in situ water quality monitoring and computational modelling to produce real-time forecasts of algal populations. The use of pattern and sensitivity analysis by KANN, RANN and HEA provides information on parameters that significantly influence the algal population. Both RANN and HEA have demonstrated their capability to perform short-term forecasting of algal population growth and provide information of factors driving these forecasts in many studies (eg. (Bobbin & Recknagel 2003; Cao *et al.* 2005; Jeong *et al.* 2003; Recknagel *et al.* 2004; Recknagel *et al.* 2005; Talib *et al.* 2005; Welk 2003). These forecasts should allow the timely implementation of an operational control measure to prevent or control an algal bloom.

As previously discussed, the current management regimes at Myponga and Happy Valley reservoirs continue to require the use of $CuSO_4$ as they are failing to make an impact upon cyanobacteria bloom frequency or intensity using continuous artificial mixing from October to March, and therefore a different approach should be tested. Benefits such as reduced internal loading show that artificial destratification can improve water quality in reservoirs, and it is one of the most sustainable methods to do so (Brookes et al. 2002a), thus the continuation of this technique is suggested, though with some tweaking of the implementation. Intermittent artificial mixing may be a possible solution in these reservoirs and it has numerous benefits over continuous mixing. Reynolds et al. (1984) explains that the alternating mixing and stratification

conditions that are achieved by intermittent mixing, will only favour particular species for short times before the conditions change and a different group of species is promoted. This way, no group of algae would get to realise its potential before conditions became unfavourable again. Furthermore, it has been observed that some algae can adapt to the conditions offered by continuous mixing and become able to flourish in an environment initially designed to discourage them. Intermittent mixing would avoid this issue. A 2003 report into algal mitigation options for Happy Valley reservoir found that, of all the management strategies that were modelled for the investigation, intermittent operation of mixers and aerators (on a 2 out of 4 day rate) was found to achieve very high levels of cyanobacteria growth inhibition (Ingleton 2003a). However, because sediment release is a concern in Myponga and Happy Valley reservoirs, along with cyanobacterial blooms (though not as significant) it may not be wise to completely disregard continuous mixing, as it appears to be beneficial in reducing the internal loading. Therefore, strategies involving a combination of intermittent and continuous mixing could be suggested. An example of such a regime is found in Lake Nieuwe, the Netherlands where continuous mixing had previously been used for the spring/summer period to control cyanobacteria. During the trial intermittent mixing was used in spring, and continuous mixing was initiated in summer. A 75% reduction in energy costs was achieved for the spring period using artificial mixing, and the energy costs for the whole spring/summer period were only 27% of the previous year with continuous mixing for those seasons. Importantly, during the period of intermittent mixing there were no disadvantageous consequences for the oxygen content of the water (Visser et al. 1996). Considering the results of these studies suggest that reduced cyanobacterial growth and an oxygenated water column are possible with various regimes of intermittent mixing, potential management strategies for Myponga and Happy Valley reservoirs could be: bubble plume aerators used continuously and surface mixers only initialised when forecasted abundances are above a certain threshold level. Or perhaps, as with the regime implemented at Lake Nieuwe, continuous mixing could be carried out in only Jan-Feb, the most problematic time, and intermittent mixing during the other months of concern. Ideally, the management will keep cyanobacterial populations to an acceptable level, thereby reducing the reliance on $CuSO_4$ treatment.

2.6 Online, real-time monitoring and forecasting

In order to minimise expenses to water industries and reduce public health concerns caused by harmful algal outbreaks in drinking water reservoirs, appropriate predictive tools and early warning systems need to be developed and implemented. To maintain a consistent supply of high quality, safe drinking water reserves, reservoir water quality must be continuously monitored. This continuous monitoring can only feasibly be achieved by on-line monitoring (Fogelman 2006). Traditional field monitoring uses manual water sampling, often with sampling frequencies limited to weekly- monthly sampling. In the case of low frequency manual sampling, much water quality information is missed as hazards can develop at time-scales shorter than the sampling frequency, such as algal blooms which can be very dynamic in nature and can occur over a period of days. In general there is a need for high frequency data to understand water quality processes and the time-scales over which they occur including the changes in water quality in a typical algal bloom event, and the conditions leading up to it (Wong 2004). Online, real-time monitoring could also assist in the management of water quality hazards other than algal blooms, including problems occurring over different time scales, such as soluble metal release from the sediment which develops over a scale of weeks to months (Brookes et al. 2002b). As well as being able to capture more information of water quality and algal dynamics, continuous online monitoring can be linked to real-time forecasting. Current monitoring practices are lab-based and because of the time taken for sampling, transportation to the laboratory and analysis of the samples, they are inappropriate for on-line monitoring and unable to be linked to real-time forecasting where rapid feedback is essential (Fogelman 2004). Numerous techniques have, and continue to be, developed for real-time monitoring, including optical methods through to identification of genetic signatures using PCR-based methods (Sellner et al. 2003). Online monitoring of reservoirs has not been widely practised in the past, as the real-time continuous information was not regarded as useful in operational decision-making (Brookes et al. 2002b). However, when online monitoring is linked to real-time forecasting, it can become very useful for prediction of water quality hazards and provide decision support that will enable proactive, rather than reactive, intervention and management.

Data from continuous online monitoring systems can be used to develop early warning systems for algal blooms, incorporating detection and prediction of dynamics. Forecasting methods that are capable of fast response to highly non-linear and rapid events, and allow preventive and operational control are required. Traditional methods of forecasting have proven to be unsuitable

for use in operational management of freshwater / reservoir systems. Real time forecasting, based on continuous *in situ* measurements of lake water quality and climate data in conjunction with forecasting models, will allow the prediction of sudden growth and high abundance of algal populations several days in advance. To optimise the efficiency of algal bloom control methods and minimise their impact on the ecosystem, it is essential they be applied preventatively or at a very early stage of bloom development, at the appropriate level or dosage. Early warning at an initial stage in algal bloom development based on real time forecasting makes the application of alternative operational control options such as intermittent mixing (NRA 1990) or pH shock (Klapper 1991) possible, and can be used by water management authorities for decision support and to make provisions to alter treatment plant operation in the case of impending blooms (Maier et al. 1998). In recent years, significant advances have been achieved in both the range and accuracy of electronic water quality measurements and the predictive modelling of algal blooms, confirming that this approach is a viable option for future algal dynamics and water quality management.

Examples in the literature of systems using continuous online water quality monitoring, to be linked with modelling methods include:

Romero et al. (2003) included real-time monitoring, databases, numerical models and visualisation tools in a decision support system, Aquatic Real-time Management System (ARMS), which was applied to two Sydney water supply reservoirs. The system had use of large data sources of real-time and historical data for both Lake Burragorang and the Prospect reservoir including stream data flow rates and water quality, in-lake meteorology and water column temperature data, and withdrawal rates (Ewing et al. 2004). Three process-based models were integrated into the system, namely DYRESM, ELCOM (one and three dimensional hydrodynamics models respectively) and CAEDYM (an ecological model), with the aim to increase understanding of reservoir water quality processes, to evaluate management scenarios and to forecast effects of perturbations. DYRESM was applied to long-term simulations between one month and several years, for example a two-year DYRESM-CAEDYM simulation of the response of Prospect reservoir to changes in its flow regime. ELCOM was used for event modelling tasks between one week and one month, such as an ELCOM-CAEDAM simulation of the fate of a flood inflow to Burragorang reservoir, tracking contaminants through the weeks after a rain event. DYRESM-CAEDYM was tested on Prospect reservoir (1988-1990), with the aim of simulating one-dimensional seasonal dynamics. It was shown to simulate stratification patterns, dissolved oxygen and nutrient levels well in general but simulations of hypolimnetic temperatures,

epilimnetic phosphate concentrations and nitrate levels did not match field observations (Romero et al. 2003). ELCOM-CAEDYM simulations were found to be computationally intense, taking many hours or days to run. The three dimensional model ensemble was tested on a flood event of 1997 in Lake Burragorang and successfully simulated the primary underflow dynamics in the reservoir. ARMS is available as a desktop software package.

Wong (2004) linked online, continuous monitoring of two coastal locations in Hong Kong with Lagrangian modelling for the prediction of harmful red tide events caused by motile dinoflagellates. A remote automatic measuring and recording system was used to provide high frequency data including algal and dissolved oxygen dynamics along with key water quality and hydrographic parameters. In vivo chlorophyll concentration was measured in real-time via flurometric measurement and translated to Chl-*a* for use (Wong 2004). Wong used several years of online data to develop the mathematical model for Chl-*a* prediction that considered algal growth and decay, settling and swimming and vertical turbulent mixing. With a nutrient competition consideration, the type of bloom likely to occur can be predicted of either the non-motile or the harmful motile species. The system was shown to be successful in providing quick assessment of the likelihood of red tide occurrence and increasing understanding of controlling factors of bloom formation.

Lee et al. (2004) presented real-time monitoring and forecasting of algal bloom dynamics using artificial neural networks, based on continuous high frequency measurements of water quality and hydro-meteorological parameters at Kat O field monitoring station in Hong Kong. A feedforward supervised artificial neural network was used, specifically a MLP trained with back propagation, to forecast chlorophyll fluorescence (representing algal biomass) with different lead times of up to three days in advance. Daily values of chlorophyll fluorescence, dissolved oxygen, water temperature, solar radiation and wind speed were used as input variables but multivariate statistical analysis of the input variables suggested that the use of past chlorophyll fluorescence values alone was sufficient for algal prediction. Testing was carried out with chlorophyll fluorescence data from July 2001 to October 2001 (approximately 120 samples), and the remaining samples (between 500 and 700) were used for training. The most useful model was found to use chlorophyll fluorescence measurements at times t, t-1, t-2 and t-7 to give forecasts two days in advance, which gave a testing r^2 value of 0.74 and an RMSE of 216.

A study by Muttil and Lee (2005) for the real-time prediction of coastal algal blooms using Genetic Programming (GP), has delivered encouraging results for the use of evolutionary algorithm techniques in real-time forecasting (Muttil & Lee 2005). The evolutionary algorithm was designed

to discover an equation for forecasting chlorophyll fluorescence at the Kat O field station, Hong Kong. Initially GP was used to identify the most significant input variables for predicting chlorophyll fluorescence, at different lead times, from 3 years of high frequency data including chlorophyll fluorescence and related hydro-meteorological and water quality parameters with different lag periods. From the available chlorophyll fluorescence data (March 2000 to March 2003), July 2001 to October 2001 were used for testing (approximately 120 samples) and the remainder (about 800 samples) was used as the training set. The best performing equation (Eq. 3) was found to use the inputs chlorophyll fluorescence (Chl) at time *t* and *t*-1, solar radiation (SR) at *t*-1 and wind speed (WS) at *t*-4 and provide chlorophyll fluorescence forecasts one day in advance. This model provided a correlation coefficient of 0.85 and an RMSE of 165 on the testing data set (Fig. 14). The authors highlighted the advantage of the GP providing an equation relating input and output variables, which facilitates the interpretation of results in comparison with ecological reasoning. It was concluded that GP is a feasible method of algal bloom modelling that can be applied in a real-time situation and give predictions with reasonable accuracy for one day in advance.

Eq.3

$$\mathrm{Chl}(t+1) = \mathrm{Chl}(t) + \frac{[\mathrm{SR}(t-1) - \mathrm{Chl}(t)][\mathrm{Chl}(t) + 0.4\,\mathrm{SR}(t-1) + 0.4]}{20 + \mathrm{SR}(t-1)\left[\mathrm{Chl}(t) + \frac{2}{[\mathrm{SR}(t-1) - \mathrm{Chl}(t)+2][\mathrm{SR}(t-1) - \mathrm{Chl}(t)]}\right] - \mathrm{Chl}(t) + 0.12\,\mathrm{WS}(t-4)\mathrm{Chl}(t)}$$

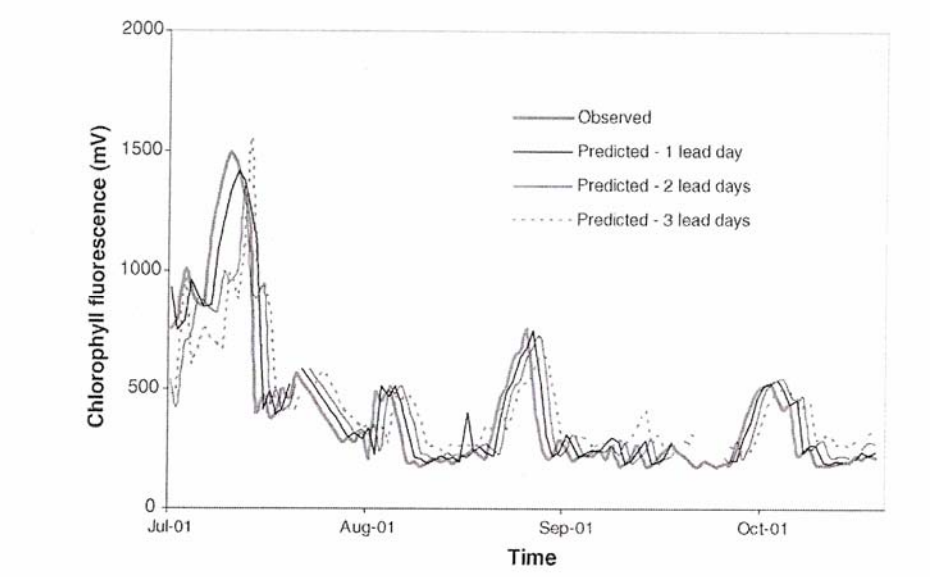


Figure 14. Results of chlorophyll fluorescence forecasts, with 1, 2 and 3 days lead time, using a predictive equation (from Muttil and Lee (2005)

In this study, machine learning methods, RANN and HEA, are proposed to be trialled as potential methods to link with real-time water quality monitoring data (or simulated real-time data) in order to enable real-time forecasting of algal blooms in two South Australian drinking water reservoirs.

2.7 Summary

Ecological modelling has long been recognised as a very powerful approach for analysis and prediction of ecological systems, processes, interactions and occurrences. In recent years, it has been found that computational, data driven methods can successfully be applied to time-series data from complex, non-linear systems such as freshwaters. In particular, ANN and EA have shown that they can be effectively used to discover and predict patterns within datasets, and provide further information and understanding of the causal relationships driving algal dynamics.

Myponga and Happy Valley reservoirs are the study sites for this project. In the past both reservoirs have experienced thermal stratification in summer and nuisance cyanobacteria blooms. Although the incidence and intensity of thermal stratification has decreased since the introduction of artificial mixing in the reservoirs, cyanobacteria blooms continue to plague both locations.

This study aims to apply ANN and EA to water quality databases from Myponga and Happy Valley reservoirs to develop predictive and explanatory models with focuses including algal dynamics, both event based and seasonal, and management regimes. Information gained from such models can be used to assist decision-making and management.

3. MATERIALS, MODEL DESIGN AND APPLICATION

This chapter provides a general overview of the data and methods; however, specific data sets and modelling approaches used for each case study or experiment are described in each relevant chapter.

3.1 Myponga reservoir site and data summaries

3.1.1 Myponga reservoir site description

Myponga reservoir was built as a concrete arch-dam with a ski jump spill way and completed in 1962 (Government 1962). It is located approximately 60km south of Adelaide, South Australia, on the Myponga river and is a vital water supply to southern metropolitan Adelaide and the Fleurieu Peninsula (Lewis et al. 2003). The reservoir has a storage capacity of 26,800 ML, a maximum depth of 35m, an average depth of 15m and the area of water spread is 2.8km^2. The mean water retention time based on abstraction is approximately three years and water is removed from the reservoir via an offtake valve on the dam wall. Myponga reservoir has a shallow average euphotic depth of 3m, due to its highly coloured nature. The average light extinction coefficient is about 1.5m, although Myponga is generally low in turbidity (Lewis *et al.* 2003; Velzeboer *et al.* 1991).

The catchment area covers 124km^2 and is located within the Adelaide foothills, accommodating an intense dairy and cattle farming community. Although the dominant land use within the catchment is estimated to be 62% livestock grazing and 24% dairying (Thomas et al. 1999), pine plantation, grassy pastures and small fragments of native vegetation can be found in the area surrounding the reservoir, supported by an average annual rainfall of 750mm (Smalley 1998).

Some of the most dominant phytoplankton types found in Myponga reservoir are *Scenedesmus*, *Cryptomonas*, *Anabaena*, *Nitzschia*, *Cyclotella*, *Chroomonas*, *Chlamydomonas* and *Monoraphidium* (Lewis et al. 2002) (Lewis 2004). The reservoir is generally well mixed from late April to late September, and artificial mixing is employed from October to March in an effort to maintain mixed conditions throughout the warmer months, though this is not always achieved. Cyanobacteria, such as *Anabaena*, can usually only dominate during periods of thermal stratification, when they can exploit the stable water column by using their buoyancy control to

access both light and nutrients. They are therefore restricted to summer and autumn. Being non-motile, some of the green algae and diatoms prefer mixed conditions and can therefore theoretically dominate at any time in the year. However, diatoms are highly perceptible to sedimentation losses due to their heavy nature and rely on re-suspension via turbulence. Therefore they tend to persist during extended periods of strongly mixed conditions. Green algae prefer higher light intensities, even though they can handle variable light conditions and thus prefer the warmer seasons with higher solar radiation. Green algae tend to make up most of the algal biomass in Myponga reservoir (Lewis 2004). Although no serious water quality hazards are created by green algae, high abundances can cause problems in the water treatment plant, e.g. filter clogging.

Due to the employment of artificial destratification, overall the conditions at Myponga reservoir in relation to algal growth can be considered as well mixed with poor light conditions (Lewis 2004). However, as discussed above, in the warmer months of most years the reservoir undergoes thermal stratification, of varying intensity and duration, due to the increased energy from radiation and decreased water inflow from the catchment. In the past this stratification combined with the highly coloured and nutrient rich water has created optimal conditions for the growth of algae, particularly cyanobacteria (Kelly 1998). Major bloom events are often in mid-January, although management is on alert from December to March annually (Burch 2005b). *Anabaena circinalis* has been identified as the key nuisance species at these times.

3.1.2 Historical data from Myponga reservoir

Data was collected by water quality management authorities since the 1960's, however much of it was too inconsistent for this particular use, thus only 18 years (1986-2003) of data was selected to be used in the present study. The data used was collected from Location 1 in the reservoir, as it was nearest the off-take valve on the dam wall and also supplied the longest, most consistent dataset. The dataset contains 14 variables, describing the physical, chemical and biological conditions of the water body (see Tab. 1).

Table 1. Water quality data from Myponga reservoir Sampling Location 1

VARIABLE	ABBREVIATION	UNIT	USEABLE DATA	MEAN	MIN	MAX	STND. DEV.
Chlorophyll-*a*	Chl-*a*	ug/L	1986-2003	7.84	0.2	41.6	6.92
Water temperature	-	°C	1986-2003	16.1	8.4	25	3.81
Turbidity	-	NTU	1986-2003	4.4	1.2	30	2.58
Colour	-	HU	1986-2003	72.9	28	166	22.71
Total Phosphorus	Total P	mg/L	1986-2003	0.06	0.011	0.218	0.026
Phosphate	PO_4	mg/L	1986-2003	0.022	0	0.09	0.019
Nitrate	NO_3	mg/L	1986-2003	0.11	0.001	0.37	0.092
Iron (soluble)	Fe	mg/L	1986-2003	0.404	0.05	1.11	0.2
Manganese (soluble)	Mn	mg/L	1986-2003	0.025	0.005	0.12	0.028
Dissolved Oxygen	DO	mg/L	1993-2003	8.78	3	13.16	1.31
Conductivity	-	uS/cm	1999-2003	620.39	525	781	47.69
Anabaena	-	cells/mL	1991-1993 1995-2003	185.9	1	5933	667.4
Scenedesmus	-	cells/mL	1986-1991 1996-2003	8443.9	1	188000	19460.5
Nitzschia	-	cells/mL	1996-2003	953.53	1	13100	2205.78

3.2 Happy Valley reservoir site and data summaries

3.2.1 Happy Valley reservoir site description

Happy Valley reservoir was built between 1892 and 1897, 15km south of Adelaide. Happy Valley reservoir is a key drinking water supply for metropolitan Adelaide as it provides the water for the Happy Valley water treatment plant (WTP), completed in 1991, that serves 40% of the metropolitan area (SAWater 2002; United-Water 2005).

The reservoir is relatively shallow with an average depth of 6.8m and a maximum depth of 19m. The usable capacity of the reservoir is approximately 12,000ML, with the area of water spread being 1.88 km^2 at full supply level (Burch 1987).

In 2002, a rehabilitation project was initiated to ensure Happy Valley met or exceeded national and international guidelines for best practice management of dam structures (SAWater 2003). These works increased flood storage capacity and reduce risk of failure and leakage. The dam was constructed of earth with a clay core, with the dam wall being approximately 25m high and 1155m long (SAWater 2002).

Happy Valley reservoir receives water from not only the Onkaparinga Catchment, but also from the Murray River. The Onkaparinga Catchment is a multiple use catchment including settlements and agriculture such as dairy farming, grazing, market gardening, and viticulture (Ingleton 2003a).

The most important phytoplankton genera in Happy Valley reservoir, based on their regular occurrence as dominants or subdominants, are *Cyclotella*, *Dictyosphaerium*, and *Anabaena*. Records show two potentially toxic species, *Microcystis aeruginosa* and *Anabaena circinalis*, present in the reservoir but *Anabaena circinalis* is the main species of concern (Burch 1987; Ingleton 2003b).

Happy Valley reservoir experiences at least weak thermal stratification for up to 5 months each year (November – March). Whilst well mixed isothermal conditions usually exist during May to September, the warmer months often experience intense stratification accompanied by sediment nutrient release and cyanobacterial blooms (Burch 1987). As with Myponga reservoir, the key nuisance species during these times in *Anabaena circinalis.* Artificial destratification is employed from around October to March in an attempt to prevent anoxic conditions at the sediment and cyanobacterial bloom events.

3.2.2 Historical data from Happy Valley reservoir

Even though some data was collected in Happy Valley reservoir from the 1940's onwards, only 8 years (91, 97-2003) of data was comprehensive and consistent enough for use in the present study. The data used was collected from Location 1 in the reservoir as it provided the longest dataset with maximum variables. The dataset contains 14 variables, describing the physical, chemical and biological conditions of the water body (see Tab. 2).

Table 2. Water quality data from Happy Valley reservoir Sampling Location 1.

VARIABLE	ABBREVIATION	UNIT	USEABLE DATA	MEAN	MIN	MAX	STND. DEV.
Chlorophyll-*a*	Chl-*a*	ug/L	1991-2003	9.08	0	66.45	7.81
Water temperature	-	°C	1990-2003	16.75	6	36	3.51
Turbidity	-	NTU	1990-1991 1994-2003	9.04	1.8	41	6.97
Colour	-	HU	1990-1991 1994-2003	43.54	5	111	24.25
Total Phosphorous	Total P	mg/L	1990-1991 1994-2003	0.098	0.017	0.246	0.051
Phosphate	PO_4	mg/L	1990-1991 1994-2003	0.024	0.006	0.076	0.015
Nitrate	NO_3	mg/L	1990-1991 1994-2003	0.226	0.005	0.63	0.157
Iron	Fe	mg/L	1990-1991 1994-2003	0.151	0.006	0.433	0.098
Manganese	Mn	mg/L	1990-1991 1994-2003	0.015	0.005	0.166	0.018
Dissolved Oxygen	DO	mg/L	1990-2003	9.24	2.2	21.4	1.33
Conductivity	-	uS/cm	1991, 1995-2003	650.34	436	1030	80.78
Anabaena	-	cells/mL	1991-2003	130.9	0	6660	423.1
Dictyosphaerium	-	cells/mL	1990-2003	1008.2	0	26000	2233
Cyclotella	-	cells/mL	1990-2003	474.3	0	18600	1486.1

3.3 A comparison of study sites

Both reservoirs are located within 60km south of Adelaide, South Australia (see Fig. 15), and experience Mediterranean climate. Both serve as drinking water supplies providing water to on-site water treatment plants that service large parts of metropolitan Adelaide. Both reservoirs experience problem algae blooms particularly cyanobacteria, and employ aeration and $CuSO_4$ dosing in an attempt to reduce the frequency and severity of bloom events.

Both reservoirs can be classified as eutrophic systems that require management to maintain water quality. Forsberg and Ryding (1980) recommend that when classifying trophic state, an approach be used that considers multiple variables. Three methods were used to classify Myponga and Happy Valley reservoirs, considering Chl-*a*, TP, PO_4, NO_3 and DO measurements, with all methods giving a eutrophic outcome (see Appendix A for calculations). However, Myponga was often only slightly over the eutrophic threshold, whereas Happy Valley was classified as distinctively eutrophic.

This study allows comparisons between Myponga and Happy Valley reservoirs as both are classified into the same lake ecosystem category, that being warm monomictic and eutrophic.

Aside from the obvious differences of size and volume (see Tab.3), Myponga and Happy Valley reservoirs also differ by the origins of water that they store. Unlike the remaining South Australian reservoirs, Myponga is entirely catchment fed and relies only on the Myponga River and runoff from the 124km^2 surrounding catchment. Happy Valley, on the other hand, is largely isolated from its surrounding natural catchment of only 6km^2, but is fed by controlled releases from Mount Bold Reservoir (an upstream impoundment of the Onkaparinga River) and also receives water pumped from the River Murray when required. Therefore, flow into Myponga reservoir is seasonal and consequently, most rainfall and resulting runoff and nutrient influx are received during winter and into spring, with potentially lower water levels and depleted nutrient concentrations in summer and autumn. Flow into Happy Valley, although most often would follow a similar seasonal pattern, can be altered by upstream releases of River Murray or Mount Bold Reservoir water and therefore can sometimes display patterns (with regard to nutrient levels, phytoplankton assemblages and abundances) dissimilar to what would be expected.

Table 1. Comparison of reservoir attributes

	Myponga	Happy Valley
Surface area	2.8km^2	1.88 km^2
Max. volume	26,800 ML	12,700ML
Max. depth	36m	19m
Mean depth	15m	6.8m
Catchment area	124km^2	6km^2

NOTE: This figure is included on page 53 of the print copy of the thesis held in the University of Adelaide Library.

Figure 15 - Locations of Myponga and Happy Valley reservoirs in South Australian (from SA Water Drinking Water Quality Report 2003-2004)

3.3.1 Water Quality Time-series Graphs of Myponga and Happy Valley reservoirs

Data is central to machine learning, data driven techniques used throughout the project, and results are influenced by the quality of data used. Therefore it is desirable to have high quality data, in a sufficient quantity, to gain the maximum benefits from the methods.

Interpolated data is displayed in the form of time-series graphs (Fig. 16-19) to reveal annual dynamics and long-term trends of both reservoirs, as well as allow comparisons between the two study sites.

3.2.2.1 Chl-*a*

The time series graph of Chl-*a* concentrations in Myponga and Happy Valley reservoirs (Fig. 16a) shows that both study sites experience similar levels, with Happy Valley having somewhat higher concentrations during extreme events. It appears that there is a slight increase in Chl-*a* concentration over time. As expected, Chl-*a* concentrations are maximal during the early part of each year, coinciding with warmer weather that stimulates algal growth.

3.2.2.2 Water temperature

Fig. 16b shows that both reservoirs have similar water temperatures over the period studied and exhibit expected seasonal fluctuations.

3.2.2.3 Turbidity

Myponga and Happy Valley reservoirs appeared to have similarly moderate turbidity up until 1997, where Happy Valley has become significantly more turbid with some extreme events giving high turbidity levels. Fig. 16c reveals that in Myponga reservoir, from 1997 onwards, the turbidity appears to have decreased and remained at low levels. The peak turbidity largely occurs mid-year.

3.2.2.4 Colour

Colour measurements in Fig. 16d show some interesting patterns particularly in Myponga reservoir throughout the 1990's. The reservoir is characteristically high in colour and levels seem to have slightly increased over time. Happy Valley has lower water colour but appears to also slightly increase toward the end of the period (2000's).

3.2.2.5 Total Phosphorus

The concentration of total Phosphorus in Myponga reservoir remains relatively steady throughout the study period, with maximal concentrations in the 2000's (Fig. 17a). Happy Valley had similar total Phosphorus conditions until the mid 1990's where concentrations began to increase significantly.

3.2.2.6 Phosphate

Higher phosphate levels were recorded for both reservoirs from the late 1990's onwards, with Myponga appearing to have slightly higher concentrations. Prior to that, Fig. 17b shows that

Myponga and Happy Valley recorded similar phosphate levels with peak values around 0.05mg/L. Concentrations appear to peak around August/ September annually.

3.2.2.7 Nitrate

The time series graph in Fig. 17c shows that nitrate concentrations are clearly much greater in Happy Valley reservoir, with highest levels being recorded in the 1990's. It also shows an apparent decreasing trend from the late 1990's in Happy Valley reservoir. Concentrations recorded in Myponga reservoir have largely remained stable, with a few years of noticeably lower concentrations (1993, 1994, 1997-1999, 2002). As with phosphate, nitrate concentrations appear to peak around August/ September.

3.2.2.8 Iron

Iron concentrations are shown in Fig. 17d to be highest in Myponga reservoir, with maximal concentrations from 1986-1993, followed by a drop in concentration levels. Happy Valley data shows a stable level of iron throughout the period. Annual peak iron concentrations are shown to occur at the end of a year and beginning of the next year, i.e. over the summer.

3.2.2.9 Manganese

Fig. 18a shows the time series graph for Manganese concentrations in both reservoirs and reveals that Myponga has the highest levels of the reservoirs. There is no clear long-term trend in the data for this reservoir, although concentrations are minimal for the first three years (1986-1988), from then onwards years of high and low concentrations seem to occur with no identifiable pattern. Happy Valley reservoir also recorded relatively low concentrations during the first few years of data (1990-1991, 1994-1996), increased to an extreme high level in 1997, then decreased to a level that remained consistent for the rest of the period. Annual peak concentrations are shown to occur around March-April.

3.2.2.10 Dissolved oxygen

Similar concentrations are shown for dissolved oxygen concentrations in Myponga and Happy Valley reservoirs (Fig. 18b). No long-term trend occurred as the concentrations remain at a consistent level for the entire duration of the studied period.

3.2.2.11 Conductivity

From limited conductivity data (Fig. 18c), Happy Valley reservoir is shown to have higher conductivity levels than Myponga reservoir, and has maximal values during the 2002-2003 period. Myponga data is too limited to exhibit any trends. Seasonal patterns are difficult to interpret, however the peak events in both reservoirs occurred over the 2002-2003 summer period.

3.2.2.12 *Anabaena*

Anabaena abundance is shown in Fig. 19a to reach higher levels in Myponga reservoir than Happy Valley. No distinct long-term trend can be seen for either reservoir, though seasonally it can be seen that peak *Anabaena* abundance occurs over the summer period each year. There is one major exception to this in Myponga reservoir (2001), where *Anabaena* abundance interestingly peaked in winter.

3.2.2.13 Green algae

Of the two green algae used in this study, *Scenedesmus* in Myponga reservoir is found in much greater abundances than *Dictyosphaerium* in Happy Valley reservoir (Fig. 19b). *Scenedesmus* does not present a significant long-term trend, instead there are periods of low abundance (1986-1991,1994-1996, 2000-2001) mixed with periods of high abundances (1992-1993, 1997-1999, 2002-2003). *Dictyosphaerium* does not appear to experience a change in peak abundance levels throughout the studied period. Seasonal patterns are inconsistent for both algae.

3.2.2.14 Diatoms

Cyclotella, a dominant diatom in Happy Valley reservoir, is shown to reach slightly greater abundances during peak events than *Nitzschia* in Myponga reservoir (Fig. 19c), though overall *Nitzschia* is more consistently high in abundance. *Cyclotella* has quite low populations for most years excluding 1989-1991, 1994-1995 and 2002. No long-term trends are obvious for either diatom, though seasonally *Nitzschia* can be seen to occur mostly between February and July.

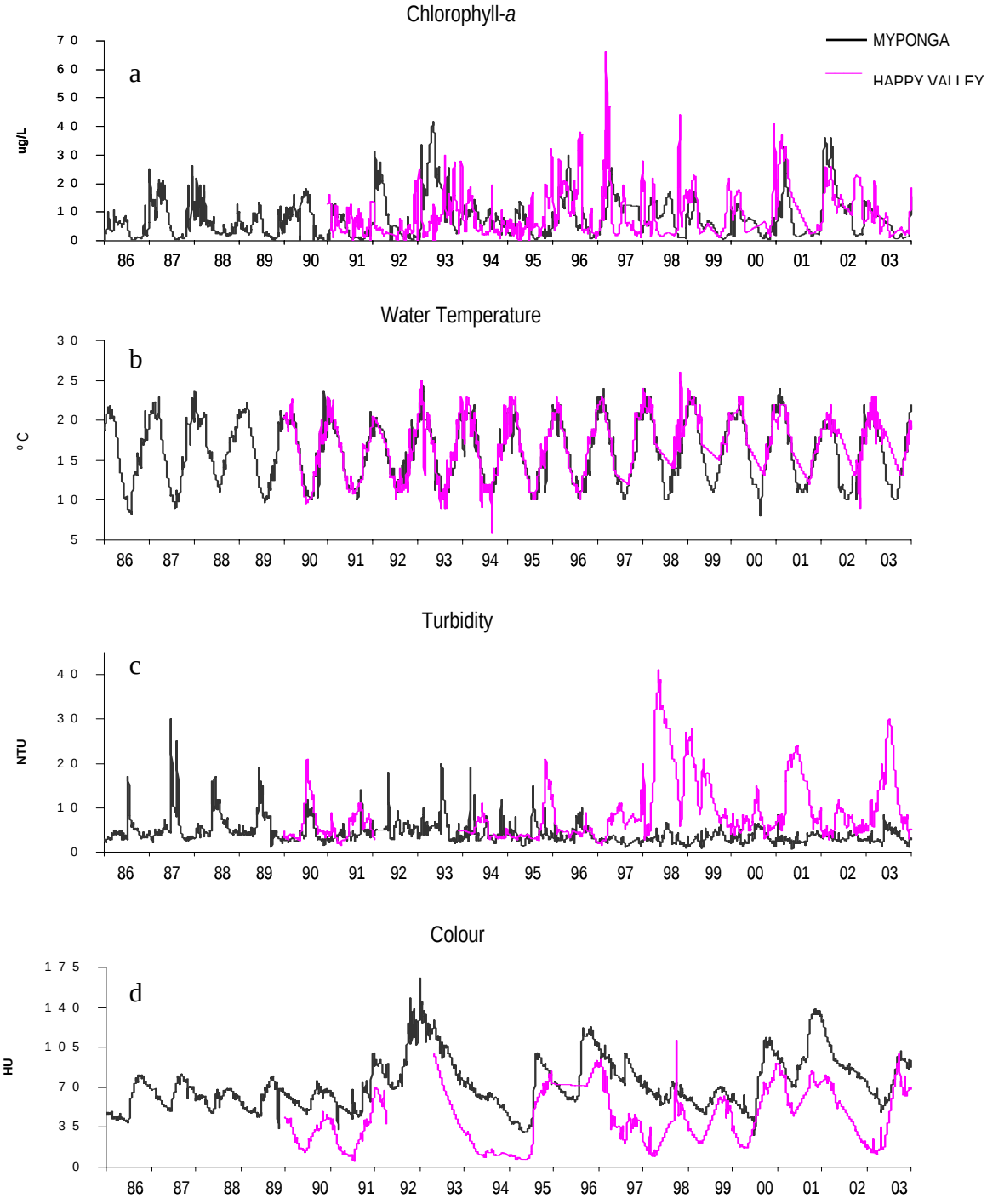


Figure 16. Time-series graph of preprocessed water quality data from Myponga and Happy Valley reservoirs, specifically a) Chl-a b) water temperature c) turbidity d) colour

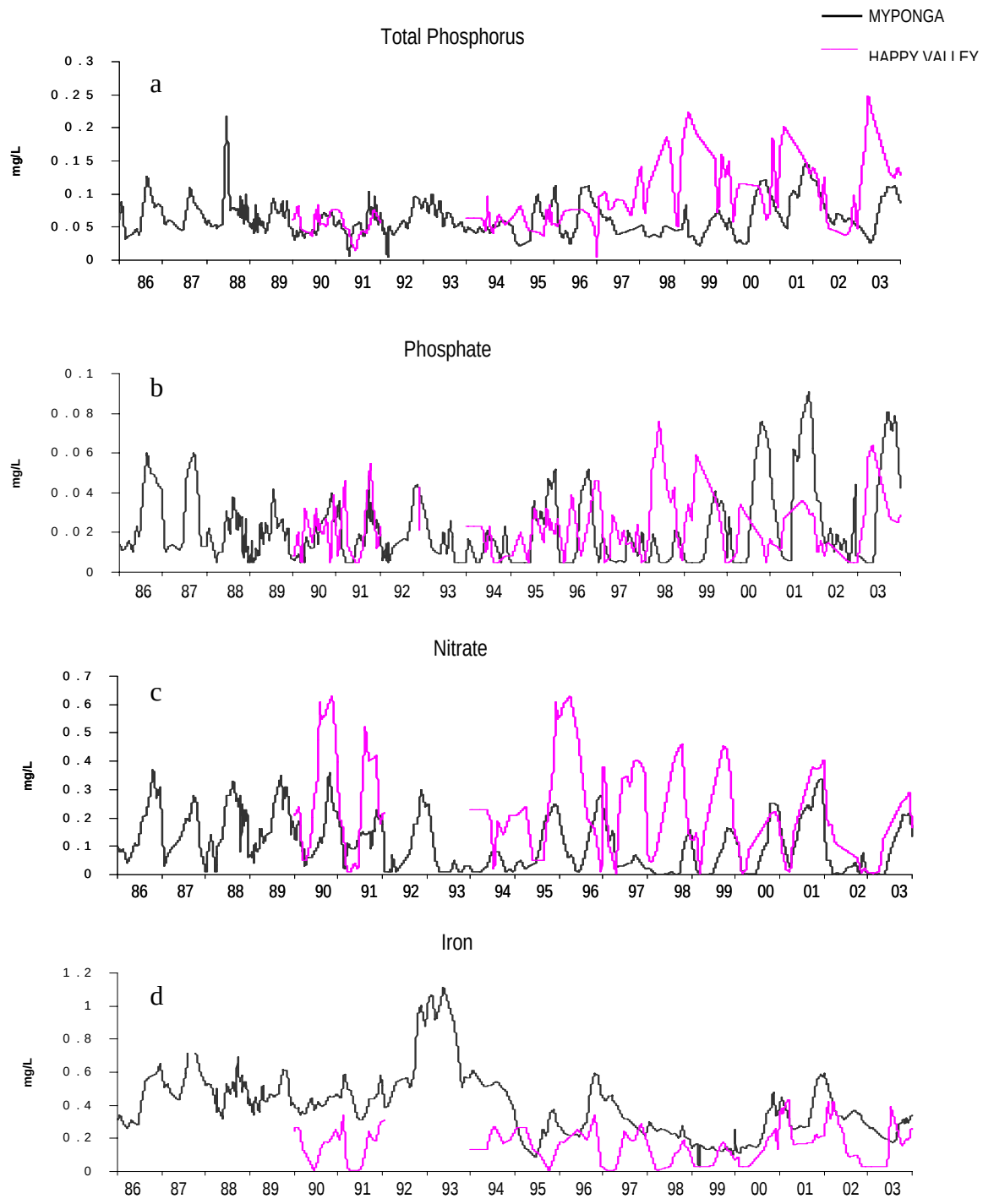


Figure 17. Time-series graph of preprocessed water quality data from Myponga and Happy Valley reservoirs, specifically a) total phosphorus b) phosphate c) nitrate d) iron

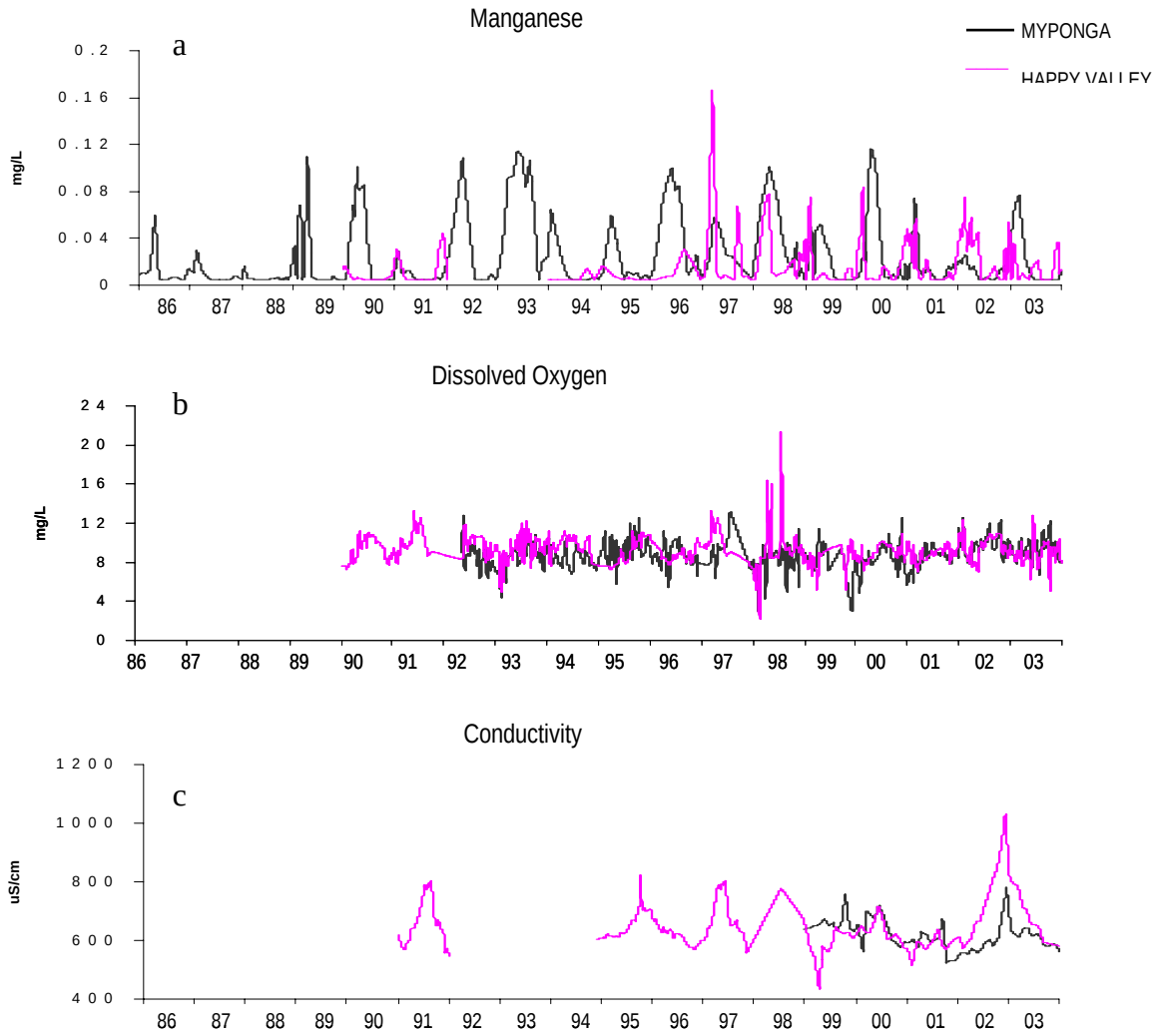


Figure 18. Time-series graph of preprocessed water quality data from Myponga and Happy Valley reservoirs, specifically a) manganese b) dissolved oxygen c) conductivity

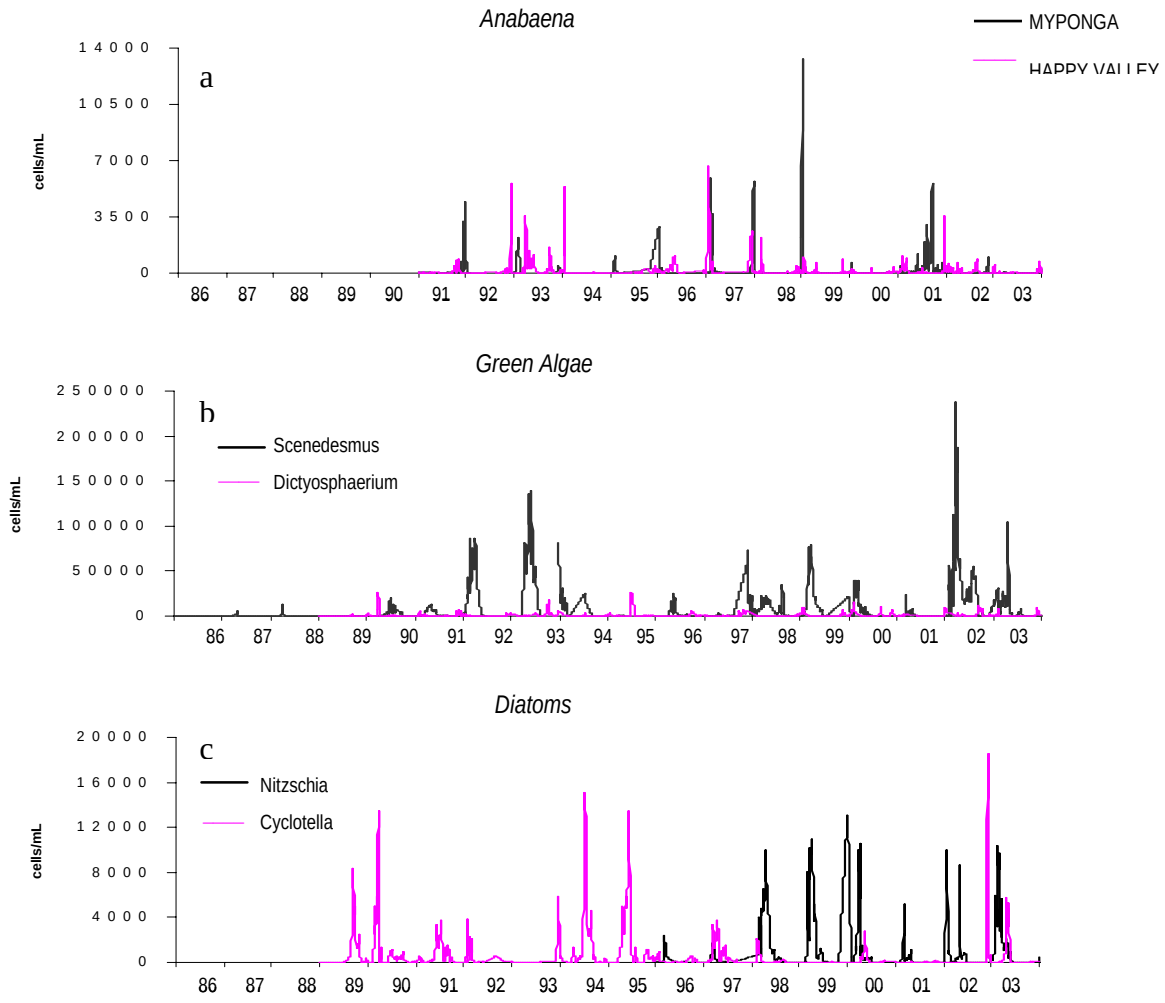


Figure 19. Time-series graph of preprocessed water quality data from Myponga and Happy Valley reservoirs, specifically a) Anabaena b) green algae c) diatoms

3.4 Hope Valley reservoir site and data summaries

3.4.1 Hope Valley reservoir site description

Real-time monitoring data from the Hope Valley reservoir was used as an independent data set for the case study aiming to test rule-based agents developed using Myponga and Happy Valley historical data (see Chapter 6). Hope Valley reservoir falls into the same lake ecosystem category as Myponga and Happy Valley reservoirs (Warm monomictic and slightly Eutrophic). It is located within metropolitan Adelaide and also used as drinking water storage. It is significantly smaller than either Myponga or Happy Valley reservoirs, with a maximum capacity of 3,700ML and water spread of 0.52 km^2.

3.4.2 Real-time data from Hope Valley reservoir

In order to test rule-based agents developed during this research with real-time data, a preliminary data set was collected (see Tab. 4) in real-time via online monitoring of Hope Valley reservoir (CRC for Water Quality and Treatment project no. 2.0.2.2.2.1). 160 days of data was available to be used for testing of the agents. As the real-time monitoring records measurements every 60 minutes, the data set initially encompassed overnight measurements. It was decided that, because the data used to develop the models was sampled between 10am and midday, it would be best to use an average of the measurements between these times for this application.

Table 4. Water quality data from Hope Valley reservoir

VARIABLE	ABBREVIATION	UNIT	USEABLE DATA	MEAN	MIN	MAX	STND. DEV.
Chlorophyll-*a*	Chl-*a*	ug/L	9/11/2006 – 20/4/2007	11.1	3.1	67.4	10.75
Water temperature	-	°C	9/11/2006 – 20/4/2007	21.5	17.2	25.4	2
Turbidity	-	NTU	9/11/2006 – 20/4/2007	4.2	0.9	10.3	1.77
Dissolved Oxygen	DO	mg/L	9/11/2006 – 20/4/2007	6.47	4.3	9.92	1.27
Conductivity	-	uS/cm	9/11/2006 – 20/4/2007	595.5	504	709	51.48

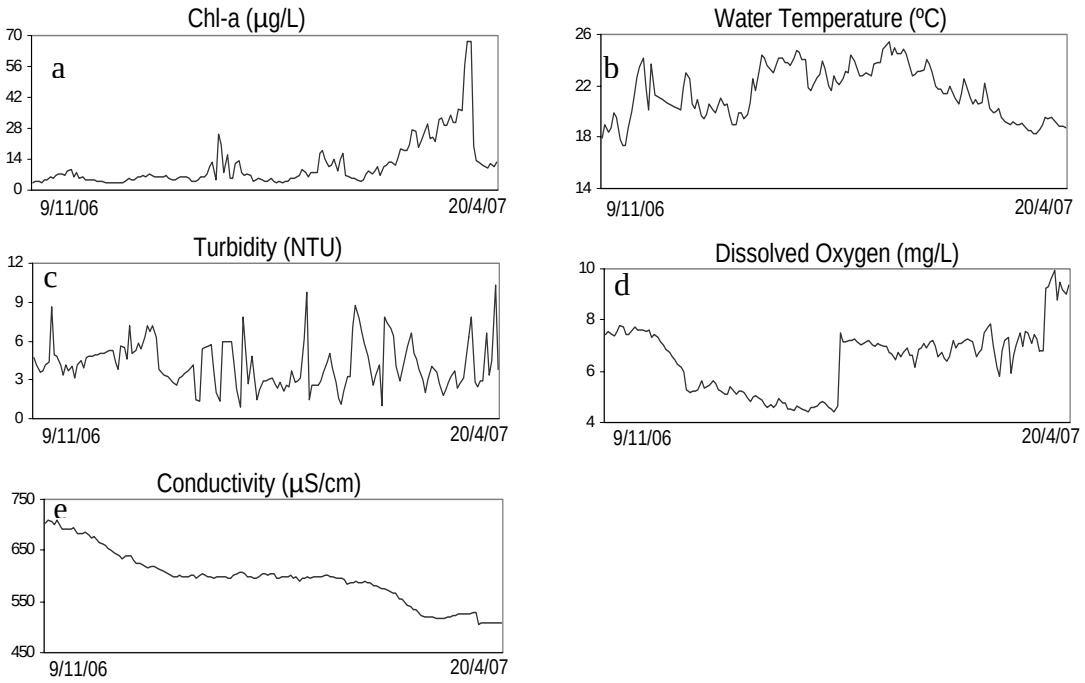


Figure 20. Time-series graphs of available real-time water quality data from Hope Valley reservoir

3.4.2.1 Chl-*a*

Chl-*a* concentrations in Hope Valley reservoir showed growth events on three occasions during the recorded period (Fig 20a). Small peaks are shown in January and again in February, however the major peak event (up to approximately 70µg/L) occurred in early-mid April.

3.4.2.2 Water temperature

Water temperature in Hope Valley reservoir is shown to start at mid-range temperatures of around 17ºC in early November to warm temperatures in the mid-20's over late December to early March, and then returns to mid-range temperatures (Fig.20b). Obviously the increase in water temperature coincides with the summer season.

3.4.2.3 Turbidity

Recorded turbidity values seem to be consistently higher at the beginning of the studied period (Fig.20c). The troughs reach lower levels from late December onwards, however the peak levels reach high values equal or above those in the earlier in the period.

3.4.2.4 Dissolved oxygen

Fig. 20d shows that dissolved oxygen dynamics are clearly much higher in the latter half of the studied period (around late January onwards), with maximal values achieved at the end of April.

3.4.2.5 Conductivity

Conductivity measurements show a decreasing trend throughout the studied period with recorded values reducing from approximately 700 to 500µS/cm (Fig. 20e).

3.5 Methods

3.5.1 Data Pre-processing

All historical data was pre-processed before modelling commenced. This lengthy process aims to clean the data and ensure that the data set is free of false or missing values and outliers, and is of a reliable quality. The regularity of the raw data was highly inconsistent, caused by different sampling regimes within each reservoir and for physical, chemical and biological water quality variables. To achieve equidistant intervals between all data, as required for the development of particular time-series models, the data was linearly interpolated to produce daily values and each month was standardised to 30 days. The use of interpolated data has, in the past, been a contentious issue that has been studied extensively in the statistical literature. More relevantly, recent studies have focused on the use of interpolated temporal data for time-series modelling using machine-learning methods. McKay et al. (2006) compared models trained with raw data and linearly interpolated data and validated them on the same test data set of raw measurements. It was found that the model trained with linearly interpolated time-series data actually performed better than the model trained with raw data. Cao (2006) also found that interpolated data did not increase error levels when used to develop HEA forecasting rule-sets, in fact, it gave better results than the raw data.

3.5.2 Recurrent Artificial Neural Network (RANN)

RANN (Pineda 1987) were used to forecast phytoplankton dynamics in both Myponga and Happy Valley reservoirs, using Chl-*a* data as an indicator of algal abundance or by using the specific abundance values of the main nuisance algae in each reservoir, *Anabaena*. Sensitivity analyses were then used to reveal quantitative relationships between the input variables and the output.

For the development and implementation of the SNN models, NeuroSolutions for Excel Version 4.2 was used.

3.5.2.1 Model design

Bowden (2003) formulated a methodology for the successful design and implementation of ANNs for the forecasting of water resource variables, designed to be of assistance to the many users of ANNs in the water industry who were not experts in the field computational modelling. It aims to harness the full capabilities of the ANNs by preventing the poor modelling practice due to unsystematic model development. The methodology is as follows: establish if ANNs is a better approach than more traditional statistical models; divide the available data into training and testing subsets for modeling purposes; decide on a suitable data transformation; determine optimal model inputs; choose type of ANN and architecture; select an appropriate performance measure and training of the network's weights {Bowden, 2003 #8}.

Bowden's methodology was considered as a guide, particularly when developing the SNN models in this project, though much of it also applied to HEA. It can be safely assumed that the databases used in this study contain non-linear data that is best suited to use by ANNs or machine learning methods rather than more conventional statistical models. Data division, input selection, training and testing methodology is described below, and data transformation was not thought to yield significant improvement for this particular application (Maier 2005) and therefore was not considered further. RANN was chosen for all SNN use in this project as it has proven to be particularly powerful for time-series modeling (see section 2.2.1.2) and is recommended by NeuroSolutions for more difficult temporal problems (NeuroDimension 2003). When building an SNN model numerous decisions have to be made regarding the network architecture. For example, number of hidden layers and number of processing elements in each layer, transfer function, learning rule and step size all need to be established. To clarify the role of some architectural components of the network, some definitions may be helpful. A processing element is a neuron-like unit that, together with many other processing elements, forms a neural network. The number of processing elements directly affects the overall computing power of the network. The transfer function is the component of a processing element through which the sum is passed (transformed) to create net output. Finally, the learning rule is used to calculate the weight updates. It specifies how weights adapt in response to a learning example (NeuralWare 1993). Some of these network parameters depend on the type and amount of input data; therefore, throughout the project slightly different architectures may have been used from model to model. However, some parameters were kept static throughout all experiments. All RANN were

designed with one hidden layer and used the Tanh Axon (hyperbolic) transfer function, with the Momentum learning rule set to 0.7. One hidden layer was considered sufficient for these problems, the TanhAxon is largely considered the non-linear axon of choice and was appropriate for this research and momentum was chosen as the learning rule although it is not the fastest learning rule, it is thought to be generally the most stable (NeuroDimension 2003). Different input layer structures are better suited to certain types of data and applications. A simple Axon structure was predominantly used throughout this work. The exact architecture of each model will be detailed in the relevant chapters.

3.5.2.2 Input selection

Optimal input variables can be determined in a number of ways including correlation analysis and expert knowledge. Data-driven methods are largely assumed to be able to identify critical model inputs, however it is best not to rely solely on the model to determine significant input, but to do some input refining before hand. Selection of the most relevant variables to include as model input can lead to reduced computational effort and easier to manage models (Schleiter et al. 2006).

In this study, no one specific method was used throughout the project to select the optimal model inputs, it was dependent upon the availability of data and the focus of the specific experiment. In some initial cases, there was very limited data and therefore, the variables that were available were used as inputs. In most cases, variables recognised in relevant literature to have significant relationships with phytoplankton were selected. In most cases, the potential input variables were first applied to KANN so that obvious relationships and interactions could be seen. Often a 'trial and error' technique was applied to find the best inputs, where RANN would be run with all potential inputs and then a sensitivity analysis would be carried out and inputs that had little to no influence on the output were excluded from the next run.

3.5.2.3 Data division, training and testing

A RANN requires two sets of data - one for training and the other for testing. Data set division is important and each sub set must be representative of the same population, though they do not have to be the same size. It is important that the test data be independent from the training data, so that the model can be assessed by its performance on new data. Power (1993) states that the testing of a model on new or independent data is clearly the most robust method for testing models predictive capabilities.

Split-sample validation (see Weiss and Kulikowski, 1991) was used for RANN models where a subset of the data is kept separate for testing the model, whilst the remaining data is used for training. This method was selected as it allows testing of the model on an independent data set and can be considered as a way of simulating real-time situations i.e. by presenting the model with unseen data. This approach restricts the amount of data available for training and testing, which can be a disadvantage when the data set is small, but in this case there is enough data for it not to be of concern. It was considered more appropriate than cross-validation techniques such as bootstrapping and leave-one-out, which are particularly useful for small and/or spatial data sets, but require the model to be retrained numerous times (Recknagel & Cao 2007 (in press)). Depending on the experiment, one or two years of data were selected to be left out of the training process so that the model could be tested on independent data. Selection of testing years was based on challenging the model with an event that it was designed to predict i.e. the year must contain an algal bloom. If two test years were being used they should be different in nature, further challenging the models applicability to different conditions.

Length of training necessary for RANN models was mostly discovered using the 'trial and error' technique. Criteria upon which ideal training time was determined include reasonably accurate results when the model is tested on the training data (extremely good training results often demonstrate that the network has been overtrained which results in poor forecasting abilities, whilst poor training results likely demonstrate that the network is yet to be trained sufficiently), and testing results within an acceptable error range when the network is applied to unseen data.

3.5.2.4 Performance measures

Validation or testing of a model is necessary for model acceptance and aims to establish two things: whether the model is acceptable for its intended purpose, and how much confidence can be placed in the models results and applications in the real world (Rykiel Jr 1996). Validation techniques are many and varied, and there is a lack of agreement upon which, if any, is best. Therefore, when possible, it is best to use numerous validation techniques (Mayer & Butler 1993; Power 1993). Validation techniques can be grouped into four categories; subjective assessment or face validity, visual techniques, deviance measures and statistical tests (Mayer & Butler 1993). For RANN models developed throughout this project, one method from each category has been selected for use. Subjective or face validity involves experts or knowledgeable people in the field assessing whether the model and its results are reasonable. For this study, numerous people in the field of freshwater ecology and management, and computational modeling assess the model before its selection for inclusion in this thesis. Visual techniques include the use of graphical

displays, in this case, a time series plot of observed data vs predicted data. This is considered a useful and informative method of data presentation (Mayer & Butler 1993; Power 1993)). Root Mean Square Error (RMSE), a deviance measure, was also used and finally, r^2 values demonstrating the goodness of fit, were selected as an appropriate and well accepted statistical performance measure for this type of modeling (Mayer & Butler 1993).

For forecasting models, the performance criterion is predictive accuracy. RMSE and r^2 values are very common prediction accuracy measures used in machine learning but it is also recognised that visual inspection of fit and knowledge of the modeled system is still essential in assessing a model. For example, a model that predicts an algal bloom early is a more useful model than one that predicts an algal bloom late, yet the error measures do not take this into account.

3.5.2.5 Sensitivity analyses

Sensitivity analyses can be useful for extracting information on the relationships between the input and output variables of an SNN. Throughout this research two types of sensitivity analyses will be used in association with the RANN models, they are 'most influencing parameter' (MIP) and 'sensitivity on wide-ranged disturbance' (SWD). Both are a product of the 'sensitivity about the mean' option offered by NeuroSolutions for Excel software. Using this method, all input variables, except the one of interest, are kept constant at their mean value. The variable being investigated is varied within the range of its mean by +/- 1 to 2 standard deviations. Jeong *et al.* (2001) cites Zar (1984) to explain that +/- 1 standard deviation represents commonly occurring variation, and +/- 2 standard deviation covers approximately 95% of total data variation. Therefore, sensitivity analyses with +/- 1 standard deviation encompass the general conditions of the water body, whereas sensitivity analysis with +/- 2 standard deviations will include specific and infrequent events and interactions. MIP reveals which inputs the output is most sensitive to. SWD allows further examination of specific input-output relationships by plotting the behaviour of the output variable in response to variations of an input across 2 standard deviations, above and below the mean. Resulting graphs of the input – output relationships over the range of the varied inputs demonstrate how the changing environmental parameters impact upon the output. Although these methods are very useful in extracting information that, in conjunction with ecological knowledge, can give good insight into relationships between input and output variables, it must be acknowledged that the methods have limitation. They only consider the relationship between a single input and a specific output. In ecology, relationships are often not isolated in such a way and are rarely as simple, more often variables are interrelated and

complex. This is why results of sensitivity analyses should be interpreted with knowledge of the system and bearing in mind that other factors may be related and influencing the demonstrated relationship.

3.5.3 Hybrid Evolutionary Algorithm (HEA)

HEA (Cao et al. 2006b) were used with a similar objective as the RANN models; to forecast phytoplankton dynamics in both Myponga and Happy Valley reservoirs, using Chl-*a* data as an indicator of algal abundance or by using the specific abundance values of the dominant nuisance algae in each reservoir. However, in addition to forecasts, the actual rule that has been discovered from the historical data to produce the forecasts is given as output. This allows further insight into the processes and relationships involved in the modelled system.
All HEA experiments were performed on a Hydra supercomputer (IBM eServer 1350 Linux) using the C programming language.

3.5.3.1 Model design

Data division, input selection and training and testing methodology are discussed in detail below. Throughout this study all HEA models began with an initial population of 200 (*N*), with a maximum of 100 generations (MAXGEN) and repetitive runs of 50 or 100. A detailed description of the algorithm was given in section 2.3.1, Fig. 10. The parameter settings of the HEA for rule-set discovery are shown in Fig. 21, where FL = the logic function set, FC = the comparison function set, FA = the arithmetic function set, MAXK = the maximum size of the rule set, DIF and DTHEN/ELSE = the maximum tree depth for the IF and THEN/ELSE trees respectively. For parameter optimisation by GA, M = the number of individuals randomly selected from the old population and MAX is the number of iterations comparing the fitness of new individuals to the worst individual from the old population.

Structure Optimization (GP)	$N = 200$ $F_L = \{AND, OR\}$ $F_C = \{>, <, \geq, \leq\}$ $F_A = \{+, -, *, /, \exp, \ln\}$ $MAXK = 4$ $D_{IF} = D_{THEN/ELSE} = 4$ $MAXGEN = 100$
Parameter Optimization (GA)	popsize = 50 $a = -0.5$ $b = 1.5$ $M = 8$ $MAX = 500$

Figure 21. Parameter settings of HEA for rule set discovery (from Cao et al, 2006)

Although the main outcome of the HEA applications was forecasting rule-sets, Chapter 6 suggests a procedure developed to extract a single rule-based agent for forecasting of Chl-*a* concentrations in both study sites and, potentially, other sites within the same lake ecosystem category.

3.5.3.2 Input selection

Data-driven methods are largely assumed to be able to identify critical model inputs, and this is particularly true for EA applications. If we assume this is the case, then it should be acceptable to make all potential inputs available to the algorithm, which should disregard the unimportant variables. Though in most experiments throughout this research only variables recognised in relevant literature to have significant relationships with phytoplankton were supplied to the algorithm for consideration. Some HEA experiments within this research considered only input variables that can be electronically measured and relayed to a computer via a data logger and telemetry system, thus simulating a real-time forecasting situation.

3.5.3.3 Data division, training and testing

Initial HEA experiments used the split-sample method for data division and model training/ testing. This method allows testing of the model on an independent data set and is considered as a way to simulate real-time situations i.e. by presenting the model with unseen data. Similarly to RANN applications, the experiments used either one or two years of independent data for model validation. Using the split-sample method for HEA also allowed for comparisons between the two forecasting methods, as it was used for RANN experiments also.

For later HEA experiments, *k*-fold cross-validation was used in a series of experiments using a reduced data set of only electronically measurable inputs. *K*-fold testing requires that the data be split into *k* equal sized partitions, in this case, year long blocks of data have been used. One of these sets is used as a test data set, whilst the others form the training data set. This process is repeated so that each set is used for testing once, and measures of performance are gained for each test. The overall performance is then based on an average of these *k* test data sets (Fielding 1999a). This method allows all data to be used for training at some point, important for smaller data sets, and can assess the models capabilities over the entire data set.

3.5.2.4 Performance measures

HEA models in this research are subject to the same performance measures as RANN models (see 3.5.2.4 for more details). One method from each of the following four categories has been

selected for use: subjective assessment or face validity, visual techniques, deviance measures and statistical tests (Mayer & Butler 1993). Model and result assessment by knowledgeable people in the field, time series plots of observed data vs predicted data, RMSE and finally, r^2 values satisfy this criteria.

3.5.2.5 Sensitivity analyses

HEA models are subject to two types of sensitivity analysis, for explanation of input-output relationships. SWD is used as well as a measure of MIP, however this MIP analysis is based on the number of times each input is selected for inclusion into a predictive rule set. The HEA is considered to be able to identify which inputs are necessary for the accurate forecasting of the target variable, therefore the inputs that have been most often included in rule sets are thought to be the most influencing parameters.

3.5.4 Kohonen Artificial Neural Networks (KANN)

KANN (Kohonen 1984) were chosen as the type of NSNN to be used throughout this project for the purpose of ordination, clustering and visualisation of water quality and algal abundance data with regard to seasons, water temperature, ranges of nutrients and artificial mixing strategies. The KANN were built and visualised by the SOM toolbox with Matlab Version 6.5.1.

3.5.4.1 Model design

The features or patterns found in the input data are expressed by Euclidean distances, which are calculated between inputs and can be visualised as a unified distance matrix (U-matrix) or as a partitioned map (k-means). The U-matrix visualises the relative distances between neighbouring data in the input data space using the colour spectrum, with the blue areas demonstrate neighbouring data with small distances and belonging to a cluster, whilst red areas indicate larger distances between neighbouring data and suggest borders between clusters. The k-means algorithm, is the most common partitative algorithm and simply partitions the input data into a specified number of clusters based on the U-matrix (Recknagel et al. 2006b). Component planes map each input variable relative to the classification criteria. Figure 22 shows the various outputs of KANN. To interpret the results, compare the component plane to the k-means map to see the relationship between the input variable (Chl-*a*) and classification criteria (seasons), i.e. the input variable Chl-*a* is highest in early and late summer.

The method was applied in a number of ways to examine features within the data by clustering with regard to seasons to show seasonal dynamics, different nutrient and water temperature ranges to show habitat preferences of algae and the novel application of dividing the data into distinct management periods, to examine changes in water quality parameters over differing management regimes.

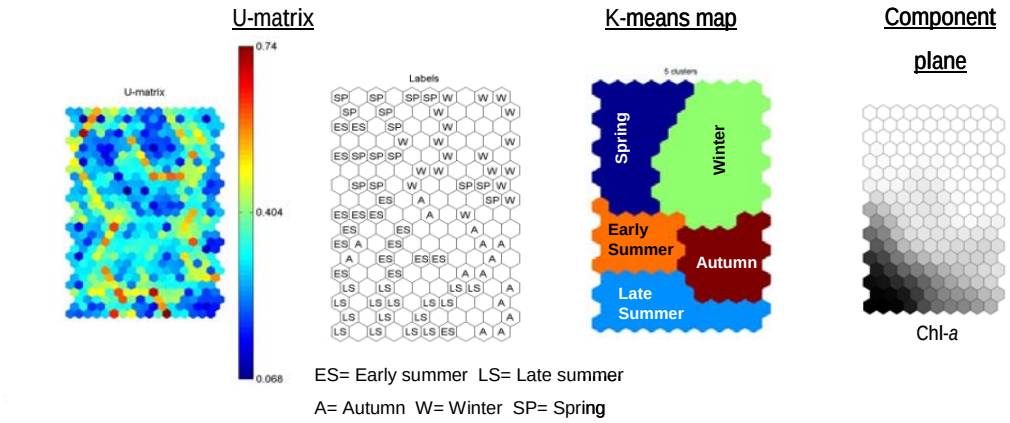


Figure 22 - Seasonal patterns visualised as a U-matrix, a k-means map and a component plane.

3.2.4.2 Input selection

Input selection is not as critically important for KANN models as it is with RANN and HEA. With RANN and HEA, the performance of the model depends on the relationship between input and output variables, for example, inputs need to be variables that drive or control algal blooms or the model won't predict well. Whereas with KANN, the inputs do not all necessarily have to be linked to the clustering criteria, if they have no relationship this will simply show up in the cluster results. This is why KANN is a good starting point for data sets where little is known about the relationships between the variables. Simply enter all variables as input and find whether there are distinct relationships or not. Of course, to examine particular relationships or occurrences it is best to prune the input variables appropriately to reduce noise. This project followed that approach by beginning with examinations of seasonal patterns across all data and refining experiments down to investigations of particular conditions, using only relevant inputs.

3.5.4.3 Data division, training and testing

Division of data into training and testing subsets is not required in this instance, as all data is used in a single data set for the production of KANN. However, data must be labelled relative to the purpose of ordination i.e. a classification criteria, for example when investigating the seasonality of the input variables, all input values are labelled and classified by a season.

3.5.4.4 Performance measures

Performance of a KANN model was measured by both the quantization and topographic error values. The quantization error measures map resolution by averaging the distance from each data vector to its best matching unit (BMU), and the topograohic error measures topology preservation by the percentage of data vectors for which the BMU and the second BMU are not neighbouring map units. Again, a visual assessment of how well the classification data was grouped into clusters can be useful when assessing the performance of a KANN, when the k-means clustering algorithm has been used.

3.6 Method integration

This project applied three data driven, machine learning techniques to water quality data sets from two South Australian drinking water reservoirs. It aimed to better understand the dominant algal populations by analysing temporal dynamics and management impacts, and to provide forecasting models for Chl-*a* and algal groups. It was thought that an integrated approach would allow broad examination of the data and provide comprehensive results. Fig. 23 shows the key steps involved in this approach. KANNs were used to discover and visualise relationships in the data set with regard to seasonality, habitat preferences and management impacts. Results from KANN provided information that aided input selection for the forecasting models. RANNs were developed to forecast Chl-*a* and specific algal populations and sensitivity analysis further explained input-output relationships and were even combined with KANN results to express relationships both qualitatively and quantitatively. HEA was also used for forecasting and sensitivity analysis and importantly, provided tangible model representations, in the form of predictive rule-sets that could potentially be applied to other similar water bodies. These methods combined provide explanation and forecasting of algal dynamics within the two studied reservoir systems.

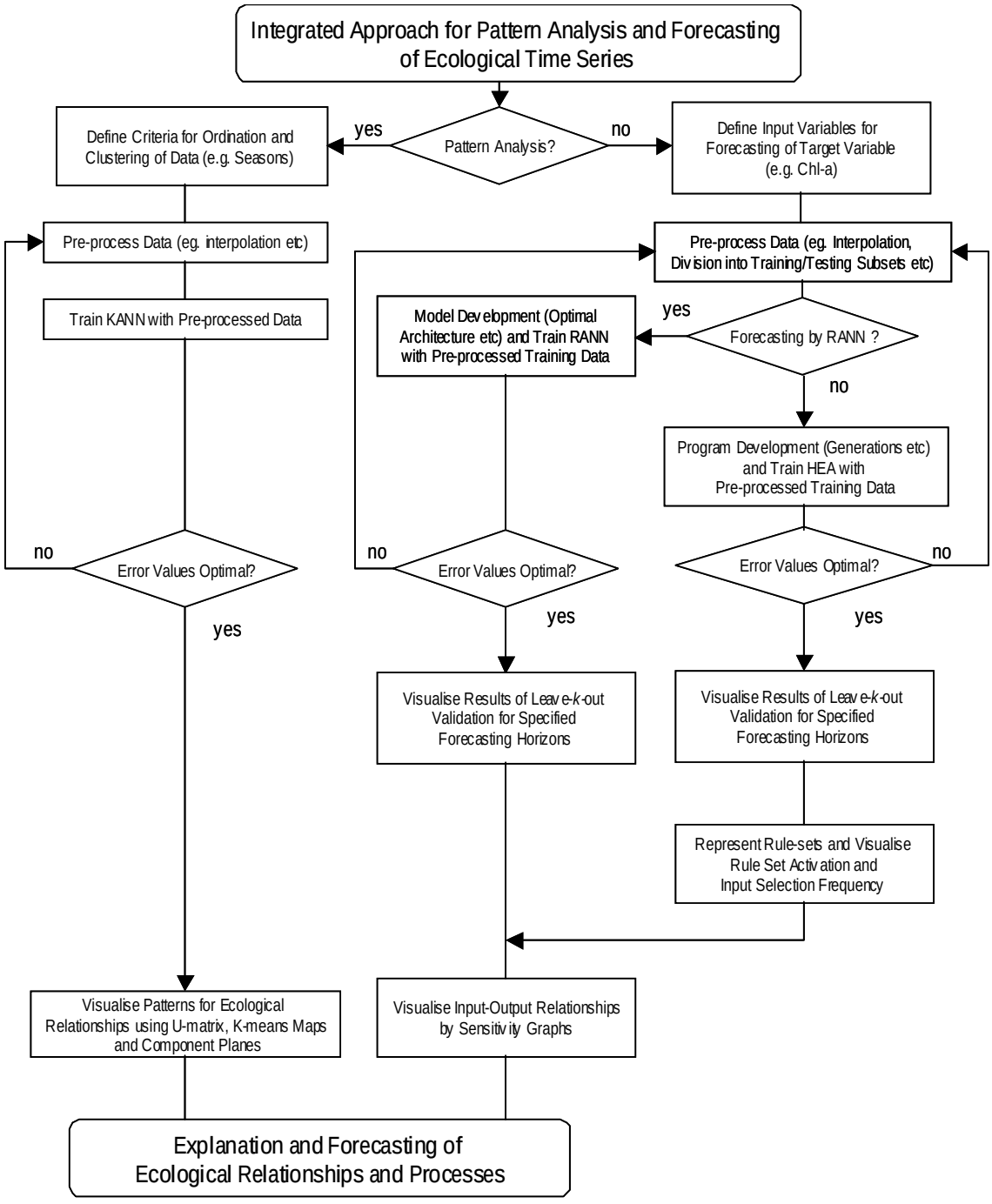


Figure 23. Framework for the integrated approach using KANN, RANN and HEA for explanation and forecasting of algal dynamics in Myponga and Happy Valley reservoirs

4. ORDINATION AND CLUSTERING OF WATER QUALITY VARIABLES

4.1 Introduction

Ordination and clustering of ecological data can reveal significant patterns and features. Traditionally, multivariate statistics such as Principal Component Analysis (PCA) have been used for this purpose. Van Tongren et al. (1992) used canonical ordination techniques of PCA and a hybrid of PCA and Redundancy analysis (RDA) on 8 years of data from the Loosdrecht lakes. The aim of the research was to analyse the trends and relationships between environmental factors and the planktonic communities. The study was able to reveal relationships between the variables within the data set, particularly the influence of temperature on the annual cycle of environmental variables, as well as on the phytoplankton/zooplankton assemblages and densities. Varis (1989) successfully used canonical correlation analysis to examine the influence of nutrient loading on the phytoplankton community in Lake Tuusulanjarvi, with particular attention to cyanobacteria. Cyanobacteria was shown to have very strong positive correlations with phosphorous and temperature and a negative correlation with dissolved inorganic nitrogen and dissolved inorganic phosphorous ratio.

There are many accounts of statistical analyses on communities using conventional multivariate analyses. However, conventional statistical methods are mainly limited to linear data and are not flexible in many aspects, for instance, data handling (Chon et al. 2006). Ball et al. (2000) discusses the poor performance of statistical methods for data modelling and Chon (1996) compares KANN to numerous statistical methods, finding KANN to be superior for classifying data. Overall, KANN can cope better than conventional techniques with the complexity and non-linearity involved with ecological data (Boddy & Morris 1999; Jongman *et al.* 1987).

In this research, KANN allowed an initial exploration of the data sets before they were used for the development of forecasting models. This permitted the identification of important relationships and potential input variables for later modelling applications.

Throughout this chapter, KANN were applied to time series water quality data from Myponga and Happy Valley reservoirs. Broadly, the aims of this chapter were to examine relationships between physical, chemical and biological factors within the reservoirs and to uncover short-term

seasonal patterns and long-term dynamics, the latter being specifically in context of management regimes.

Initially, the focus was on the seasonality of major water quality variables and their interactions with each other, as well as comparisons of patterns between the two study sites. Then KANN was applied to the more complex task of examining water quality changes between periods of different management regimes in Myponga reservoir. Throughout the long-term databases used for the study, both Myponga and Happy Valley reservoirs have attempted to control nuisance algal populations using various methods and levels of artificial mixing and aeration, in conjunction with $CuSO_4$ dosing (see sections 2.5.1 and 2.5.2 for more details). The experiment demonstrated seasonal shifts as well as changed magnitudes of nutrients, metals and biological components, in response to periods of stratification and various methods and intensities of artificial mixing. Finally, KANN was applied to determine at which level specific water quality factors seem to favour the growth and maintenance of particular algal groups.

Numerous applications of KANN have examined seasonality of various limnological variables, however clustering regarding management periods and ranges of physical and chemical factors are new concepts.

4.2 Aims and Hypotheses

The aims of this research include:

- the examination of the seasonality of water quality variables;
- discovery of conditions within Myponga and Happy Valley reservoirs which are conducive to algal growth: in general by looking at Chl-*a*, and to different algal functional groups, specifically the dominant blue-green, diatom and green algae from each reservoir; and
- to examine the impacts of management actions.

These aims will allow the testing of the following hypotheses:

Hypothesis 1 – water quality parameters would maintain natural seasonal patterns and habitat preferences as established in general ecological theory i.e. nutrients highest in winter, lowest by late summer/autumn; blue-green algae highest abundance in late summer, diatoms prefer turbulent conditions (theoretically winter/spring) and green algae, being opportunistic, would fill the gaps between the other two functional groups.

Hypothesis 2- Continued and updated management throughout Myponga reservoir's history would gradually lead to improved water quality conditions.

4.3 Methods and Materials

4.3.1 Data

The maximum amount of data available for each experiment was used to ensure that the extracted patterns could be assumed as representative for the whole dataset. However, the fact that less relevant data was available for Happy Valley reservoir meant that, for merged experiments, some available data from Myponga reservoir had to be left out to keep the input data from each reservoir equal, in accordance with requirements of the modelling software.

Table 5. Data used for each experiment in this chapter.

Experiment	Years of data used
Short-term dynamics and seasonal patterns 4.4.1.1	1996-2003 from Myponga and Happy Valley reservoirs
Long-term dynamics and management regimes 4.4.2.1	1986-1987, 1988-1989, 1990-1991, 2000-2001 from Myponga reservoir
Relationship to water temperature 4.4.3.1	1996-2003 from Myponga and Happy Valley reservoirs
Relationship to PO_4 concentrations 4.4.3.2	1996-2003 from Myponga and Happy Valley reservoirs
Relationship to NO_3 concentrations 4.4.3.3	1996-2003 from Myponga and Happy Valley reservoirs

4.3.2 Model Design

The KANN models were developed and visualised using the SOM toolbox with Matlab Version 6.5.1. A detailed explanation of the modelling method and process was given in Chapter 3, Section 3.3.4.

For the last series of experiments in this chapter (4.4.3), data from both reservoirs has been merged and used as input. Initial comparative experiments, where data from each reservoir was kept separate (including 4.4.1.1. and others not shown), demonstrated that the seasonal patterns of water quality parameters in both reservoirs were mostly similar. It was considered best to use

the merged results for this series of experiments as they displayed the same information but were able to minimise and simplify the results section.

All KANN models in this chapter use the k-means algorithm to classify the data into pre-defined groups or ranges. Seasons, ranges of nutrient concentrations and different water temperature levels were used to demonstrate how water quality variables associated with such factors. The tables below specify the ranges used for each experiment.

Table 6. Classification criterion used in 4.4.1.1 and 4.4.2.1

Seasons	Periods
Summer	1 December – 30 February
Autumn	1 March – 30 May
Winter	1June – 30 August
Spring	1 September – 30 November

Table 7. Classification criterion used in 4.4.3.1, 4.4.3.2 and 4.4.3.3

Range	Myponga and Happy Valley reservoirs
Lowest range	Water temperature: <14°C PO_4: <0.02 mg/L NO_3: <0.1 mg/L
Mid-range	Water temperature: >14°C - <19°C PO_4: >0.02 mg/L- <0.05 mg/L NO_3: >0.1 mg/L - <0.2 mg/L
High range	Water temperature: >19°C PO_4: >0.05 mg/L NO_3: >0.2 mg/L

4.4 Results

4.4.1 Short term dynamics and seasonal patterns

Many factors determine the distribution and abundance of organisms in freshwater lakes. Observed plankton communities are always the product of interactions between abiotic and biotic factors (Gower 1980). The community of pelagic phytoplankton of lakes consists of a diverse assemblage of nearly all functional groups. Despite their differences with regard to taxonomy and physiological requirements, many functional groups coexist in the same water volume at the same time. However, phytoplankton communities are usually dominated at a time by one or a small number of species, mostly restricted to a fairly precise season. Dominant genera in phytoplankton assemblages change seasonally, as physical, chemical and biological conditions in the water body change (Wetzel 1983). Change in temperature and light, or the balance between sinking and resuspension contribute to the seasonality of the phytoplankton community (Sommer 1989). This results in continuous changes in both the abundance and composition of the communities over the year, occurring in a similar sequence each year in a given lake.

In monomictic water bodies, such as Myponga and Happy Valley reservoirs, the major physical and chemical changes caused by the formation and disruption of the thermocline are thought to be an extremely influential factor with regard to both water quality issues and phytoplankton dynamics. Many consider the physical forcing due to water movements as the major influence on the particular members of the phytoplankton community i.e. diatoms. The suspension and vertical transport of the algal cells within the euphotic zone, the variations in nutrient loading from sediments and inflow, and the fluctuations in light conditions experienced by circulating algal cells, are seen to be the key factors in the relationship between hydrodynamics and phytoplankton ecology (Capblanq & Catalan 1994).

Despite the inherent variability of ecological systems and the management imposed on both Myponga and Happy Valley reservoirs, seasonal dynamics were able to be extracted from the data allowing comparisons between reservoirs.

4.4.1.1 Seasonality of major water quality variables and algal functional groups using merged data from both reservoirs

Fig. 24 shows the ordination and clustering of the water quality variables with regard to seasons. To examine the seasonality of the major water quality variables, the k-means map (Fig. 24a) visualised the 4 seasons with the k-means algorithm defining the middle left section as summer, the bottom area as autumn, the middle right section as winter and the top area as spring. The quality of the clustering was acceptable, particularly considering the many and varied input data, with quantization and topographic errors at 0.29 and 0.047 respectively.

The ordination and clustering by KANN showed that the water temperature in both Myponga and Happy Valley reservoirs was highest in summer and autumn (Fig. 24b). Turbidity was greatest in winter and spring in the Myponga Reservoir and late autumn and winter in Happy Valley, with much higher peak values found in Happy Valley (Fig. 24c). The major cause of turbidity would be runoff and mixing in the water column in and around winter (see Appendix B for graph demonstrating peak rainfall and turbidity occurring nearly simultaneously in Myponga reservoir), and it is thought by management that algae do not contribute significantly to turbidity in Myponga reservoir (Burch 2005a). The substantially higher level of turbidity (26.6 NTU) found in Happy Valley may be caused, to some extent, by turbid water from the Murray River which is episodically pumped into the reservoir.

Colour was highest in spring in Myponga reservoir, whilst Happy Valley showed high colour in spring and summer, with Myponga having much higher colour (Fig. 24d). This is in accordance with previous knowledge, as Myponga is known to be characteristically highly coloured, though not overly turbid (Velzeboer et al. 1991), and has greater mean and max colour values than Happy Valley. Colour in Myponga reservoir is clearly driven by dissolved organic carbon (DOC) and both exhibit distinct seasonality with highs in spring (see Appendix C), after DOC inflow and accumulation during winter. Happy Valley does not seem to share the obvious causal relationship between DOC and colour. DOC levels in this reservoir show no clear pattern, possibly due to inflow from not only the catchment but also diverted river Murray water at different times through the year, and cannot be said to influence colour dynamics.

Both reservoirs can be classified as eutrophic, based on several trophic state calculations that considered Chl-*a*, total phosphorous, PO_4 and NO_3 concentrations (see Appendix A). Happy Valley is considered a nutrient rich system as it receives a high nutrient load from not only the surrounding catchment, but also the water from the Murray River that enters the system. Although nutrient concentrations can mostly be considered as similar between the reservoirs,

Myponga has shown slightly higher PO_4 levels, whilst NO_3 concentrations are generally much greater in Happy Valley.

Clear clusters were shown for peak PO_4 concentrations for both reservoirs (Fig. 24e), with levels clearly shown to be highest in spring in Myponga and late autumn and winter in Happy Valley and at concentrations of 0.0792mg/L and 0.0604mg/L respectively.

The KANN results showed that in Myponga reservoir, NO_3 concentrations clearly peaked in spring, whilst in Happy Valley NO_3 concentrations appeared to be dispersed throughout winter and spring with an accumulation in early summer, with significantly higher levels in Happy Valley (0.518mg/L) than Myponga (0.285mg/L) (Fig. 24f). Perhaps peak NO_3 concentrations are dispersed throughout several seasons in Happy Valley because of the affect of Mount Bold and River Murray water being pumped into it during summer and autumn. Whereas Myponga reservoir follows the seasonal pattern of rain and consequent runoff fuelling the nutrient influx largely in and around winter, with an accumulation by early spring; Happy Valley reservoir gets nutrient influx, not only in winter and spring as the seasons dictate, but also in summer and autumn due to addition Murray water added to the system because of peak water usage around that time. To summarise the difference in seasonal nutrient patterns between the reservoirs; Myponga simply experiences a natural seasonal cycle of rain and resulting major nutrient influx, but KANN results for Happy Valley are not reflective of solely natural seasonal patterns but also the interference of management through the addition of Murray water, increasing nutrient levels when not naturally expected.

The seasonal patterns of these major water quality variables predominantly control the seasonality of phytoplankton, according to their individual taxonomy and physiological requirements. This experiment considers Chl-*a* as a representation of algal growth, to make general conclusions regarding associations between water quality variables and phytoplankton growth. Chl-*a* concentrations for both reservoirs peaked in summer with significant levels carrying into autumn, with Happy Valley having slightly higher peak values of approximately 29ug/L (Fig. 24g). This is complies with data set summaries showing that Happy Valley has a slightly higher mean Chl-*a* value and a substantially greater maximum. These results would suggest that major phytoplankton growth in both reservoirs is a result of higher solar radiation and water temperatures, periods of water column stability and high nutrient loads at the beginning of the growth stage, which is later exhausted (with KANN showing nutrient depletion coinciding with high phytoplankton abundance). Chl-*a* is a good representative for general algal growth;

however, it is interesting and sometimes necessary to examine further by investigating different types of algae.

Three main algae functional groups were examined using the dominant genera found in each reservoir. For blue green algae or cyanobacteria, the dominant genus in both reservoirs was *Anabaena* (Fig. 25a). *Nitzschia* and *Cyclotella* were the dominant diatoms in Myponga and Happy Valley respectively (Fig. 25c). For green algae, *Scenedesmus* was dominant in Myponga and *Dictyosphaerium* in Happy Valley (Fig. 25b). In both reservoirs, *Anabaena* was found to prefer the summer season and into autumn, though in Myponga reservoir the higher abundances began a little earlier at the end of spring. This concurs with much literature stating that the combination of numerous conditions in summer provides the optimal growth environment for most cyanobacteria. *Anabaena* is characteristically found in summer, and a preference for high water temperatures and pH levels is thought to contribute to this (Reynolds 1984; Shapiro 1990). *Anabaena* is particularly associated with lakes that experience stratified conditions, as Myponga and Happy Valley occasionally do, particularly those species that can most take advantage of a stable water column. Especially those species that possess gas vesicles enabling them to move vertically through the water column; upward into the surface layer to maximise exposure to light and downward to enable nutrient uptake during times of depletion in the surface waters (Moss 1998). After the increased phytoplankton growth from late spring, nitrogen levels in the water are depleted by late summer, favouring blue-greens with the ability to fix atmospheric nitrogen (Sommer 1989), with *Anabaena* being one such genus. *Anabaena* abundance in Myponga reservoir is shown to be nearly double that in Happy Valley reservoir (762 and 458 cells/mL respectively). This could be explained by the NO_3 concentrations within the reservoirs. Fig. 24f shows that Myponga reservoir has approximately half the NO_3 concentration of Happy Valley reservoir. The lower levels of NO_3 in Myponga, provide *Anabaena* with a bigger competitive advantage in the reservoir, through their ability to fix atmospheric nitrogen; whilst the higher levels of NO_3 in Happy Valley reservoir reduce this competitive advantage and therefore the genera does not out compete others so successfully, keeping its populations smaller.

Both diatom genera were shown to favour autumn, which is an interesting result considering they are often found to frequent winter and spring. Turbulence is a requirement of all diatom genera, as their cells are surrounded by a hard outer casing composed of silica, making them too heavy to remain buoyant in the water column in calm, stable conditions (Harris 1986; Kalff 2002). Turbulence is required to transport the diatoms back to the surface layer of the water where they

can receive vital sunlight and nutrients. Diatoms are noted to be quite insensitive to pH change in water (Harris 1986), and whilst it has been suggested that there is preference to lower water temperatures, the main habitat requirement is turbulence and instability in the water column. The KANN results showing the peak abundances occurring in autumn can be explained by a combination of artificial and natural mixing. The mechanical mixers and aerators used to reduce phytoplankton growth in the reservoirs during summer and into autumn are supporting these populations by providing the turbulence they require to prevent sedimentation losses. Sommer (1989) cites experiments by Reynolds et al (1984), which have shown that artificial mixing in water bodies could prolong the presence of large diatoms populations. Then toward the end of the autumn season, when rain and wind induced turbulence would be more frequent in the lead up to winter, the diatoms would also be provided with a suitable environment to support large populations. Fig. 25c shows that Myponga reservoir generally supports larger populations of diatoms than Happy Valley reservoir (4880 and 988 cells/mL respectively). Myponga reservoir is more intensively mixed (having two mixers and an aerator, compared with one mixer and an aerator at Happy Valley reservoir) and therefore is likely to provide more suitable turbulent conditions than Happy Valley reservoir.

Both green algae, *Scenedesmus* and *Dictyosphaerium*, were found in highest abundance throughout summer and autumn in Myponga and Happy Valley respectively, although *Dictyosphaerium* appeared slightly earlier in the year, showing increasing levels in spring (Fig. 25b). Myponga was shown to support larger populations of *Scenedesmus* than *Dictyosphaerium* in Happy Valley reservoir. Green algae are considered to be very opportunistic and can grow rapidly under a variety of conditions. They are quite ubiquitous and can form a large component of any phytoplankton assemblage from spring through to autumn (Happey-Wood 1988). *Scenedesmus* is a non-motile green algae and therefore prefers some mixing of the water column but can maintain reasonably large populations during thermal stratification due to extremely rapid growth rates. Therefore, its growth during the summer and autumn periods in Myponga is not surprising as the artificial mixing in use during this period limits the duration of episodes of thermal stratification, thought it can still occur regularly.

This experiment shows a general following of natural seasonal patterns one would normally expect to occur in water bodies of this nature, although management practices do appear to have caused some departures from the expected patterns for some variables.

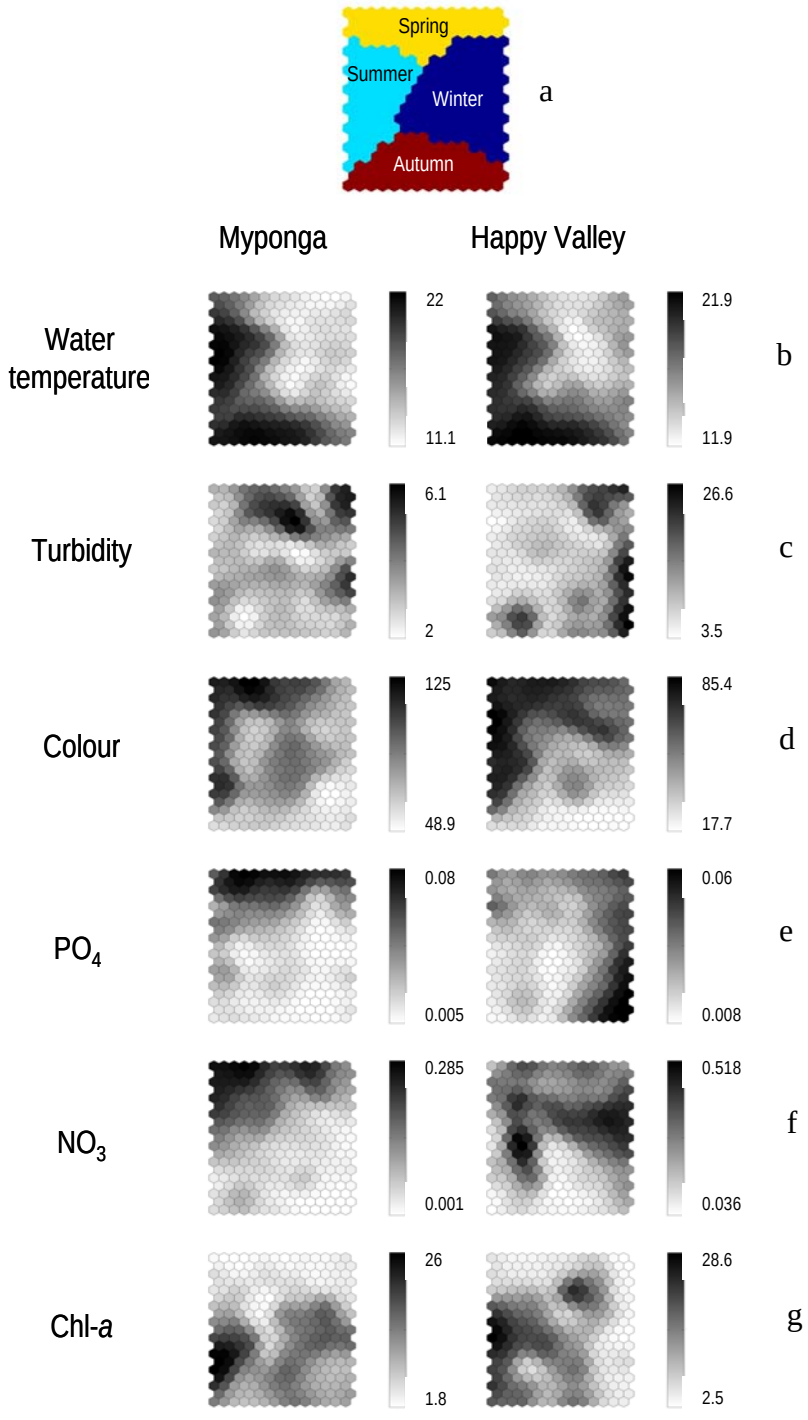


Figure 24. Ordination and clustering of main water quality variables by KANN with regard to seasonality

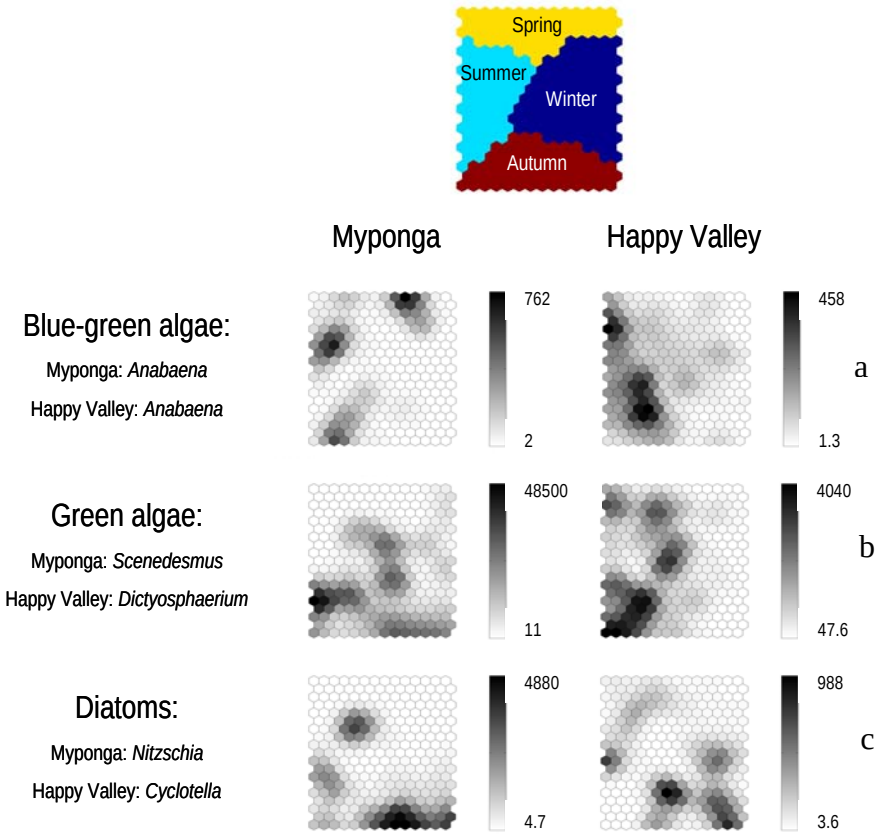


Figure 25. Ordination and clustering of dominant algal groups by KANN regarding seasonality

4.4.2 Long-term dynamics and management related patterns

Long-term trends in water quality factors can be the result of many influences on a reservoir or the rivers that feed it. There is no question though that direct, in-lake management of a reservoir is significantly powerful in shaping long-term changes and patterns.

The main use of a water body largely determines the type and extent of water management. Management practices can have a wide range of impacts on a water body and have the potential to dramatically change the state of the system.

As both Myponga and Happy Valley reservoirs are used to store water that will be supplied to households as potable water, they are very highly managed. Artificial mixing and $CuSO_4$ dosing have been used in both Myponga and Happy Valley reservoirs in an effort to control nuisance algal blooms, particularly the potentially toxic *Anabaena* known to cause taste and odour problems.

In the following section, both short-term seasonal dynamics and long-term patterns of water quality and phytoplankton parameters are brought into the context of various management periods and regimes.

4.4.2.1 Impacts of mixing regimes on major water quality variables using Myponga Reservoir data set

Ordination and clustering of four 2-years-periods of water quality time series, which differed in intensity and methods of artificial mixing, provided insights into effects of management on chemical and biological water quality properties. KANN, applied to this complex task of examining water quality changes between periods with different management regimes, indicated that both seasonal shifts as well as changed magnitudes of nutrients, metals and Chl-*a* occurred in response to periods with stratification and changed mixing strategies.

Fig. 26 visualises results of the seasonal ordination and clustering of major water quality variables in the Myponga reservoir for four periods with different mixing conditions. The k-means map visualised the four seasons (Fig. 26a), with the algorithm establishing the top left corner as summer, the top right corner as autumn, the bottom right corner as winter and the bottom left corner as spring. The quality of the clustering was good, particularly considering the many and varied input data, with quantization and topographic errors at 0.3 and 0.008 respectively.

As expected, the highest water temperature occurred in summer and autumn in all periods and obviously it is not a factor that can be changed by management (Fig. 26b). The results show that during the period from 1986 to 1987 (Fig. 26, left column), when no artificial mixing was used and the reservoir experienced thermal stratification, Chl-*a* was highest in summer and then autumn with a maximum of 15.9ug/L (Fig.26c). *Anabaena* reached highest abundance at 1320 cells/ml in autumn (Fig. 26d) and PO_4 concentration was greatest in winter and spring with a maximum of 0.057mg/L (Fig. 26e). The concentrations of Manganese (Mn), which also has implications for drinking water quality and is thought to encourage algal growth (Boney 1989), were high in summer and autumn with a maximum of 0.034mg/L (Fig. 26g).

During the period from 1988 to 1989, when 3 submersible mixers were used (Fig. 26, second column from left), the abundance of *Anabaena* was highest in spring with a much lower maximum of 852 cells/ml compared to the previous period with stratification. This result suggests that the introduction of artificial mixing succeeded in creating an unsuitable environment for the species, leading it to peak in a different season, at a level of half its abundance in the previous period

where no mixing was used. Although Chl-*a* was lower as well, it still peaked in autumn, and was obviously not caused by *Anabaena*. The lower *Anabaena* and Chl-*a* abundances in 1988 to 1989 were also reported by Velzeboer et al. (1991), with the conclusion that the mixers were successful in reducing total algal biomass. Interestingly, the results in Fig. 26e indicate a slight decrease of PO_4 concentrations in 1988 to 1989 compared to the previous period with stratification. This result may hint at lower internal PO_4 loading from anaerobic sediments since mixing aims to maintain aerobic conditions at the sediment. In contrast, the Mn concentrations increased in the same period and peaked in summer. During the period from 1990 to 1991 (Fig. 26, third column from left) an aerator was implemented in the Myponga reservoir, which resulted in an increase of Chl-*a* to similar concentrations as the period with stratification, but different regarding seasonality. Maximum concentrations were in autumn and winter whilst maintaining reasonable concentrations in summer. The PO_4 and Mn concentrations also increased slightly.

Finally, the period from 2000 to 2001 (Fig. 26, right column), where a combination of surface mixers and aerators was implemented, showed very similar patterns as the period with stratification because Chl-*a* peaked in summer and autumn, PO_4 peaked in spring and winter and Mn peaked in autumn, though all were at a higher concentrations than the initial period. Mn concentration consistently increased throughout the periods and peaked in autumn in all periods. This is possibly due to sediment release that may occur during the warmer seasons of summer and autumn, when stratification can potentially occur.

The aim of changing and updating management regimes is obviously to continue to improve water quality. Two significant issues in water quality management at Myponga reservoir are the discouragement of *Anabaena* growth and presence by creating an unsuitable environment for them; and the prevention or minimisation of internal loading, which will exacerbate the *Anabaena* problem and decrease water quality. Although the results show that *Anabaena* abundances have not increased over the examined period, they are still present and cause regular, troublesome blooms. Internal loading appears to have been halted with regard in some respects, but has had not success in controlling Mn, which is at a level where it must be treated for at the water treatment plant. For aesthetics, a maximum of 0.1mg/L is the guideline value given by the Australian Drinking Water Guidelines as taste and staining can occur at this concentration (Government 2004).

Some of the results produced by this experiment suggest that the management changes at Myponga Reservoir have not necessarily resulted in improved water quality across time and it certainly does not seem to be the case for all water quality variables.

This is in contrast to claims from several papers (see section 2.5.3) that artificial mixing had convincingly improved water quality at Myponga Reservoir. Some of these articles used data only up until the mid-90's and therefore have seen trends from a different perspective and came to different conclusions. For example, Brookes et al. (2000) and Lewis et al. (2003) claim that aeration by artificial destratification has lowered metal release from the sediment. However, the consistently increasing levels of Mn shown in Fig. 26g suggest that water quality, with regard to internal loading of metals, has not completely improved with artificial mixing. PO_4 and Chl-*a* are also shown to have higher maximums at the last period (2000-2001) than they did prior to the introduction of artificial mixing (1986-1987). In support of Brookes et al. (2000), is the reduction of Fe that can be observed throughout the years (Fig. 26h).

In summary, this experiment shows that there did appear to be an initial improvement in most water quality factors after the introduction of artificial mixing, however this trend did not necessarily continue. Important variables such as Chl-*a*, PO_4 and Mn were shown to be in higher concentrations after years of artificial mixing than they were prior to its inception, whilst NO_3 remained stable and Fe concentrations decreased. These results suggest that it could be appropriate to implement a new management strategy, perhaps as described in section 2.5.5.

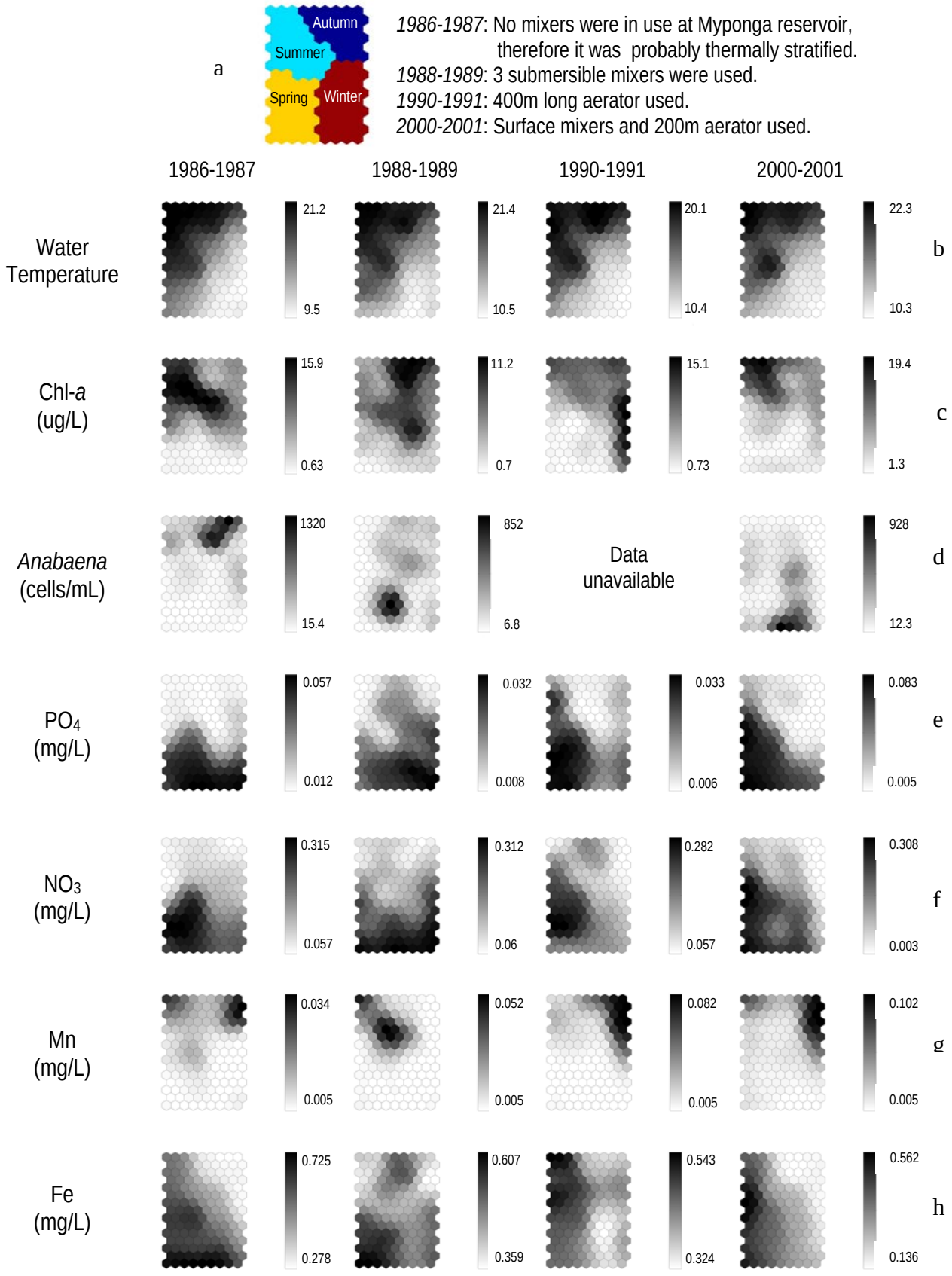


Figure 26. KANN, using k-means map, with corresponding component planes for major water quality variables clustered seasonally and separated into periods of different management at Myponga reservoir.

4.4.3 Habitat preferences established by clusters according to ranges of physical/chemical conditions using merged data from both reservoirs

To look at the relationship between the occurrence of different algae functional groups in relation to different physical and chemical habitat attributes, the dominant genera from each of blue-green, green algae and diatoms, from each reservoir were combined and clustered in comparison to 3 ranges of the specific attribute.

Whilst the word 'preference' is used here, it is sometimes more that the algal blooms coincide with a certain condition – not specifically prefer it. It is also acknowledged that it is an interconnected set of conditions that often promote and sustain algal growth and maintenance. However, this experiment offers a simplified insight into conditions that favour growth of particular algal groups, and potentially suggests a means of control of the phytoplankton by manipulating the water quality parameter to a level that is not conducive to the growth of nuisance species, and that encourages the presence of preferred species.

4.4.3.1 Occurrence of algae functional groups in relation to water temperature

Fig. 27 displays the results for the ordination and clustering of algal occurrence in relation to water temperature. The k-means map visualised the 3 clusters relating to the 3 different ranges of water temperature found in the data. The k-means algorithm identified the top cluster as related to low water temperatures of less than 14˚C, the bottom-right cluster as mid water temperatures of 14-19˚C and the bottom-left cluster as high water temperatures of above 19˚C (Fig. 27a). The k-means algorithm identified a clockwise transition of clusters with some slight overlapping areas at the borders of the clusters. The quality of the clustering was high, as the relatively small number of input variables reduced the likelihood of noise and messy clustering, giving low quantization and topographic errors of 0.07.

The component plane showing Chl-*a* concentration (Fig. 27b) clearly indicated that highest levels were found to coincide with mid water temperatures of 14-19˚C, likely to be caused by green algae and/or diatoms. *Anabaena* from both reservoirs were shown to clearly favour periods of high water temperature from 19° and above, and there is much literature to support this notion. The component plane for green algae showed it to be present in reasonable abundance

throughout all water temperature ranges, suggesting that it is not dependent on particular temperature conditions. Green algae are considered as opportunistic, and therefore it is likely that this genus proliferates whenever resources allow it – not simply under certain water temperature conditions (Happey-Wood 1988). Results for the occurrence of diatoms were interesting in that the component plane showed high abundances mainly in the high water temperature range. Often diatoms are associated with the cooler waters of winter and spring, though as mentioned in 4.4.1.1, this association with winter and spring is likely to be mostly related to the turbulence associated with these seasons rather than the water temperature. As artificial mixing, in the form of mechanical mixers and aerators, takes place throughout summer and autumn in Myponga and Happy Valley reservoirs, these warm water seasons can also provide the turbulence required to support diatoms populations. An extension of their suitability to summer and autumn is their superior nutrient competition abilities as discussed below in 4.4.3.2.

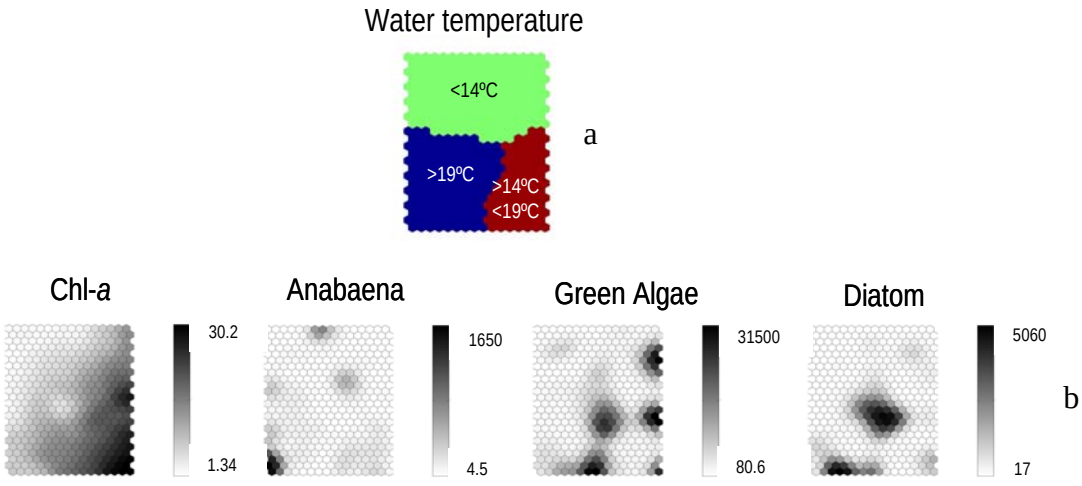


Figure 27. KANN, using k-means map showing water temperature ranges, and corresponding component planes for dominant algal functional groups in Myponga and Happy Valley reservoirs.

4.4.3.2 Occurrence of algae functional groups in relation to phosphate concentrations

Fig. 28 displays the results for the ordination and clustering of algal occurrence in relation to phosphate concentrations. The k-means map visualised the 3 clusters relating to the 3 different ranges of PO_4 concentrations found in the data. The k-means algorithm identified the top cluster as related to low PO_4 concentrations of less than 0.02mg/L, the middle cluster as mid PO_4 concentrations of 0.02-0.05mg/L and the bottom cluster as high PO_4 concentrations of above

0.05mg/L (Fig. 28a). The k-means algorithm identified some slight overlapping areas at the borders of the clusters. The quality of the clustering was high, as the relatively small number of input variables reduced the likelihood of noise and messy clustering, giving low quantization and topographic errors of 0.062 and 0.052 respectively.

The component plane showing Chl-*a* concentration (Fig. 28b) clearly showed that highest levels were found to coincide with low PO_4 concentrations of less than 0.02mg/L. As Chl-*a* is representative of algal abundance, it stands to reason that more Chl-*a* will result in greater uptake of PO_4, an essential nutrient for phytoplankton growth. Once phytoplankton growth and populations increase significantly, the PO_4 is quickly consumed, resulting in rapid depletion of the nutrient pool. This depletion is often compounded by the fact that phytoplankton can not only take what is needed of the nutrient for immediate utilisation, but can employ 'luxury uptake' of the nutrient to levels far exceeding current requirements (Reynolds et al. 1984). This 'luxury uptake', which further adds to the speed and magnitude of PO_4 depletion, allows cell growth to continue for some time after the nutrient levels in the water have become quite low.

Interestingly, clusters representing high *Anabaena* abundance correspond with areas of mid and high PO_4 concentrations. This is counter-intuitive, as you might expect peak populations to require more of the nutrient and therefore deplete the concentration. However, because *Anabaena* populations are normally prevented from reaching very high numbers for fear of blooms and toxicity issues, it is possible that the peak numbers are not often sufficient to impact upon the PO_4 concentrations. Component planes for both green algae and diatoms showed that highest abundances were reached at times of low PO_4 concentrations of less than 0.02mg/L. These genera reach much higher cells/mL levels than *Anabaena* and therefore, the PO_4 uptake required for the growth and maintenance of the peak populations would presumably deplete the nutrient levels. Interestingly, diatoms are recognised as particularly good competitors for nutrients, especially phosphorus (Sommer 1989), which also supports their populations through out the warmer, nutrient poor seasons in Myponga and Happy Valley reservoirs.

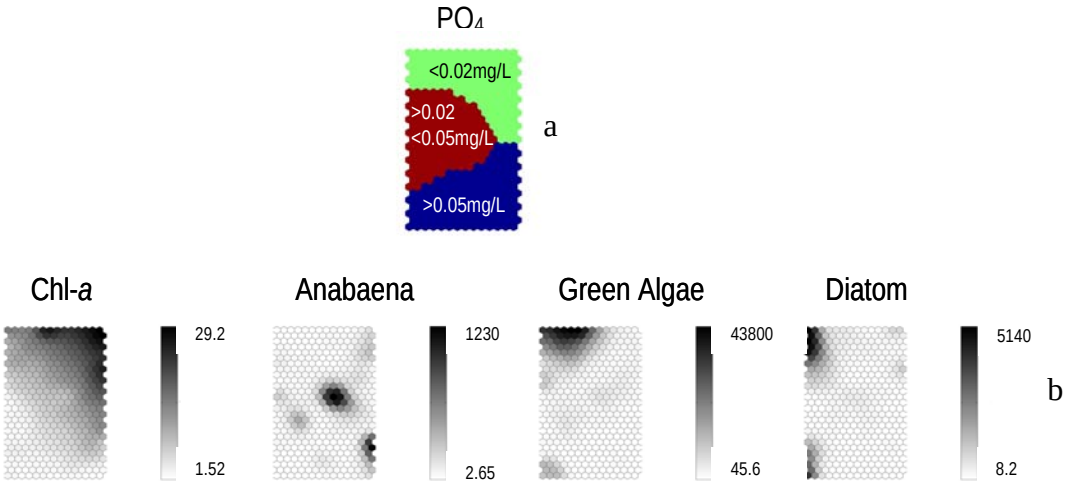


Figure 28. KANN, using k-means map showing PO_4 concentration ranges, and corresponding component planes for dominant algal functional groups in Myponga and Happy Valley reservoirs.

4.4.3.3 Occurrence of algae functional groups in relation to nitrate concentrations

Fig. 29 displays the ordination and clustering of algal occurrence in relation to nitrate ranges. The k-means map visualised 3 clusters relating to the 3 different ranges of NO_3 concentrations found in the data. The k-means algorithm identified the top cluster as related to low NO_3 concentrations of less than 0.1mg/L, the mid-left cluster as mid NO_3 concentrations of 0.1-0.2mg/L and the bottom cluster as high NO_3 concentrations of above 0.2mg/L (Fig. 29a). The k-means algorithm identified some slight overlapping areas at the borders of the clusters. The quality of the clustering was high, as the relatively small number of input variables reduced the likelihood of noise and messy clustering, giving low quantization and topographic errors of 0.058 and 0.056 respectively.

The component plane showing Chl-*a* concentration (Fig. 29b) clearly showed that highest levels were found to coincide with low NO_3 concentrations of less than 0.1mg/L. This result concurs with the seasonality of water quality factors as shown in 4.4.1.1, where Chl-*a* is shown to peak in summer, whilst NO_3 concentrations peaked in spring and then depleted rapidly in summer to the low levels which correspond with high Chl-*a* concentrations. As with PO_4, NO_3 is an essential nutrient for phytoplankton growth and therefore the more population growth and maintenance at any one time will require more uptake of the nutrient. However, not all phytoplankton are dependent on the NO_3 in the water column. Certain cyanobacteria, including *Anabaena*, are

known to possess to ability to fix atmospheric nitrogen and therefore do not have to rely on the NO_3 pool in the water (Shapiro 1990). The component plane for *Anabaena* showed that peak abundance corresponded with mid-high NO_3 concentration. Cyanobacteria as known to have high tolerance and even preference to excessive nutrient loading, which would explain the presence of *Anabaena* in high PO_4 and NO_3 concentrations (Paerl 1988). As with *Anabaena*'s relationship with PO_4 concentrations, although it would usually be expected that peak populations would require more uptake and therefore deplete nutrient reserves, the relatively low peak abundance values that this functional group is normally reduced to would prevent the depletion from occurring as it does with more abundance algal groups. Also, because *Anabaena* is able to fix atmospheric nitrogen, it may be that little nitrate uptake is occurring due to the use of nitrogen fixation, allowing the source pool to be maintained at a high level. Boney (1989) even states that in some habitats, filamentous cyanobacteria make significant contributions to in lake nitrogen budgets.

Component planes for both green algae and diatoms showed that highest abundances were reached at times of low NO_3 concentrations of less than 0.1mg/L. These genera are not capable of fixing atmospheric nitrogen for use and reach much higher population numbers than *Anabaena* and, as with PO_4 in section 4.4.3.2, NO_3 uptake required for the growth and maintenance of the peak populations would presumably deplete the nutrient levels due to reliance on the NO_3 in the water column. Happey-Wood (1988) supports this, stating that the growth of numerous types of green algae is commonly accompanied by substantial decreases in the inorganic nitrogen pool of the surface waters of lakes.

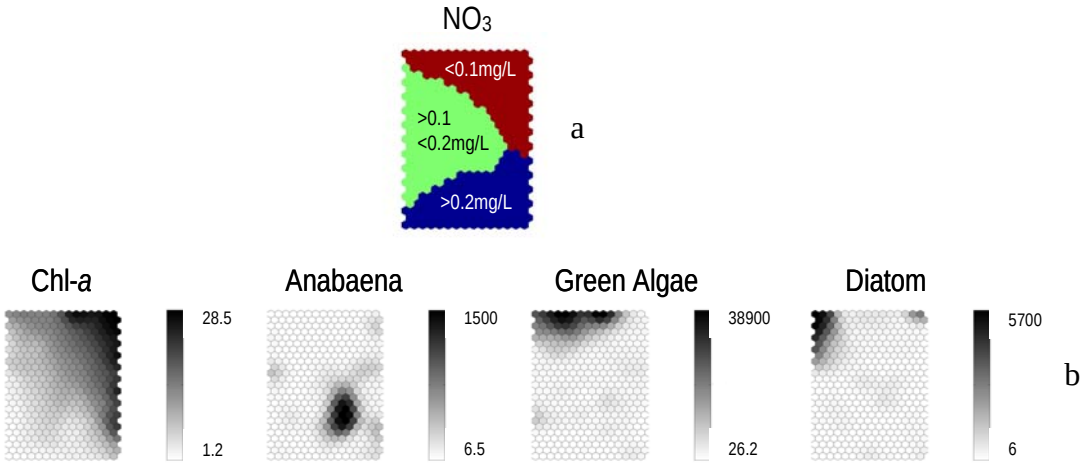


Figure 29. KANN, using k-means map showing NO_3 concentration ranges, and corresponding component planes for the dominant algal functional groups in Myponga and Happy Valley reservoirs.

4.5 Discussion

This chapter used KANN to examine physical, chemical and biological water quality data by ordination and clustering. Results offered insight into short-term dynamics, demonstrating seasonality and succession, and long-term trends in the context of management periods.

The seasonality of major water quality variables showed that, whilst many parameters followed the usual natural patterns for water bodies similar in nature to Myponga and Happy Valley reservoirs, there were some exceptions. Whilst many of the physical and chemical parameters behaved as expected such as water temperature being highest in summer and into autumn, and PO_4 and NO_3 having highest concentrations mainly in winter and spring, the succession of algal functional groups was not exactly as hypothesised. It was anticipated that blue-green algae would occur in highest abundance during summer and into autumn, diatoms would be in greatest numbers during winter and into spring and that green algae would mainly fill spring and into summer. Some overlapping was expected but not of peak populations. However, the successional order observed showed that whilst the blue-green algae did occur in highest abundance when expected in summer, the green algae peaked in late summer and into autumn and the diatoms were interestingly shown in highest abundance during autumn and into winter. It would seem that the management in place slightly disturbs the natural patterns and algal succession in these water bodies. It is known that diatoms require turbulence to move their heavy silica structure into the euphotic zone and prevent them from sedimenting out of the water column. Therefore in an unmanaged water body, early autumn would probably not support large populations of diatoms, as the water column is still relatively stable for the first part of the month until turnover. However, at Myponga and Happy Valley reservoirs, the artificial mixing and destratification means that diatom populations can start to build up in summer and peak in autumn due to the mechanical mixing that is providing the necessary turbulence for their survival. Management practices throughout summer and autumn are also what suppress blue-green algae enough to enable green algae to reach peak abundance at the same time. Therefore, the impact of management can be seen in terms of short-term dynamics as well as longer-term changes.

The examination of the long-term changes of water quality at Myponga reservoir over a period of 16 years, in context of the several different mixing regimes used throughout that time, gave interesting results. It was hypothesised that the continued management of the reservoir would lead to improved water quality conditions over time. However, this was not the case for many variables and therefore the management strategy employed at Myponga reservoir can be

improved. Particular attention needs to be paid to PO_4, Chl-*a* and Mn concentrations, which had higher levels in the last period than they had experienced in any of the previous periods.
The investigation of algae functional group occurrence in relation to changing levels of a particular variable showed that green algae and diatoms coincided with periods of low PO_4 and NO_3 concentrations. Although the nutrients are probably at reasonable levels to begin with, there is rapid uptake and depletion by these functional groups that can reach very high abundances. Blue-green algae occurred during times when the nutrients were at mid-to-high concentrations (early to mid summer, after the winter-spring accumulation but before the depletion by large phytoplankton growth) and because they are managed to quite low population numbers, they did not deplete the source pool as the other functional groups did. The fact that *Anabaena* is known to fix atmospheric nitrogen could further explain the lack of impact on NO_3 levels. These results fit in with the seasonality of the functional groups demonstrated in the initial experiment. When blue-green algae begin to build up populations in early summer, there is a high level of nutrients due to accumulation over winter and spring from increased input and reduced uptake. By mid to end of summer the nutrients deplete as blue-green and green algae populations climb, resulting in low nutrient levels by autumn when green and diatoms are seen to occur in high numbers. This last series of experiments demonstrates a method that could potentially reveal how to manipulate water quality variables to improve water quality by discouraging particular phytoplankton occurrence and encouraging others that are less harmful.

This chapter has demonstrated that KANN is a useful instrument for the exploration of complex ecological time-series data. The method has been shown to allow examination of short-term dynamics such as seasonality and succession, and permit comparisons between two study sites. Short- and long-term patterns can be viewed in context of management regimes by splitting the data into periods based on different management strategies. This allows the assessment of management efficiency and highlights particular aspects where management was successful or should be focused and improved. Examination of the conditions that promote particular groups of algae can demonstrate areas that could be manipulated to manage the algal communities for better water quality. This chapter has clearly shown KANN to be capable of being applied to ecological data sets in many different manners. The method seemed to provide ordination and clustering in a superior manner than traditional techniques such as PCA, for greater unravelling of complexities within the data. Visualisation of the clusters in a clear and instructive manner was also a great advantage compared to the traditional techniques.

In summary, KANN can be used to thoroughly explore and examine water quality data and providing pattern analysis through time and across periods of different management; highlighting important relationships, possible interactions and areas of interest for further research. It was an important and useful first step for this project as the results and information gained from this series of KANN experiments provided better understanding of the functioning of the study sites, which could then be applied to the development of forecasting models.

5. FORECASTING OF CHLOROPHYLL-A AND *ANABAENA* DYNAMICS

5.1 Introduction

The need for the development of predictive water quality models has increased due to water quality deterioration of many of the world's freshwater sources, along with an increased demand for potable water. Prediction and elucidation of changes in water quality are particularly important for water resources relied upon by human populations. Water industries and government agencies in Australia undertake enormous research and management efforts in order to prevent algal blooms in lakes and drinking water reservoirs, as manifested by the establishment of the CRC for Water Quality and Treatment (www.waterquality.crc.org.au). As recommended by the UK National Rivers Authority (NRA 1990), appropriate predictive tools and early warning systems need to be developed and implemented in order to prevent future expenses in the order of millions of dollars to water industries by harmful algal outbreaks in drinking water reservoirs, lakes and rivers. Early warning of impending algal blooms should facilitate the prevention of major bloom events by allowing the timely implementation of operational control measures such as intermittent artificial mixing.

Using time-series water quality datasets from Myponga and Happy Valley reservoirs has allowed the development of forecasting models for each site, and the performance of sensitivity analyses, which give insight into relationships that drive phytoplankton dynamics. The two datasets also allowed comparisons between the dynamics and relationships within the reservoirs, which belong to the same category of warm monomictic and eutrophic lakes.

This chapter discusses the application and comparison of Recurrent Artificial Neural Networks (RANN) and Hybrid Evolutionary Algorithms (HEA) methods for predictive modelling of phytoplankton dynamics. Both methods enable useful time-series modelling by inducing forecasting models from historical data patterns. Models discovered by the two methods allow the conduction of sensitivity analyses in order to reveal quantitative relationships between the input variables and the output, which enables the elucidation of causal relationships between water quality parameters.

The forecasting of phytoplankton dynamics in both Myponga and Happy Valley reservoirs was focused on either Chl-*a* concentration, as an indicator of total algal biomass, or the specific abundance of the dominant nuisance species, *Anabaena circinalis*.
Models developed by either of the two methods can potentially be applied to real-time monitoring data and thus provide real-time forecasting.

5.2 Aims and Hypotheses

The main aim of this chapter is, quite simply, to accurately predict the timing and magnitude of Chl-*a* concentration or *Anabaena* abundances within the water bodies. Through the development of single lake models for each study site, the study also aims to make comparisons between the reservoirs. Additionally, the construction of models trained with merged data endeavours to take a step towards generic modelling and forecasting. The use of the two forecasting techniques allows the comparison between methods with the intention of determining the best method for forecasting phytoplankton dynamics in real time. In the pursuit of these aims, the validity of the following hypotheses will be investigated:

Hypothesis 1 – models developed by RANN and HEA forecast the timing and magnitude of algal abundance reasonably well using physical and chemical input variables.

Hypothesis 2 – algal population dynamics are more rapid, distinct and challenging to model than algal community dynamics represented by Chl-*a*, and therefore forecasting models for *Anabaena* will be less accurate than those forecasting Chl-*a*.

Hypothesis 3- training models with merged data from same category reservoirs would provide the model with more patterns of similar nature for learning and thus, improve the generalisation of the models.

5.3 Methods and Materials

5.3.1 Data

Interpolated daily values of water quality time series data from both reservoirs were used for the training and testing of RANN and HEA models, allowing daily forecasts of phytoplankton

dynamics. Data selection was dependent upon what parameters were available at the time of the experiments, the duration of their availability and the outcome of sensitivity analyses. Tab. 8 shows the years of data used in each experiment.

Table 8. Data used for each experiment in this chapter

Experiment	Years of data used
Forecasting aim: 7-days ahead Chl-a forecasting in Myponga reservoir Section: 5.4.1.1 and 5.4.1.2 Inputs: 1 test year experiments: PO_4, NO_3, water temperature, turbidity, colour and DO 2 test years experiments: PO_4, NO_3, water temperature, turbidity and DO	1993-2003
Forecasting aim: 7-days ahead Chl-a forecasting in Happy Valley reservoir Section: 5.4.1.3 and 5.4.1.4 Inputs: 1 test year experiments: PO_4, NO_3, water temperature, turbidity, colour and DO 2 test years experiments: PO_4, NO_3, water temperature, turbidity and DO	1991, 1994-2003
Forecasting aim: 7-days ahead forecasting of Anabaena in Myponga reservoir Section: 5.4.2.1 and 5.4.2.2 Inputs: 1 test year experiments: PO_4, NO_3, water temperature, turbidity, colour, DO and Chl-*a* 2 test years experiments: PO_4, NO_3, water temperature, turbidity, DO and Chl-*a*	1995-2003
Forecasting aim: 7-days ahead forecasting of Anabaena in Happy Valley reservoir Section: 5.4.2.3 and 5.4.2.4 Inputs: 1 test year experiments: PO_4, NO_3, water temperature, turbidity, colour, DO and Chl-*a* 2 test years experiments: PO_4, NO_3, water temperature, turbidity, DO and Chl-*a*	1991, 1994-2003
*For merged applications the years listed for each reservoir were combined.	

5.3.2 Model Design

Using each method, single lake models were developed for each reservoir separately. Later, merged lake models were created using merged data from both reservoirs, to see whether the accuracy of the forecasts would improve through increased training samples or patterns, or decrease in accuracy due to the more generic nature of the model (see Fig.30 for experimental progression).

Model evaluation, for both methods, was based on r^2 and RMSE values along with visual assessment of the nearness of the measured and predicted data.

Time-lagged inputs were used for all RANN and HEA experiments, with a 7-day forecast horizon used throughout this chapter. Time-lagged inputs are considered useful for several reasons, including that there is generally a slight lag between the impacts any current conditions will have on an algal population and also because it allows the prediction of the output with an advance of a pre-determined time frame. Preliminary experiments examining models using time lags of 3, 7, 14 and 21 days were carried out. 7-day lags were seen as the ideal time frame for the purposes of this work as it is a compromise between the 3-7day lags which are of interest to water managers, and longer lags which provide more opportunity to challenge and demonstrate the model's capabilities. 3 and 7-day lags gave the best results and were of a similar accuracy level and 7-day lags were chosen as they allowed sufficient demonstration of the models forecasting abilities and would provide ample time for water management to decide upon, organise and implement the necessary operational control and management strategies.

In both methods the split-sample validation method (see Weiss and Kulikowski, 1991) was used with one or two testing years. Initially one testing year was selected, however it was later decided that this could be misleading regarding the models predictive capabilities and therefore, two testing years were used in later experiments. Selection of the testing years was based on providing the model with the greatest challenge and opportunity to prove its use in different conditions by the selection of two years that were quite different in timing and magnitude of the output (Chl-*a* or *Anabaena*).

As explained in 3.3.1.4, sensitivity analyses are used in association with each method, to provide insight into the relationships between the input and output data. The 'sensitivity on wide-ranged disturbance' (SWD) approach is used in this research (described in 3.3.2.5). Using this method, all input variables, except the one of interest, are kept constant at their mean value. The variable being investigated is varied within the range of its mean by +/- 1 standard deviation, which is thought to encompass commonly occurring variation, and therefore the general conditions of the water bodies. Most influencing parameter (MIP) analysis (described in 3.3.2.5) was also carried out on the RANN models, and results are mentioned briefly in text with graphs displayed in Appendix E.

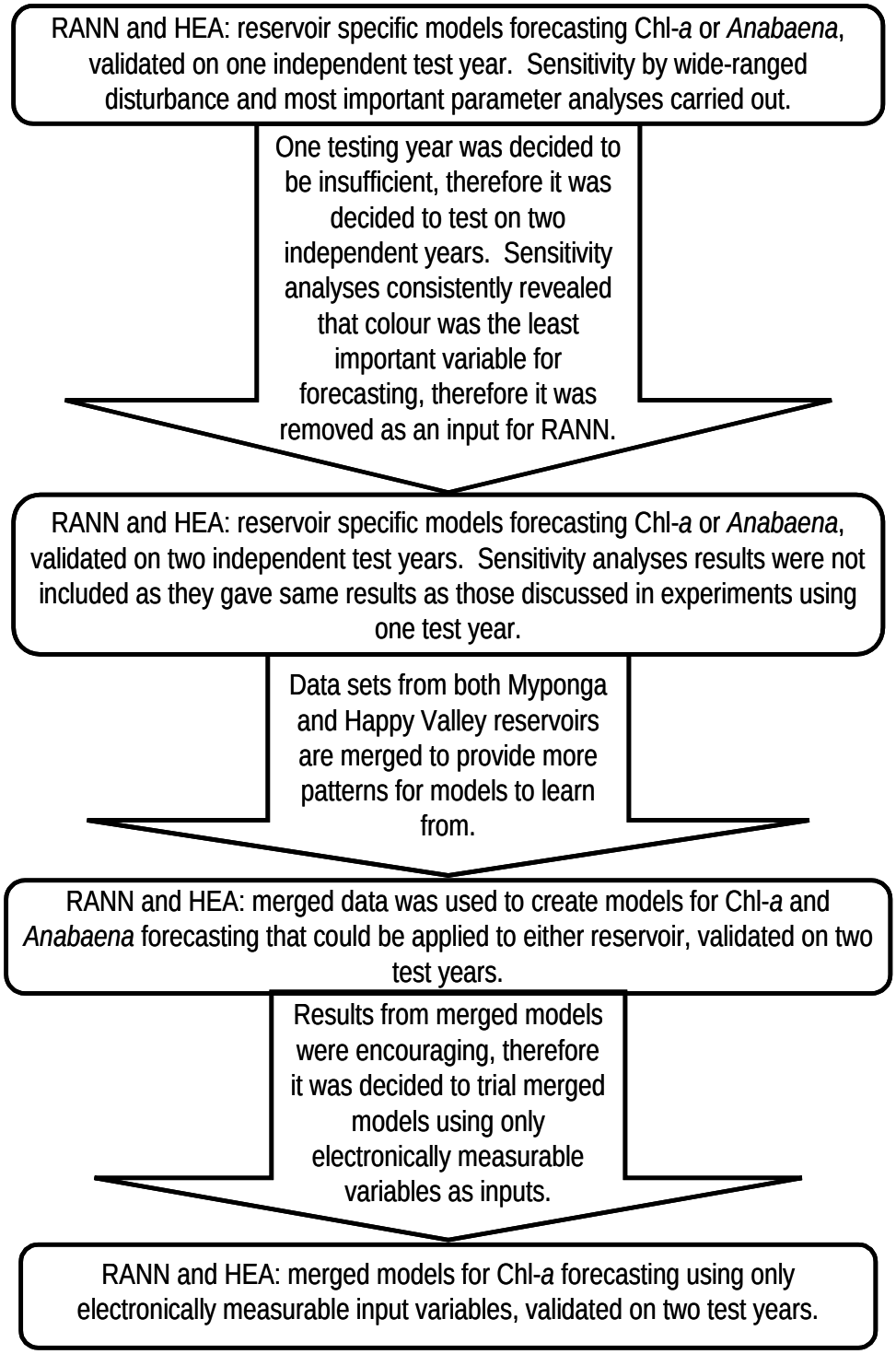


Figure 30. Experimental progression of Chapter 5

5.3.2.1 RANN

RANN (Pineda 1987) was designed to predict future patterns or values by learning from patterns and examples presented to the network. Being a 'supervised' ANN, its learning is guided by known outputs, and its recurrent nature means that the activations from the time step before are fed back into the network as another form of input to improve accuracy of time-series modelling (see Fig.2).

For the development and implementation of the RANN models, NeuroSolutions for Excel Version 4.2 was used.

Architecture

RANN requires the selection of appropriate network architecture (see 3.5.2.1 for descriptions of architecture components listed here). Some of the network architecture parameters of the RANN models remained constant for all experiments; others were optimised to suit the particular situation. All RANN were designed with one hidden layer and used the Tanh Axon (hyperbolic) transfer function, with the Momentum learning rule set to 0.7 for both the hidden and output layer, and a step size of 0.01 and 0.1 for the hidden and output layers respectively. For most experiments a simple Axon input layer structure was used but for the models forecasting *Anabaena,* Laguarre Axon - a memory axon, was found to be necessary to gain acceptable results. The input layer structure, number of processing elements or nodes and training time varied between the models and will be indicated appropriately.

Input selection

Initial input selection was based on knowledge of important factors relating to algal growth and population maintenance. The sensitivity analyses were then examined to see if there were any redundant inputs that could be removed to potentially improve the results. This trial and error technique continued and inputs were adjusted accordingly.

Training time

Optimal training time can differ with each model and the trial and error method was largely used to find the ideal amount of training for these models. Criteria upon which ideal training time was determined include reasonably accurate results when the model is tested on the training data (extremely good training results often demonstrate that the network has been overtrained which results in poor forecasting abilities, whilst poor training results likely demonstrate that the network

is yet to be trained sufficiently), and testing results within an acceptable error range when the network is applied to unseen data.

5.3.2.2 HEA

HEA (Cao et al. 2006b) was designed to extract predictive rule sets from within the training data set in the form of an IF-THEN-ELSE statement. HEA uses genetic programming to evolve the structure of the rule set and then uses a general genetic algorithm to optimise the random parameters in the rule set (see Fig.10).
All HEA experiments were performed on a supercomputer (IBM eServer 1350 Linux) using the C programming language.

Design

HEA model design requires the initial determination of several factors such as initial population number, number of generations, rule complexity etc. (see 3.5.3.1 for explanation of all parameter settings). Throughout this study all HEA models began with an initial population of 200, with a maximum of 100 generations. This was thought to allow sufficient searching for the best rule-set. The model output was set as a single rule to avoid over-complicating application of the rule set.

Input selection

For this method the algorithm carried out input selection by determining key driving variables and disregarding unimportant inputs; therefore all available input variables were supplied, though not all were necessarily used (see Methods 3.3.3.2 for more detail).

5.3.2.3 Chl-a as an input

Using Chl-*a* concentration as an input into models predicting Chl-*a* is a contentious issue. Traditionally, Chl-*a* concentrations have been left out of the input variables as there is the obvious effect of auto-correlation. From a management perspective, the importance is on the accuracy and reliability of the predictions, and the fact that the inclusion of Chl-*a* as an input to models predicting Chl-*a* improves the accuracy of the forecasts cannot be argued. However, from a modelling perspective, the use of Chl-*a* as an input would not show the power of the methods to draw significant causal variables from the input data and extract relationships. The tendency would be to largely rely on the Chl-*a* input data for the predictions. It is important to make the distinction between the intention of predicting Chl-*a* as a function of other inputs, in which case

Chl-*a* can obviously not be used as an input, or forecasting Chl-*a* in advance, where Chl-*a* can be an input if there is no interest in causal relationships and model interpretation. Recent applications using Chl-*a* as an input include Coad *et al.* (2005), who used only online Chl-*a* data from the previous week to determine the Chl-*a* concentration 1 – 7 days in advance in the Berowra estuary, using SNN. This study found that the 'Chl-*a* input only model' produced the best predictive results ($r^2 = 0.89$) but concluded that the causative parameters for inducing algal blooms could not be deduced from the predictive models developed in that manner. Lee *et al*. (2003) also used solely Chl-*a* concentrations with time lags of 7 –13 days for input to ANN for prediction of coastal algal blooms. This study concluded that all the ecological information required to make accurate forecasts appeared to be embodied by the Chl-*a* from the week or so prior. This may be the case, but it does not allow any insight into the ecological information. Experiments conducted by this author showed that when Chl-*a* concentration from one week prior was offered to the HEA as input, it was heavily relied upon to predict the Chl-*a* concentration. Although the result was good with regard to magnitude and timing, the model mostly suggested the Chl-*a* concentrations from the week before as forecasts of the concentration at any particular time which gave a 7-day delay effect (see Appendix D for an example).

For the purposes of this study, where the relationships driving algal dynamics are also of interest, as well as the forecasting results, it was considered best to not include Chl-*a* concentration as an input for models predicting Chl-*a* concentration. It was thought that this would enable better testing of the abilities of the modelling methods and allow the examination of the driving variables for Chl-*a* concentration. However, for models developed to forecast *Anabaena* abundances, Chl-*a* was made available as an input.

5.4 Results

5.4.1 7-days ahead forecasting of Chl-*a*

5.4.1.1 Myponga reservoir validated for 2001

The results of Chl-*a* forecasting in Myponga reservoir using reservoir specific models developed by RANN and HEA can be seen in Fig. 31a. Information on the models developed for this application is available in Tab. 9. The validation results produced from the application of these models to Myponga reservoir data from 2001 show that both models do well in forecasting the major peak event, with the timing of the maximum values being excellent, although the magnitude was underestimated in each case, but slightly more so by the HEA derived model. The RANN model best forecasts the timing and magnitude of the first event, where Chl-a concentrations reached a peak of 15µg/L, while the HEA model gives a delayed forecast of that concentration. The lower Chl-*a* concentrations experienced outside of the peaks are best forecasted by the HEA model, which performs well regarding magnitude, where the RANN model forecasts a close to 0 concentration. Although this would not be a problem when forecasting for management, as these values are not a concerning level; from the assessment of the model purely from a modelling perspective, it is not ideal. The small Chl-*a* peak of approximately 10µg/L at the end of the test year is predicted by both models, though both are slightly delayed in timing and, whilst the timing of the maximum values are correct, the concentration was underestimated in both cases. The HEA derived rule-set gave the best forecasting results with an r^2 value of 0.89 and RMSE of 2.88, compared with the RANN model which gave of r^2 value of 0.77 and an RMSE of 3.85 Overall both models were successful in identifying the major peak Chl-*a* concentrations.

Table 9. Information table for RANN and HEA models developed to forecast Chl-*a* concentration in Myponga reservoir (1test year)

RANN		HEA	
Input layer structure	Axon	IF ((NO_3*((NO_3*144.294)*(191.268-WT)))>=98.396)	
Processing Elements	25	THEN Chl-*a* = ((WT/470.057)*((Colour/(NO_3*64.755))*WT))	
Transfer function	TanhAxon	ELSE Chl-*a* =((Colour/(189.292-Colour))*WT)	
Iterations	8000	Runs	50

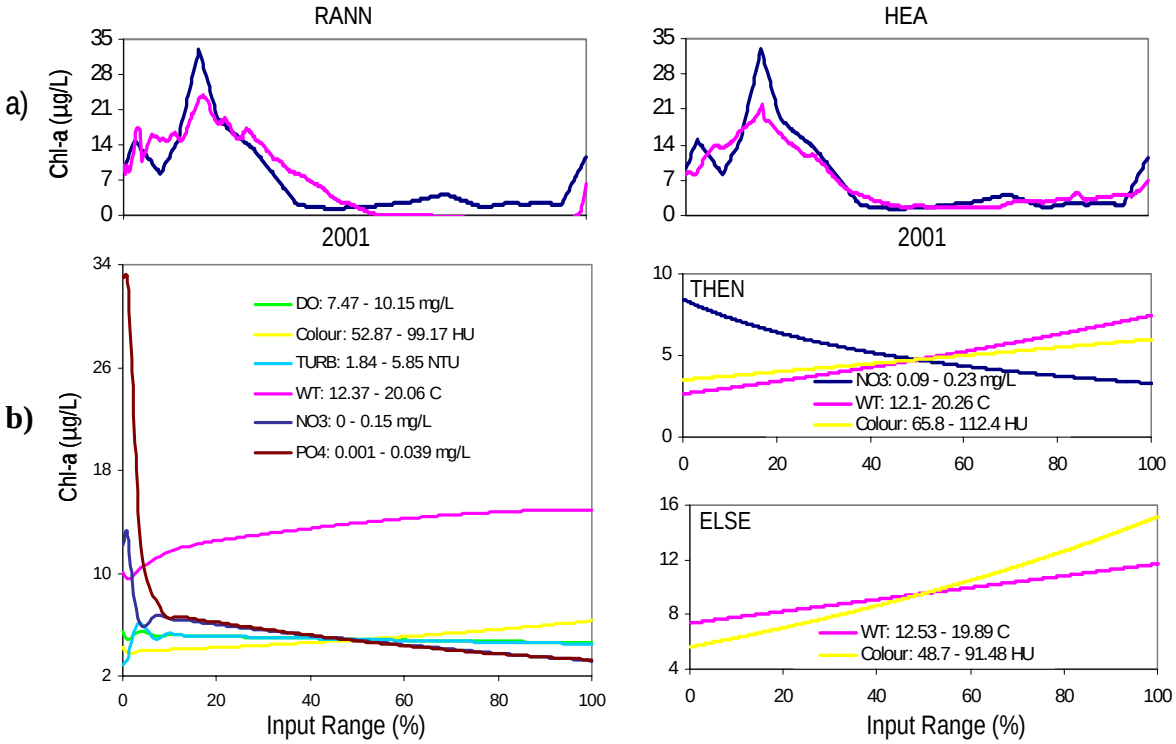


Figure 31. a) Chl-*a* forecasting results for RANN (left) and HEA (right) models tested on one year of data from Myponga reservoir, b) sensitivity analyses results from RANN (left) and HEA (right).

The sensitivity analyses produced from the methods can be seen in Fig. 31b. Of the six input variables, MIP analysis by RANN identified PO_4 concentration as the most influential, distantly followed by NO_3 concentration (see Appendix E). From the six possible input variables, HEA detected that water temperature, colour and NO_3 concentrations as driving variables for the accurate prediction of Chl-*a* concentrations in Myponga reservoir. The sensitivity graph obtained by RANN shows that highest Chl-*a* concentrations coincide with very low PO_4 concentrations and that increasing PO_4 levels occur in conjunction with decreasing Chl-*a* concentrations. This occurs because PO_4 is the limiting nutrient for algal growth and maintenance and as the algal community (represented by Chl-*a*) increases in number, the uptake of the nutrient increases which results in depletion of the PO_4 concentrations within the water column (Boney 1989). In fact, the concentration of PO_4 in a lake can be readily related to Chl-*a* concentrations (Moss 1998). Similarly, NO_3 is another important nutrient for most algae and is shown by the sensitivity results from both RANN and HEA to follow a similar pattern to PO_4. Again it is due to increasing algal numbers placing further demand on the nutrient reserves and the increased uptake resulting in depletion of the nutrient to very low concentrations during times of large algal populations. Increasing water temperature was shown by both RANN and HEA to stimulate increasing Chl-*a*

concentrations. Water temperature greatly influences the rate of metabolic processes in algal cells and warmer water temperature results in much higher levels of photosynthesis (Reynolds 1989). Both methods demonstrated a positive relationship between colour and Chl-*a* concentration. Colour in water can be due to the presence of metallic ions such as Fe and Mn, lignin compounds, dissolved organic carbon (DOC) and phytoplankton among other things, and Myponga reservoir is characteristically very highly coloured predominantly by DOC (see Appendix C). Increasing turbidity in Myponga reservoir was shown by RANN to correspond with a slight decline in Chl-*a* concentration. Whilst phytoplankton as well as suspended and colloidal matter can influence turbidity, in Myponga it is largely non-algal (Burch 2005a). In light of this, the results can be interpreted as showing that increasingly turbid water discourages algal growth, probably by light limitation. The RANN sensitivity graph shows that Chl-*a* concentration does not significantly respond to increasing DO concentrations. The relationship between DO and phytoplankton is a complex one. Oxygen is a product of algae photosynthesis, and in this way increased phytoplankton could result in increased dissolved oxygen in the water column. However, respiration and the decomposition of dead algal cells increase the biological oxygen demand and reduced dissolved oxygen levels. In this particular sensitivity result, no significant relationship can be determined.

5.4.1.2 Myponga reservoir validated for 1995 and 2001

The results of Chl-*a* forecasting in Myponga reservoir using reservoir specific models developed by RANN and HEA can be seen in Fig. 32. Information on the models developed for this application is available in Tab. 10. The validation results produced from the application of these models to Myponga reservoir data from 1995 and 2001 show that both models do well in forecasting the major peak event at the beginning of 2001, although the magnitude was underestimated in each case, but slightly more so by the HEA derived model. The timing and magnitude of the first event, where Chl-*a* concentrations reached a peak of 15µg/L, is best forecast by the RANN model. The RANN model predicts high concentrations at the beginning of the 1995 year, which is often the case though the pattern was disturbed by CuSO4 dosing in this case. The lower Chl-*a* concentrations experienced outside of the peaks are best forecast by the RANN model, which does well regarding timing and magnitude, where the HEA model forecasts underestimates providing concentrations of close to 0. Although this would not be a problem for forecasting for management, as these values are not concerning, from the assessment of the model purely from a modelling perspective, it is not ideal. The small Chl-*a* peak of approximately 10µg/L at the end of the 2001 test year is predicted well by RANN, whilst HEA indicates some

growth at that time but underestimates the concentration. The RANN model gave the best forecasting results both visually and by RMSE of 3.68, with an r^2 value of 0.62 compared with the HEA derived model which gave of r^2 value of 0.66 and an RMSE of 4.17. Overall both models were successful in identifying the major peak Chl-*a* concentrations.

Table 10. Information table for RANN and HEA models developed to forecast Chl-*a* concentration in Myponga reservoir (2test years)

RANN		HEA	
Input layer structure	Axon	IF (((PO4*30.6)*147.2)>260.186)	
Processing Elements	23	THEN Chl-*a* = ((WT-1.3)*((PO4+(WT/57.56))- NO3))	
Transfer function	TanhAxon	ELSE Chl-*a* =(Colour*((PO4+(WT/88.7))- NO3))	
Iterations	4000	Runs	50

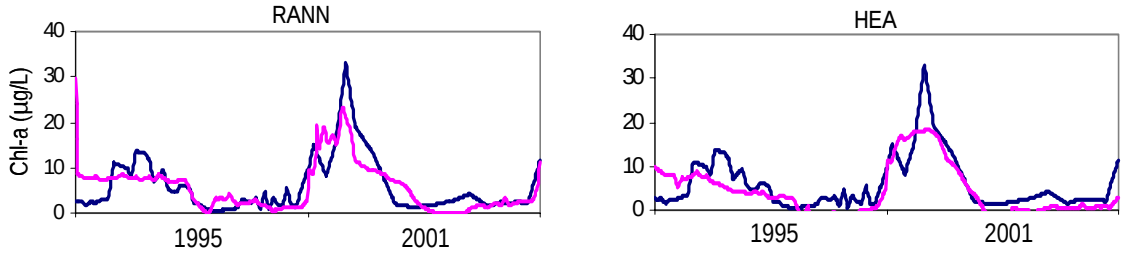


Figure 32. Chl-*a* forecasting results for RANN (left) and HEA (right) models tested on two years of data from Myponga reservoir

5.4.1.3 Happy Valley reservoir validated for 1999

The results of Chl-*a* forecasting in Happy Valley reservoir using reservoir specific models developed by RANN and HEA can be seen in Fig. 33a. Information on the models developed for this application is available in Tab. 11. The validation results produced from the application of these models to Happy Valley reservoir data from 1999 show that both models do well in forecasting the two main algal growth events. The first of these events has two peaks, the first of which is approximately 18µg/L and is well forecast by the HEA derived model, but not predicted by the RANN model. For the second peak, the opposite is true, as the RANN model predicts the correct magnitude in advance, whilst the HEA rule-set does not forecast this event well. The other key growth event of up to 23µg/L at the end of the test year was clearly identified by both models, though the RANN model gives a slightly delayed overestimation, whereas the HEA rule-set suggests the event is of a smaller magnitude. The lower Chl-*a* concentrations between the peak events was forecast particularly well by the RANN model, except for a small false peak.

The RANN model gave the best forecasting results with an r^2 value of 0.81 and RMSE of 3.23, compared with the HEA derived rule-set which gave of r^2 value of 0.76 and an RMSE of 3.56 Overall both models were successful in identifying the major peaks in Chl-*a* concentrations.

Table 11. Information table for RANN and HEA models developed to forecast Chl-*a* concentration in Happy Valley reservoir (1test year)

RANN		HEA	
Input layer structure	Axon	IF (WT<21.01)	
Processing Elements	17	THEN Chl-*a* = (((Colour/PO_4)+942.33)/412.701)	
Transfer function	TanhAxon	ELSE Chl-*a* =((((Colour/PO_4)+515.197)+(TURB*68.590))/254.190)	
Iterations	9000	Runs	50

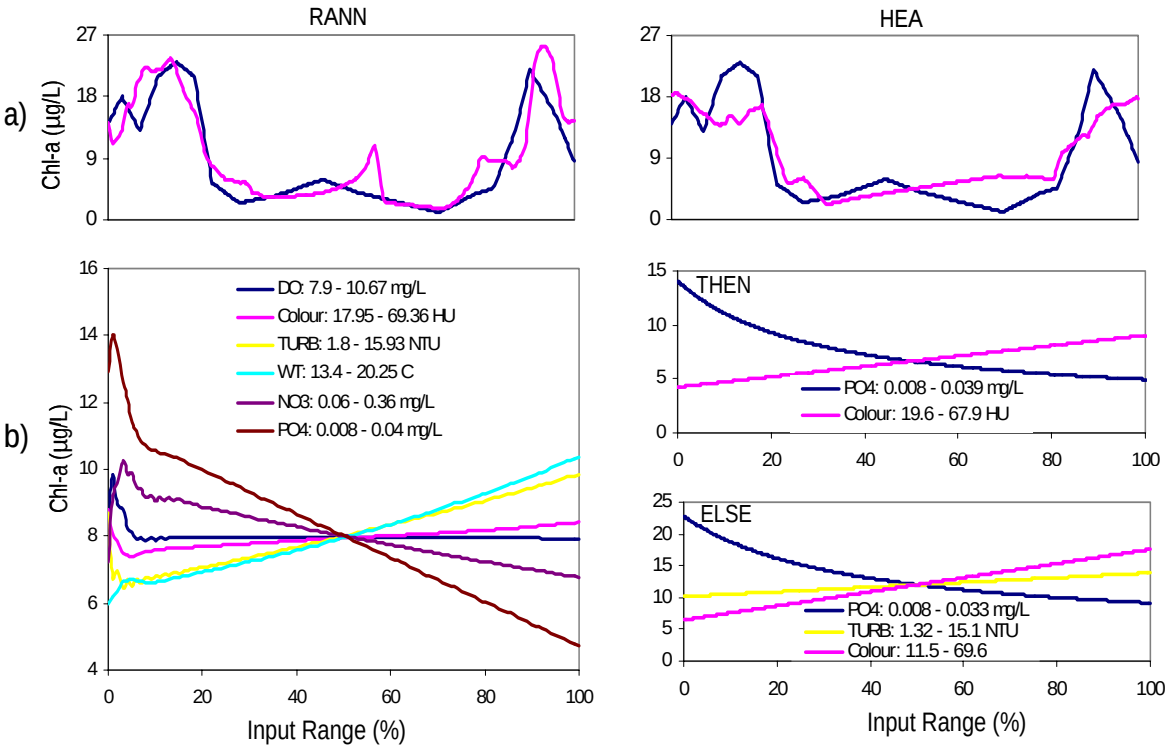


Figure 33. a) Chl-*a* forecasting results for RANN (left) and HEA (right) models tested on one year of data from Happy Valley reservoir, b) sensitivity analyses results from RANN (left) and HEA (right).

The sensitivity analyses produced from the methods can be seen in Fig.33b. Of the six input variables, MIP analysis by RANN identified PO_4 concentration as the most influential, distantly followed by NO_3 concentration (see Appendix E). From the six possible input variables, HEA detected that turbidity, colour and PO_4 concentrations were necessary for the accurate prediction of Chl-*a* concentrations in Happy Valley reservoir. The sensitivity graphs obtained by RANN and HEA show that the highest Chl-*a* concentrations coincide with very low PO_4 concentrations and

that increasing PO_4 levels occur in conjunction with decreasing Chl-*a* concentrations. Results from RANN show that NO_3 also behaves in this manner. As in Myponga reservoir, the patterns shown by these nutrients are due to increasing algal numbers placing further demand on the nutrient reserves and the increased uptake resulting in depletion of the nutrient to very low concentrations during times of large algal populations. In keeping with long established ecological theory, rising water temperature was linked by RANN to increasing Chl-*a* concentrations as it stimulates algae growth. As Chl-*a* contributes to the colour of a water body, it is logical that as Chl-*a* levels increase so does the colour intensity, and this was demonstrated by both RANN and HEA sensitivity analyses. Turbidity in Happy Valley reservoir was shown by RANN and HEA to coincide with increasing Chl-*a* concentration. Happy Valley reservoir is more productive than Myponga and algae contribute to turbidity more so in this reservoir, and therefore as algal abundance increases, so too does turbidity. As with Myponga reservoir, the RANN sensitivity graph shows that Chl-*a* concentration in Happy Valley reservoir does not significantly respond to increasing DO concentrations and, again, no significant relationship can be determined. As with the sensitivity analysis results for Myponga reservoir, here *Anabaena* was found to contribute to Chl-*a* levels.

5.4.1.4 Happy Valley reservoir validated for 1997 and 1999

The results of Chl-*a* forecasting in Happy Valley reservoir using reservoir specific models developed by RANN and HEA can be seen in Fig. 34. Information on the models developed for this application is available in Tab. 12. The validation results produced from the application of these models to Happy Valley reservoir data from 1997 and 1999 show that both models forecast the major Chl-*a* trends. The first of these events had many peaks, and while neither model forecast this variation, they were both successful in identifying the timing of the highest peak, although both significantly underestimate it. Of the 3 remaining events of approximately 20µg/L, the first is not well identified by either method, the second is indicated by both methods, but most successfully by the RANN model and the third was also indicated by both, most effectively by HEA which forecasts the correct magnitude slightly delayed. The HEA derived rule-set which gave the best forecasting results with an r^2 value of 0.55 and an RMSE of 6.44 compared with the RANN model which gave an r^2 value of 0.48 and RMSE of 7.11. Overall both models were successful in identifying the major trends in Chl-*a* concentrations.

Table 12. Information table for RANN and HEA models developed to forecast Chl-*a* concentration in Happy Valley reservoir (2 test years)

RANN		HEA	
Input layer structure	Axon	IF ((Colour+227.534)<=259.087)	
Processing Elements	23	THEN Chl-*a* = (Colour/((PO_4*73.5)+exp(NO_3)))	
Transfer function	TanhAxon	ELSE Chl-*a* =(Colour/((PO_4*227.02)+exp(NO_3)))	
Iterations	1000	Runs	50

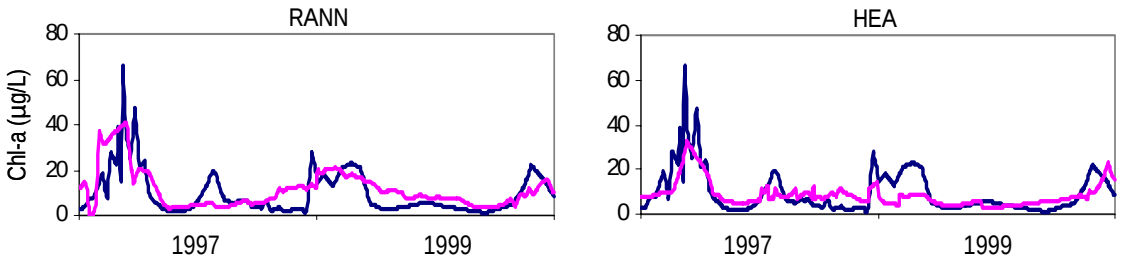


Figure 34. Chl-*a* forecasting results for RANN (left) and HEA (right) models tested on two years of data from Myponga reservoir

5.4.1.5 Myponga and Happy Valley reservoirs validated for 2001 and 1999 respectively

Models were developed using RANN and HEA and trained by merged data from Myponga and Happy Valley reservoirs, then tested on one year of data from each reservoir.

The results of Chl-*a* forecasting using the merged models developed by RANN and HEA can be seen in Fig. 35. Information on the models developed for this application is available in Tab. 13. The validation results produced from the application of these models to reservoir data from Myponga reservoir (2001) and Happy Valley reservoir (1999) show that both models do well in forecasting the 3 major peak events at the beginning, middle and end of the testing period. Both RANN and HEA models indicate the timing of the events quite well, thought the prediction of the third event is slightly delayed in both cases. The magnitude is best forecast by the RANN model, which forecast concentrations up to 25 µg/L, where as the HEA model underpredicted the major events by only forecasting concentrations up to approximately 15 µg/L. Overall the RANN model provided the best results with an r^2 value of 0.66 and an RMSE of 4.37, compared with the results from the HEA rule-set which gave an r^2 value of 0.52 and an RMSE of 5.07.

Table 13. Information table for RANN and HEA models developed using merged data to forecast Chl-*a* concentration in both reservoirs (2 test years)

RANN		HEA	
Input layer structure	Axon	IF (WT>18.076)	
Processing Elements	27	THEN Chl-*a* = = ln(|(Colour*(Colour*(TURB*Colour)))|)	
Transfer function	TanhAxon	ELSE Chl-*a* = ln(|((((-1.068)/PO_4)+14.390)*((-1.068)/PO_4))|)	
Iterations	15000	Runs	50

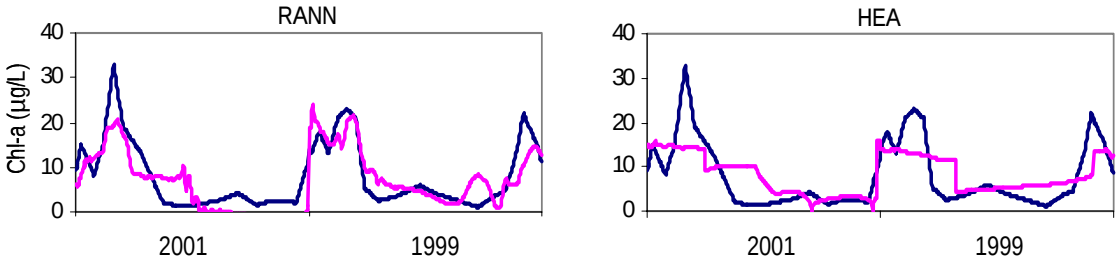


Figure 35. Chl-*a* forecasting results for merged RANN (left) and HEA (right) models tested on two years of data, one from each reservoir

5.4.1.6 Myponga and Happy Valley reservoirs validated for 2001 and 1999 respectively, based on electronically measurable data only

The promising results from the merged models designed to forecast Chl-*a* concentration gave encouragement for the concept of developing a merged model to forecast Chl-*a* concentration using only electronically measurable inputs. The input variables made available to the models were: DO, water temperature, turbidity and conductivity. Information on the models developed for this application is available in Tab. 14.

The results of Chl-*a* forecasting in both reservoirs using merged models developed by RANN, using only electronically measurable inputs can be seen in Fig. 36. The RANN model was clearly able to identify the three major Chl-*a* growth events throughout the two year testing period. The timing of the onset of the first two bloom events were well forecast, with the third event being slightly delayed. The first and last Chl-*a* spikes were underestimated with regard to magnitude, whilst the middle peak was somewhat overestimated. For the purposes of management and decision-making, overestimates of concentrations (within reason) are preferred to underestimations, because they at least mean that management will be prepared for the level of growth that does eventuate. The validation results gave an r^2 value of 0.22 and an RMSE of 8.04.

The results of Chl-*a* forecasting in both reservoirs using merged models developed by HEA, using only electronically measurable inputs can be seen in Fig. 36. The HEA model was also able to clearly identify the three major peaks in Chl-*a* concentration. The timing and magnitude of the middle peak event was predicted very well, though the model falsely predicts another peak immediately after. The first and last Chl-*a* concentration spikes were underestimated as they were by the RANN model (r^2 value of 0.22, RMSE of 8.04), but were closer to the actual concentrations. Visually the HEA model appears to perform the best out of the two methods and this is confirmed by the error values, with an r^2 value of 0.5 and an RMSE of 5.3.

Table 14. Information table for RANN and HEA models developed using merged data, with only electronically measurable input variables, to forecast Chl-*a* concentration in both reservoirs (2 test years)

RANN		HEA	
Input layer structure	Axon	IF (WT<18.623)	
Processing Elements	19	THEN Chl-*a* = exp((WT/((47.394/DO)+3.055)))	
Transfer function	TanhAxon	ELSE Chl-*a* = exp((WT/((35.538/DO)+(COND/172.917))))	
Iterations	4000	Runs	50

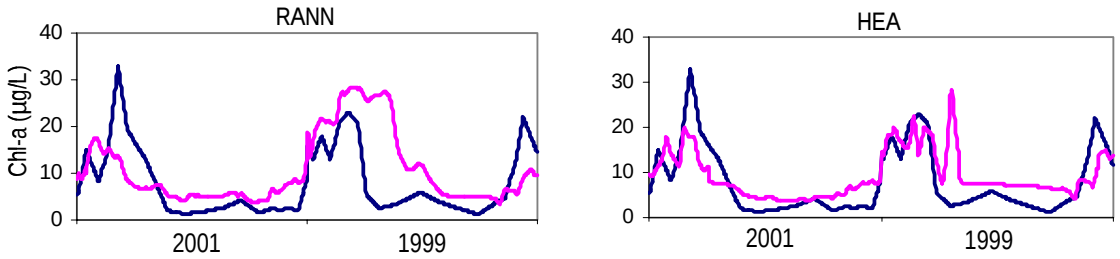


Figure 36. Chl-*a* forecasting results for merged RANN (left) and HEA (right) models developed using only electronically measurable inputs, tested on two years of data, one from each reservoir

Considering the limited input variables of DO, water temperature, turbidity and conductivity, both models did well to successfully identify the major trends in the Chl-*a* concentration over the two year testing period. Compared with the merged models created using additional nutrient data (see section 5.4.1.5), the RANN model had lower accuracy, however the HEA model produced results of the same standard and quality.

5.4.2 7-days ahead forecasting of *Anabaena* abundance

5.4.2.1 Myponga reservoir validated for 2002

The results of *Anabaena* forecasting in Myponga reservoir using reservoir specific models developed by RANN and HEA can be seen in Fig. 37a. Information on the models developed for this application is available in Tab. 15. The validation results produced from the application of these models to Myponga reservoir data from 2002 show that both models do well in forecasting the major peak event, with the timing being correct, although the magnitude was underestimated in each case, but more so by the HEA derived model. The low *Anabaena* abundance outside of the peak event was best forecast by the HEA rule set. Although it forecasts the cells/mL at approximately 50-100cells/mL above the actual abundance, these levels are not concerning for management and the models' output is therefore acceptable. The RANN model, however, forecasts two false growth events of about 200cells/mL above the actual measured abundance. Again, the levels are quite low and would not be of concern to management anyway, but false peaks of larger magnitudes could waste management efforts and are therefore not ideal. Although the RANN model provided some small false positives, visually it is the best model of the two as its forecasts are closer to the actual magnitude. This is reflected by a lower RMSE than the HEA rule set but interestingly not by the r^2 values. The results of validation of the RANN model produced an r^2 value of 0.36 and an RMSE of 174.95, whereas the HEA rule set produced an r^2 value of 0.49 and RMSE of 119.38.

Table 15. Information table for RANN and HEA models developed to forecast *Anabaena* abundance in Myponga reservoir (1test year)

RANN		HEA	
Input layer structure	Laguarre Axon	IF (((NO3*(70.856/PO4))+PO4)<=310.669)	
Processing Elements	15	THEN *Anabaena* = (((NO3*(TURB*23.904))*DO)+83.660)	
Transfer function	TanhAxon	ELSE *Anabaena* = (DO*((NO3*(TURB*18.027))*DO))	
Iterations	9000	Runs	50

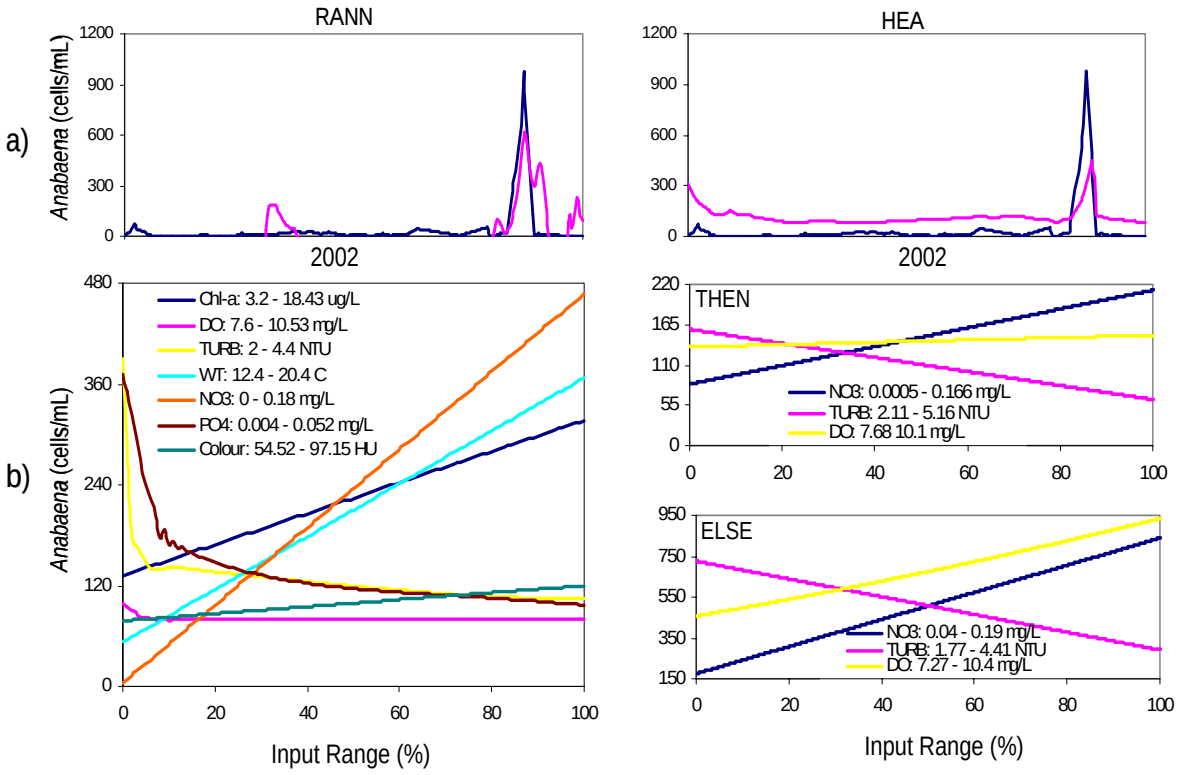


Figure 37. a) *Anabaena* forecasting results for RANN (left) and HEA (right) models tested on one year of data from Myponga reservoir, b) sensitivity analyses results from RANN (left) and HEA (right).

The sensitivity analyses produced from the methods can be seen in Fig. 37b. Of the six input variables, MIP analysis by RANN identified PO_4 concentration as the most influential (see Appendix E). From the six possible input variables, HEA detected that turbidity, DO, PO_4 and NO_3 concentrations were necessary for the accurate prediction of *Anabaena* abundance in Myponga reservoir. The sensitivity graph obtained by RANN show that highest *Anabaena* concentrations coincide with very low PO_4 concentrations and that increasing PO_4 levels occur in conjunction with decreasing *Anabaena* populations. Interestingly, NO_3 does not follow this same pattern of depletion with increasing *Anabaena* abundance. The sensitivity graphs produced by both RANN and HEA show that *Anabaena* abundance and NO_3 rise in unison. It is known that *Anabaena* is one of several cyanobacteria genera that are not solely reliant on the NO_3 concentration in the water, as they can fix atmospheric nitrogen for use. In fact, Boney (1989) states that cyanobacteria can be significant contributors to nitrogen budgets, and it is possible that this can occur in Myponga reservoir, explaining why NO_3 concentration increases as *Anabaena* populations grow. Results from the RANN sensitivity analysis show that *Anabaena* abundances increases with increasing water temperature. This follows what is known regarding *Anabaena* preferring the summer season, predominantly due to the warm water temperature and

associated thermal stratification. Both RANN and HEA demonstrate that *Anabaena* decreases with increasing turbidity. Turbidity, through light limitation, appears to be limiting *Anabaena* growth. *Anabaena* populations themselves are most often not large enough contribute significantly to the turbidity levels. The RANN and ELSE HEA analysis results showed no significant relationship between *Anabaena* and DO. The result from the THEN HEA analysis shows a positive relationship between DO and *Anabaena*. Colour and Chl-*a* concentration are both also shown to have a positive relationship with *Anabaena* abundance, as *Anabaena* can contribute to both variables.

5.4.2.2 Myponga reservoir validated for 1997 and 2002

The results of *Anabaena* forecasting using reservoir specific models developed by RANN and HEA can be seen in Fig. 38. Information on the models developed for this application is available in Tab. 16. The validation results produced from the application of these models to reservoir data from Myponga reservoir (1997 and 2002) show that the RANN model was most successful in forecasting *Anabaena* abundance. The RANN model indicates the presence of three large populations throughout the testing period, though underestimates two of the three and forecasts a false peak. The HEA model only clearly identifies one growth event, with some unclear indication of the first event and no forecast of any elevated population near the second peak event. Interestingly, while the HEA model does not appear to be useful it gives better error levels, r^2 value of 0.1 and an RMSE of 823.91, whilst the RANN model is visually much better but gives an r^2 value of 0.1 and an RMSE of 900.23.

Table 16. Information table for RANN and HEA models developed to forecast *Anabaena* abundance in Myponga reservoir (2 test years)

RANN		HEA	
Input layer structure	Laguarre Axon	IF ((Nitrate/PO4)<=3.873)	
Processing Elements	18	THEN *Anabaena* = (NO_3*((DO*(TURB*6.03))*TURB))	
Transfer function	TanhAxon	ELSE *Anabaena* = (DO*(NO_3*((DO*TURB)*(WT+TURB))))	
Iterations	20000	Runs	50

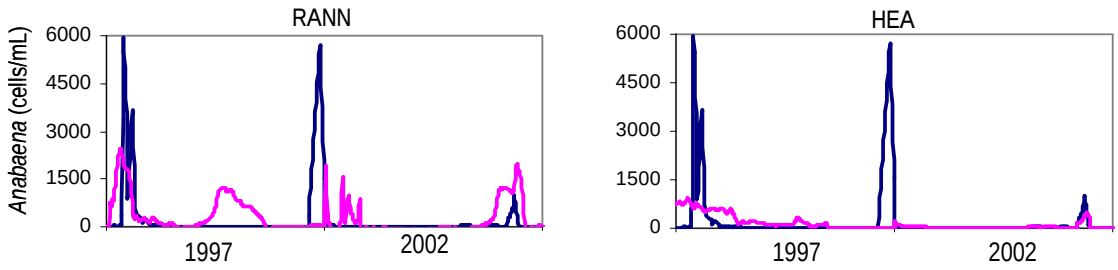


Figure 38. *Anabaena* forecasting results for RANN (left) and HEA (right) models tested on two years of data from Myponga reservoir

5.4.2.3 Happy Valley reservoir validated for 1997

The results of *Anabaena* forecasting in Happy Valley reservoir using reservoir specific models developed by RANN and HEA can be seen in Fig. 39a. Information on the models developed for this application is available in Tab. 17. The validation results produced from the application of these models to Happy Valley reservoir data from 1997 show that both models are able to identify where the major growth events occur throughout the year, although the magnitude was largely underestimated in each case, but more so by the HEA derived model. Although the RANN model provides a few false positives, visually it is the best model of the two as its forecasts are closer to the actual magnitude. Similarly to the results found for Myponga reservoir, this is reflected by a lower RMSE than the HEA rule set but interestingly not by the r^2 values. The results of validation of the RANN model produced an r^2 value of 0.56 and an RMSE of 858.7, whereas the HEA rule set produced an r^2 value of 0.63 and RMSE of 1017.9.

Table 17. Information table for RANN and HEA models developed to forecast *Anabaena* abundance in Happy Valley reservoir (1test year)

RANN		HEA	
Input layer structure	Laguarre Axon	IF ((NO_3*WT)>6.557)	
Processing Elements	25	THEN *Anabaena* = (((35.509/(TURB+PO_4))*25.062)+105.758)	
Transfer function	TanhAxon	ELSE *Anabaena* = (((189.596/exp(DO))*105.758)+56.489)	
Iterations	25000	Runs	50

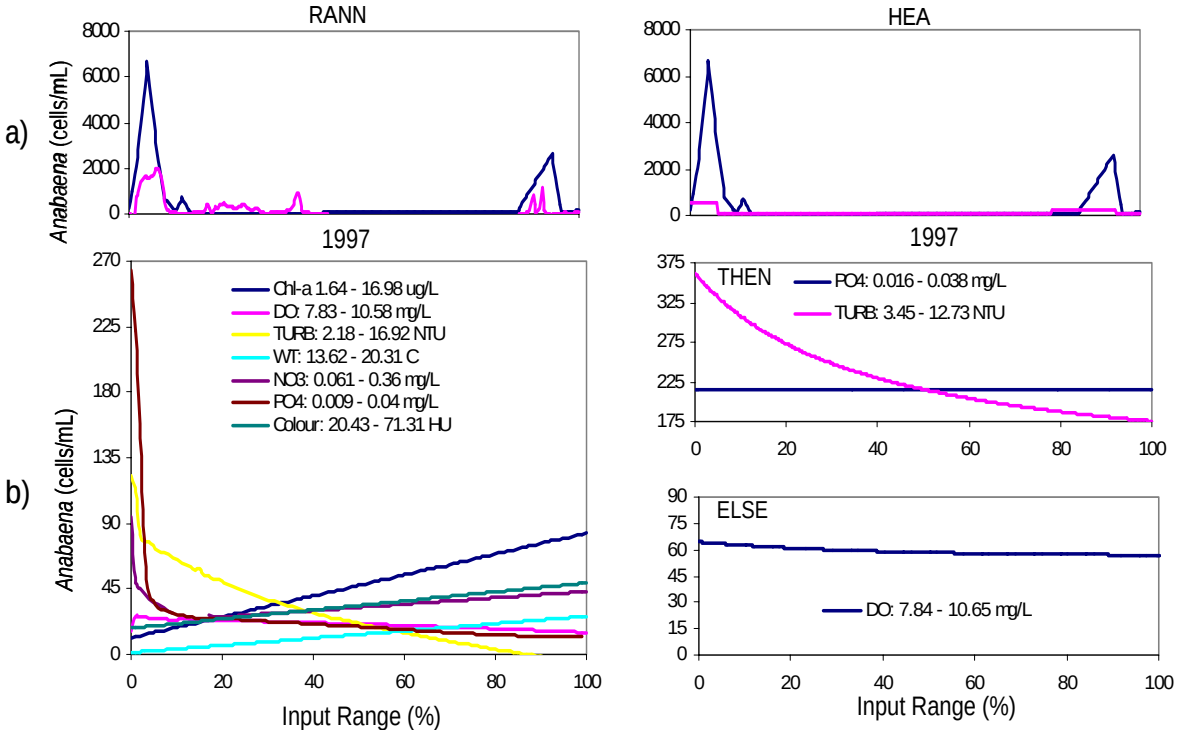


Figure 39. a) *Anabaena* forecasting results for RANN (left) and HEA (right) models tested on one year of data from Happy Valley reservoir, b) sensitivity analyses results from RANN (left) and HEA (right).

The sensitivity analyses produced from the methods can be seen in Fig. 39b. Of the six input variables, MIP analysis by RANN again identified PO_4 concentration as the most influential (see Appendix E). From the six possible input variables, HEA detected that water temperature, turbidity, DO, PO_4 and NO_3 concentrations were necessary for the accurate prediction of *Anabaena* abundance in Happy Valley reservoir. The sensitivity results for Happy Valley reservoir were similar to those for Myponga reservoir. The sensitivity graph obtained by RANN shows that highest *Anabaena* concentrations coincide with very low PO_4 concentrations and that increasing PO_4 levels occur in conjunction with decreasing Anabaena populations. As with Myponga reservoir, in Happy Valley reservoir NO_3 does not follow this pattern of depletion with increasing *Anabaena* abundance. The sensitivity graphs produced by RANN show that

Anabaena abundance and NO_3 rise in unison. Like the results for Myponga reservoir, *Anabaena* populations were shown to increase with warming water temperature and to decrease with rising turbidity. Both RANN and HEA show *Anabaena* abundance to not be significantly related to DO concentrations. As expected, increasing *Anabaena* abundance coincides with increasing Chl-*a* concentrations.

5.4.2.4 Happy Valley reservoir validated for 1997 and 2001

The results of *Anabaena* forecasting using reservoir specific models developed by RANN and HEA can be seen in Fig. 40. Information on the models developed for this application is available in Tab. 18. The validation results produced from the application of these models to data from Happy Valley reservoir (1997 and 2001) show that both models have difficulty in forecasting the spikes in *Anabaena* populations throughout the two test years. The RANN model was able to indicate slightly elevated abundances at the time of the major growth events, but underestimated them so much that the forecasts are useless from a management perspective. Interestingly, whilst RANN produced the best model visually, which gave an r^2 value of 0.12 and an RMSE of 802.3, the HEA model gave better error values of r^2 value of 0.46 and RMSE of 794.65 even though the models forecasts failed to indicate any of the growth events except the first one.

Table 18. Information table for RANN and HEA models developed to forecast *Anabaena* abundance in Happy Valley reservoir (2 test years)

RANN		HEA	
Input layer structure	Laguarre Axon	IF ((exp(DO)>=51.747)	
Processing Elements	20	THEN *Anabaena* = (((Colour/TURB)*((-7.635)/ln(|NO_3|)))+34.7)	
Transfer function	TanhAxon	ELSE *Anabaena* = ((Colour/DO)*329.66)	
Iterations	15000	Runs	50

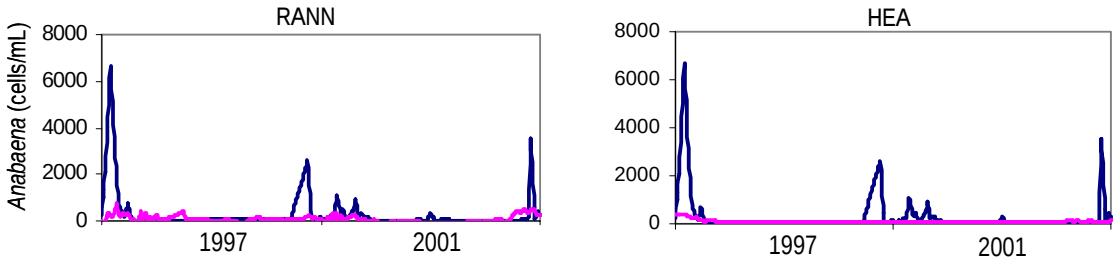


Figure 40. *Anabaena* forecasting results for RANN (left) and HEA (right) models tested on two years of data from Happy Valley reservoir

5.4.2.5 Myponga and Happy Valley reservoirs validated for 2002 and 1997 respectively

The results of *Anabaena* forecasting using the merged models developed by RANN and HEA can be seen in Fig. 41. Information on the models developed for this application is available in Tab. 19. The validation results produced from the application of these models to reservoir data from Myponga reservoir (2002) and Happy Valley reservoir (1997) show that both models have successfully identified the timing of the major *Anabaena* bloom, yet RANN significantly underestimated the magnitude of the population, whilst HEA was much closer to the actual abundance level. The first small peak of approximately 1000 cells/mL was suggested by both models, but RANN was most successful with regard to both timing and magnitude. The bloom event at the end of the testing period was poorly forecast by both methods. Although the HEA rule-set produced the best error statistics of an r^2 value of 0.55 and an RMSE of 578.8, visually the RANN model provided good results, except with regard to magnitude of the main peak, and gave an r^2 value of 0.34 and an RMSE of 656.9.

Table 19. Information table for RANN and HEA models developed using merged data to forecast *Anabaena* abundance in both reservoirs (2 test years)

RANN		HEA	
Input layer structure	Laguarre Axon	IF ((Chl-*a**(TURB*(TURB*Chl-*a*)))<=175.333)	
Processing Elements	31	THEN *Anabaena* = ((((WT*NO_3)/TURB)*(Chl-*a**(WT*Chl-*a*)))+243.84)	
Transfer function	TanhAxon	ELSE *Anabaena* = ((((WT/TURB)*WT)*((WT* NO_3)/TURB))+97.139)	
Iterations	15000	Runs	50

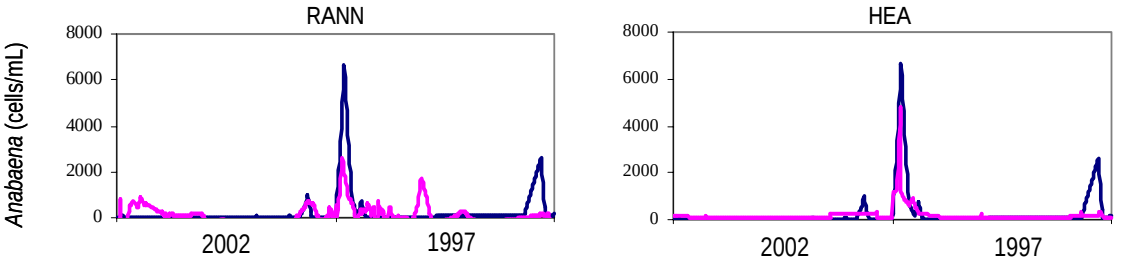


Figure 41. *Anabaena* forecasting results for merged RANN (left) and HEA (right) models tested on two years of data, one from each reservoir

Table 20. Summary of error levels for forecasting models developed in Chapter 5.

	RANN r^2, RMSE	HEA r^2, RMSE
Chl-*a* Myponga 1 test year (section 5.4.1.1)	0.77, 3.85	0.89, 2.88
Chl-*a* Myponga 2 test years (section 5.4.1.2)	3.68, 0.62	0.66, 4.17
Chl-*a* Happy Valley 1 test year (section 5.4.1.3)	0.81, 3.23	0.76, 3.56
Chl-*a* Happy Valley 2 test years (section 5.4.1.4)	0.48, 7.11	0.55, 6.44
Chl-*a* merged 2 test years (section 5.4.1.5)	0.66, 4.37	0.52, 5.07
Chl-*a* merged electronically measurable inputs only 2 test years (section 5.4.1.6)	0.22, 8.04	0.5, 5.3
Anabaena Myponga 1 test year (section 5.4.2.1)	0.36, 174.95	0.49, 119.38
Anabaena Myponga 2 test years (section 5.4.2.2)	0.1, 900.23	0.1, 823.91
Anabaena Happy Valley 1 test (section 5.4.2.3)	0.56, 858.7	0.63, 1017.9
Anabaena Happy Valley 2 test years (section 5.4.2.4)	0.12, 802.3	0.46, 794.65
Anabaena merged 2 test years (section 5.4.2.5)	0.34, 656.9	0.55, 578.8

5.4.3 Research progression

This section explains the progression of the research from RANN and HEA forecasting models validated on one year of independent data, through to two testing years, merged reservoir data sets for training and testing, and models developed using only electronically measurable merged reservoir data.

The Chl-*a* forecasting results from both 7-days ahead RANN and HEA models tested on one year of data were of high-quality. The results from applying RANN and HEA models to 7-days ahead forecasting of *Anabaena* abundance for one testing year were reasonably successful, particularly the model for Myponga reservoir. Most of the sensitivity analyses results revealed relationships complementary to literature findings; however two variables gave some interesting results. The sensitivity analyses suggested differences in the relationships with turbidity between the reservoirs. Increasing turbidity coincided with a decrease in algal abundance in Myponga reservoir and an increase in Happy Valley reservoir. Further research explained this result by revealing that turbidity in Myponga reservoir is largely caused by non-algal matter, and that turbidity can inhibit algal growth probably through light limitation. Happy Valley reservoir is a more productive system than Myponga reservoir and therefore algae are thought to contribute more to turbidity, which could explain why turbidity and algal abundance increase in unison. The sensitivity results looking at the relationship between DO and algal abundance were also interesting. It was shown by both methods that algal abundance was not particularly sensitive to changes in DO. With regard to most influencing parameters, RANN found PO_4 concentration to be the most important factor by far, regardless of whether the model was predicting Chl-*a* concentration or *Anabaena* abundance. NO_3 concentrations were also found to be an influencing factor. The least important factor was shown to be colour, though DO and turbidity were also shown to have relatively little influence.

After testing models on just one year of independent data, it was decided that this would not be a sufficient way of testing a models' forecasting capabilities. Results gained from just one test year could not give information on the models' capabilities in different conditions. In order to further test the models, yet maintain several years for training, two years of data were used for testing the models developed in this section. The test year from the original experiments were used in addition to another year, selected because of different patterns in the target variable. The second test year was not immediately prior or after the first test year, to ensure that conditions were not too similar.

The results of the RANN sensitivity analyses for models validated on one test year consistently showed that colour was the least influencing of all the available input variables. As such, colour was removed as an input from RANN models to be validated on two years of data, in an effort to reduce noise within the data set and promote better results. As HEA selects its own inputs, colour was kept available for inclusion. As a comprehensive sensitivity analysis, incorporating

both reservoirs and both Chl-*a* concentrations and *Anabaena*, was conducted on the models with one test year, and the relationships shown were largely consistent – it was felt unnecessary to include another replicated set of sensitivity analyses results for the models tested on two years of data.

Applying models developed by RANN and HEA to two test years showed that accuracy levels decreased overall (see Tab. 20). The models predicting Chl-*a* concentration continued to give useful results, with only slightly increased error levels. In contrast, the models developed to forecast *Anabaena* abundance lost a significant amounts of accuracy and the models became useless from a management perspective.

Developing a model that can be successfully tested on more than one year of data is necessary if the model may be used for management purposes. Whilst the results for the Chl-*a* models when tested on two years were good, the *Anabaena* forecasting models could not be used without improvement and it was thought that the addition of more training data could improve the generality of the models, for Chl-*a* and *Anabaena,* and allow them to better predict over longer test periods. The maximum amount of available data from each reservoir was already being used, so to achieve this increase in training data, the data sets from both reservoirs were combined to develop merged models. It was thought that a model would likely perform better on multiple test years if it could be trained with more data, and be able to learn a greater range of patterns and conditions. We have seen in the previous chapter that similar patterns were expressed in each reservoir and in this chapter the similarities between causal relationships regarding algal abundance could easily be seen by the sensitivity analysis results. Therefore, it was thought that the combination of these data sets would provide more similar patterns for the models to learn from and could produce a model with increased accuracy on multiple test years from both reservoirs. The development of merged data models like this require less development time than two reservoir specific models and can be applied to either reservoir.

The results of the models developed using merged data from both reservoirs and applied to a testing year from each reservoir were reasonably good, particularly for Chl-*a* concentration. The Chl-*a* forecasting results provided encouragement for the concept of creating a forecasting tool that would be generic for all lakes within a particular lake ecosystem category. Whilst the visual assessment of the *Anabaena* forecasting results is reasonably good, the error level indicated by RMSE is quite significant, suggesting that it will be more difficult to develop an accurate forecasting tool specifically for *Anabaena* abundance, in lakes similar in nature.

The promising results from the merged models designed to forecast Chl-*a* concentration gave encouragement for the concept of developing a merged model to forecast Chl-*a* concentration using only electronically measurable inputs. Electronically measurable inputs are those that can be measured and acquired, in real time, by a data logger and telemetry system for online monitoring, which is necessary for real-time forecasting. Although the RANN sensitivity analyses clearly showed that the nutrients, PO_4 and NO_3, were the most influential input variables, they were not able to be electronically measured and therefore could not be included in any model to be applied to real-time forecasting. The increase in training data enabled by the merged models could potentially compensate for the lack of nutrient input variables. The input variables made available to the models were: DO, water temperature, turbidity and conductivity.
Considering the limited input variables of DO, water temperature, turbidity and conductivity, both models did well to successfully identify the major trends in the Chl-*a* concentration over the two year testing period. Compared with the merged models created using additional nutrient data (see section 5.4.1.6), the RANN model had lower accuracy, however the HEA model produced results of the same standard and quality (see Tab. 20).

5.5 Discussion

Bio-inspired computational methods, RANN and HEA, were used throughout this chapter to develop forecasting tools for 7-day ahead predictions of Chl-*a* concentration and *Anabaena* abundance. The models were tested on 1 or 2 years of independent data, which had not been used during the training of the models.
Initially, reservoir specific models for both Chl-*a* concentration and *Anabaena* abundance were created and tested on one year of data. Results showed that these two variables could be well predicted using chemical and biological input variables (hence Hypothesis 1 can be accepted). Associated sensitivity analyses showed that relationships between the input and output variables were mostly aligned with literature findings and were consistent between reservoirs, bar one highlighted difference. PO_4 concentration was unanimously found to be the most influential input, which suggested that PO_4 management could be an effective means of controlling algal abundance in these reservoirs. Methods such as having inflow pass through phosphorus removing filters or the addition of a phosphorus precipitating or binding agent can potentially limit phosphorus in the reservoirs, thereby limiting algal growth. NO_3 was also considered influential, to a lesser degree, but it was very clear that nutrients were extremely important in the accurate

prediction of algal abundance. This suggested that it would be difficult to develop a forecasting tool that could accurately forecast algal abundance without relying on nutrient data.

The ability to trust a model's overall capability was questioned when the model's application has only been demonstrated for one year of data. In a compromise between further testing the models and keeping sufficient training data, two test years were used to assess the models. The results showed slightly higher error values, but still gave reasonably good forecasts for Chl-*a* concentrations. Results for *Anabaena* abundance were poor and demonstrated that the models could not be widely applied. Throughout the chapter it was found that *Anabaena* populations were significantly harder to predict than Chl-*a* concentrations, therefore Hypothesis 2 can be accepted. *Anabaena* abundance is more challenging to model as algal population dynamics are faster and more distinct than algal community dynamics (represented by Chl-*a*). The natural pattern of the genera would be to build in early summer for peaks in late summer and early autumn whilst warm water and stratification were present. However, management interrupts this by artificial mixing and, more so, by $CuSO_4$ dosing. For example, the nutrient conditions may suggest an impending bloom to the model, but in reality $CuSO_4$ dosing may occur when the population is in early stages of growth and the bloom that might have occurred without the interference of management does not take place, which introduces substantial error to the testing results. Although the training data involves dosing events there is no pattern the model is able to learn. Even though dosing mostly occurs in summer and autumn, it can happen at any time during the year and conditions may be different each time, and this makes the pattern of *Anabaena* occurrence quite different each year. The addition of some sort of measure of water stability to be included as model input could potentially greatly improve the forecasting ability and accuracy; however no data on stability of the water column was available at the time. In an attempt to improve the models with the data that was currently available, it was thought that by increasing the training data, the likelihood of similar patterns being learnt be the models would be increased. All available data for each reservoir was already being used for each reservoir specific model, so the data from the two reservoirs was combined to develop single models that could be applied to both reservoirs. The models developed using the merged data sets gave an overall improvement in the forecasting accuracy when tested on two years of data from both reservoirs compared to the reservoir specific models tested on two years of data. Applications to *Anabaena* forecasting in Happy Valley reservoir were significantly improved compared to previous applications using only one data set for training. Pattern analysis by KANN in Chapter 4

showed that the major water quality variables exhibit similar patterns in both reservoirs, therefore merging the data sets effectively doubled the examples from which the models could learn and extract relationships, making the models more robust and improving overall forecasting accuracy for application to both reservoirs.

The improvement in the accuracy for models applied to more than one test year offered by the merged data set method encouraged the investigation of developing a model that could be linked to online, real-time monitoring data. This required the use of electronically measurable input variables only. Using data from both reservoirs and inputs of water temperature, conductivity, turbidity and DO only, models were developed by both RANN and HEA. Although sensitivity analyses from earlier experiments showed that nutrient concentrations were the most important factors for algal abundance forecasting, the models using no nutrient data produced useful results. The HEA rule set was particularly successful, and produced equally good results as the merged models that included nutrient data as an input. For this reason, HEA was chosen as the best method to employ for the development of a real-time forecasting tool, continued in Chapter 7. Additionally, HEA derived rule-sets are explicitly represented, allow sensitivity analyses, are easier to implement and more widely functional than an RANN model. They can be easily applied in Excel and therefore do not require any costly specialised software and are easy to apply by an untrained user. Therefore, HEA provides a more simple and user-friendly option for application of a forecasting tool, and this is important.

A few interesting points arose from the results found in this chapter. In all results from RANN models, Happy Valley reservoir was always the most successfully forecast. The patterns in this reservoir appear to be easier for the network to learn than those in Myponga reservoir, which is interesting because Happy Valley dynamics are sometimes complicated by the addition of River Murray water.

In all RANN sensitivity analyses PO_4 concentration was found to be the most influential input variable, distantly followed by NO_3 concentration, in both reservoirs. Reservoir specific HEA models for Happy Valley reservoir were found to select PO_4 concentration to be part of the rule-set most often and models for Myponga reservoir selected NO_3 concentration to be used most in the rule-sets. This is interesting, as Happy Valley has higher levels of NO_3 and Myponga has higher levels of PO_4, therefore the models are selecting the most limiting nutrient as an important determining factor with regard to levels of algal abundance. Management have confirmed this

stating that PO_4 is the main factor determining the occurrence of a bloom event in Happy Valley reservoir (Daley & Ingleton 2006). Colour was found by RANN to be the least important variable, and although HEA often included colour in the forecasting rule-sets, it was never used in the threshold IF branch as the major determining factor.

The results in this chapter also highlighted the need to use more than one assessment technique or error measure and the importance of visual appraisals. On occasion, r^2 values have been high, yet RMSE values are also high. RMSE penalises errors in peaks and magnitude more than r^2 value, which explains such results but can make it difficult to rely on the statistics to demonstrate which models are superior. Also there are some aspects that are important to forecasting that the statistics cannot take into account. For example for management purposes, a model that forecasts growth events slightly before they occur is better than a slightly delayed forecast. Therefore it is always necessary to consider all error measures in conjunction with a thorough visual analysis of the forecast in relation to the actual observed data.

In conclusion, this chapter used RANN and HEA to develop models to forecast Chl-*a* concentration and *Anabaena* abundance. Both methods were found to be capable of producing useful forecasts 7-days in advance. Sensitivity analyses by both methods contributed to the understanding of relationships driving the algal dynamics. The progression of the chapter culminated in the development of models that required only electronically measurable inputs and could be linked to online monitoring data for real-time forecasting. Split-sample validation was useful for the purposes of comparison between reservoirs and forecasting methods, but ultimately can only provide snapshot results. A more thorough approach to model development and testing than those demonstrated in this chapter would be required to deliver models that could reliably be used operationally. This line of investigation is continued in Chapter 7.

6. RELATIONSHIPS OF CHLOROPHYLL-A AND *ANABAENA* DYNAMICS WITH PHYSICAL AND CHEMICAL INPUT VARIABLES

6.1 Introduction

This chapter features results from the combination of ordination and clustering by KANN and sensitivity analyses by RANN for knowledge discovery of phytoplankton dynamics as demonstrated by Welk et al. (2005) and Recknagel et al. (2006d). The research utilises water quality data of five variables from Myponga reservoir from 1997 to 2003 and focuses at relationships between water temperature, PO_4 and NO_3 concentrations and algal dynamics demonstrating algal preferences for specific temperature and nutrient ranges. The results demonstrate that qualitative relationships between water quality parameters discovered by KANN could be quantitatively determined by sensitivity analyses based on RANN. It can be concluded from this study that the combined applications of RANN and KANN provide a useful framework for rapidly extracting and explaining complex ecological relationships driving these dynamics. The so gained information can facilitate early warning and improve causal understanding of algal blooms.

6.2 Aims and Hypotheses

The aim of this research is to provide an understanding of relationships between algal abundance and certain input variables. Information on inter-variable relationships will be achieved by the combined results from two different analyses, utilising KANN and RANN, and results should provide convincing support of this technique for knowledge discovery.

The hypothesis is that qualitative relationships between water quality parameters discovered by KANN can be quantitatively determined by sensitivity analyses based on RANN, providing convincing information on the interactions between two variables.

6.3 Methods and Materials

6.3.1 Data

Research in this chapter utilises water quality data of five variables from Myponga reservoir from 1997 to 2003 (see Tab.21). The relationship between Chl-*a* and *Anabaena* abundance and the important water quality variables of water temperature, PO_4 and NO_3 concentrations was examined using this data.

Table 21. Myponga Reservoir database details

Experiment	Years of data used
Relationship: Water temperature and algal abundance Section: 6.4.1 Inputs: water temperature, Anabaena, Chl-*a*	1997-2003
Relationship: PO_4 concentration and algal abundance Section: 6.4.2 Inputs: PO_4, Anabaena, Chl-*a*	1997-2003
Relationship: NO_3 concentration and algal abundance Section: 6.4.3 Inputs: NO_3, Anabaena, Chl-*a*	1997-2003

6.3.2 Model Design

KANNs were developed to ordinate, cluster and visualise Chl-*a* or *Anabaena* abundance data with respect to water temperature and ranges of nutrients (as developed in 4.4.3).

RANN models were developed to forecast Chl-a concentration and *Anabaena* abundance (though the forecasting results are not of interest for this particular research and are not shown) and comprehensive sensitivity analyses were conducted to discover relationships between the input variables water temperature, PO_4 and NO_3 concentrations and the output variables (as demonstrated in Chapter 5).

The combination of the results from these two methods allows the discovery and analysis of relationships between the input and output variables in both a quantitative and qualitative manner.

6.4 Results

Sensitivity analyses produced in conjunction with the RANN models demonstrated relationships between the input variables and the output (Chl-*a or Anabaena*), and were combined with qualitative ordination and clustering by KANN.

6.4.1 Water temperature and algal abundance

Fig. 42 shows the relationship between *Anabaena* or Chl-*a* and water temperature. Both *Anabaena* and Chl-*a* are shown to increase with increasing water temperature.

Cyanobacteria are thought to prefer high water temperatures (Reynolds, 1984) and both the KANN (Fig. 42 top) and sensitivity curve (Fig. 42 bottom) confirm this in Myponga Reservoir, showing that *Anabaena* reaches maximum abundance above 19°C in both cases.

The RANN sensitivity curve (Fig. 42, bottom) also indicates that Chl-*a* increases steadily between 13 and 19°C, and levels off at its maximum between 19 and 20°C. These findings are backed up by the KANN-based ordination and clustering of Chl-*a* regarding 3 temperature ranges (Fig. 42, top). Whilst Chl-*a* concentration is lowest in the temperature range below 14°C, it reaches its maximum in the range above 19°C. The stimulating effect of temperature can be both direct and indirect. It may not be exclusively the water temperature that attracts the higher algal abundances, it may also be other conditions which coincide with the warmer weather. Thermal stratification enables buoyancy regulation by some types of cyanobacteria possessing gas vacuoles, allowing them to adjust their vertical position in the water column for access to solar radiation at the surface layers and nutrients near the thermocline (Reynolds, 1984).

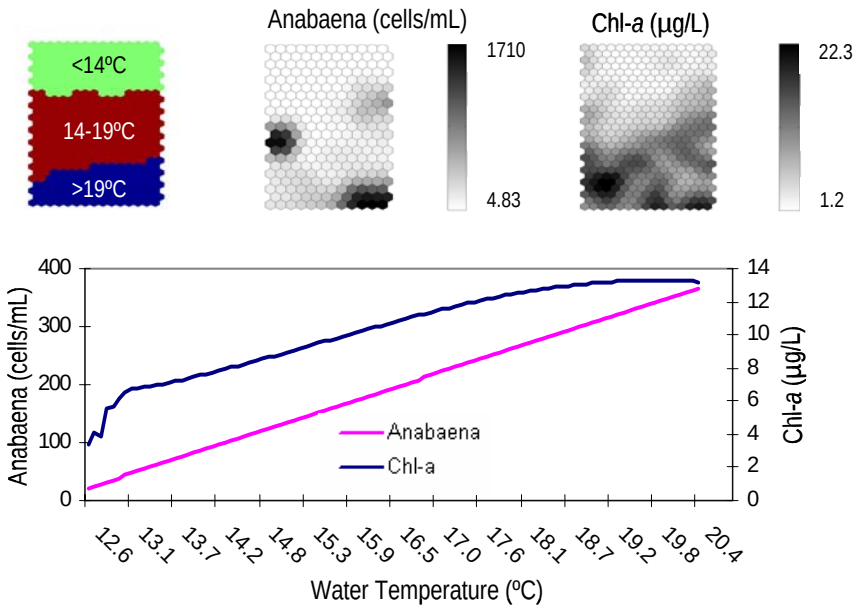


Figure 42. K-means map for temperature ranges (top left) with corresponding component planes for *Anabaena* and Chl-*a* (top middle and right) in Myponga Reservoir; sensitivity curve for *Anabaena* and Chl-*a* in response to temperature change (bottom)

6.4.2 PO_4 concentrations and algal abundance

Fig. 43 reflects the relationship between PO_4 concentrations and *Anabaena* or Chl-*a* concentrations. It reveals that highest Chl-*a* concentrations coincide with low PO_4 concentrations. The most probable explanation for this is that the large population of algae (represented by Chl-*a*) consumes large amounts of the nutrient and depletes the source pool. In Myponga Reservoir, minimal nutrient influx from the catchment is obtained during summer due to low rainfall and no runoff (Smalley, 1998), so the fact that highest Chl-*a* occurs at times of low PO4 levels, would be based on consumption of the nutrient by algae and no replenishment of supplies due to no inflow. Also contributing would be the fact that non-flagellate colonial green algae such as *Scenedesmus*, a dominant species in Myponga reservoir, are known to develop more often at lower phosphate levels.

Anabaena often peaks in summer at low PO_4 concentrations, which may be due to their capability to maximise nutrient uptake during times of depletion (Sommer, 1989). Fig. 43 shows major peaks for *Anabaena* coincide with low PO_4 concentrations in the water. PO_4 is similar to NO_3 in that, it is much more abundant during the rainy seasons when runoff increases nutrient concentrations and the reduced uptake by phytoplankton allows accumulation. Once phytoplankton growth increases significantly phosphate is quickly consumed, resulting in rapid depletion of the nutrient pool. This depletion is compounded by the fact that phytoplankton do not simply take only what is needed of

the nutrient for immediate utilisation. When phosphate levels are high many phytoplankton accumulate phosphate reserves greatly exceeding current requirements. This phenomenon, known as 'luxury uptake', further adds to the speed and magnitude of phosphate depletion. The phosphate reserves facilitated by luxury uptake, allows cell growth to continue for some time after the nutrient levels in the water have become very low (Boney, 1989). Cyanobacteria are known to store this excess uptake as polymeric polyphosphate (Paerl, 1988) and use it for maintenance and continued growth once phosphate concentrations have become limited, however this does not mean populations will continue to increase rapidly, as the new individuals will not have the phosphate reserves previously acquired by existing members of the population. This storage capability explains the presence of *Anabaena* during periods of low phosphate concentrations as demonstrated by both the NSNN (Fig.43 top) and the sensitivity analysis (Fig.43 bottom) showing maximum *Anabaena* abundance coinciding with phosphate concentrations below 0.02mg/L.

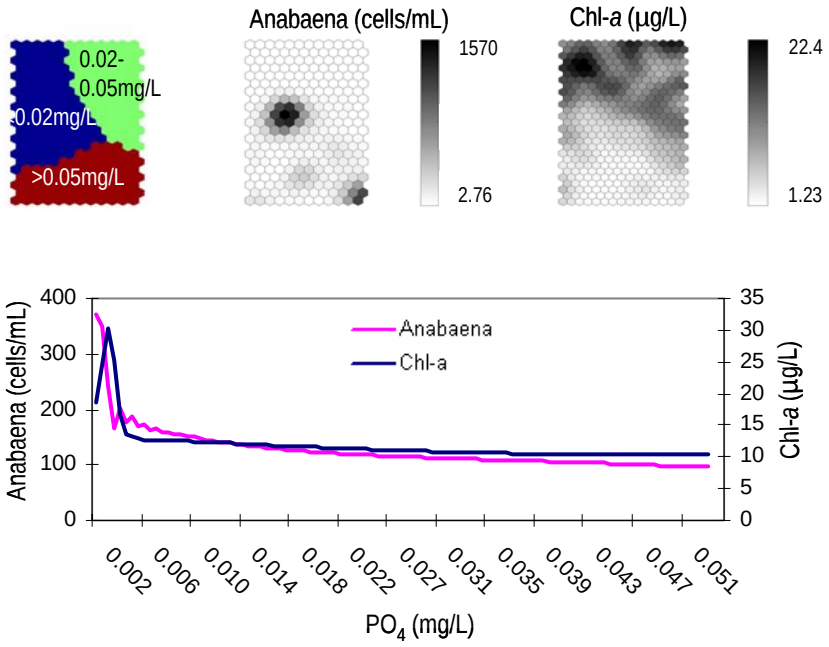


Figure 43. K-means map for PO_4 ranges (top left) with corresponding component plane for *Anabaena* and Chl-*a* (top right) in Myponga Reservoir; sensitivity curve for *Anabaena* and Chl-*a* in response to PO_4 change.

6.4.3 NO_3 concentrations and algal abundance

The sensitivity analysis in Fig. 44 (bottom) shows that Chl-*a* decreases with increasing NO_3 concentrations, and peaks at concentrations of 0.02 to 0.03mg/l. The corresponding ordination and clustering of Chl-*a* related to 3 ranges of NO_3 concentrations (Fig. 44, top) confirms these findings with Chl-*a* abundance peaking in an area corresponding to the lowest NO_3 concentration range (<0.1mg/L) on the k-means map. Like PO_4, NO_3 is an essential nutrient for algal growth, it is rapidly depleted when phytoplankton growth increases and peaks in spring and summer, causing the concentration in the water column to decrease (Boney, 1989), which is reflected by the results. Happey-Wood (1988) supports this, stating that the growth of numerous types of green algae (major contributors to algal composition in Myponga reservoir) is commonly accompanied by substantial decreases in the inorganic nitrogen pool of the surface waters of lakes.

However, not all phytoplankton are dependent on the nitrate availability in the aquatic environment. Certain cyanobacteria, including *Anabaena*, are known to possess the ability to fix atmospheric nitrogen and, therefore, do not have to rely on the nitrate pool in the water. As KANN (Fig. 44 top) and sensitivity analysis (Fig. 44 bottom) show, *Anabaena* peaks in abundance during times of high nitrate concentrations of 0.2m/L and above. It may be that little nitrate uptake is occurring due to the use of nitrogen fixation, allowing the source pool to be maintained at a high level. Boney (1989) even states that in some habitats, filamentous cyanobacteria make significant contributions to in lake nitrogen budgets.

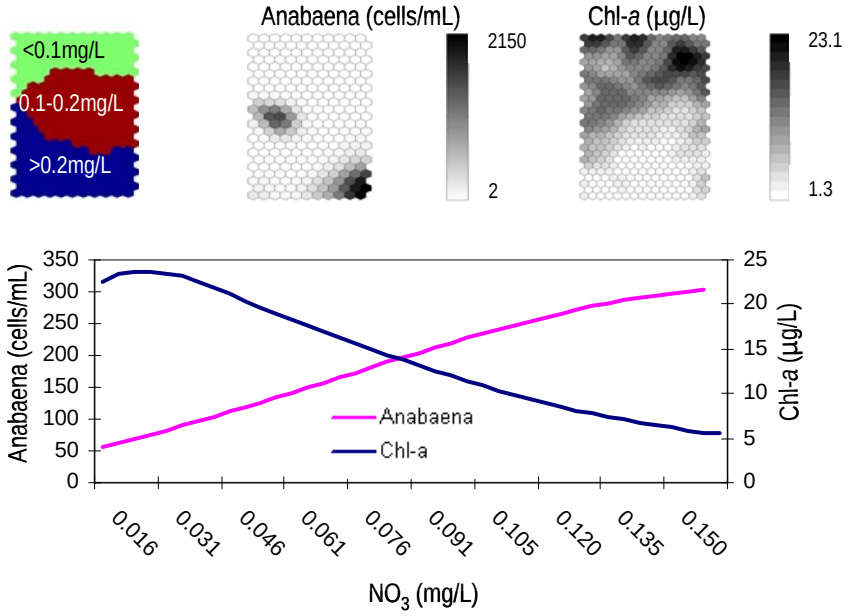


Figure 44. K-means map for NO_3 ranges (top left) with corresponding component plane for *Anabaena* and Chl-*a* (top right) in Myponga Reservoir; sensitivity curve for *Anabaena* and Chl-*a* in response to NO_3 change.

6.5 Discussion

The present study used a new approach for the exploration of ecological time-series both qualitatively and quantitatively. It can be concluded that the combined applications of RANN and KANN provide a useful framework not only for forecasting phytoplankton dynamics but also explaining complex ecological relationships driving these dynamics. Thus, the hypothesis that relationships described quantitatively by RANN can be illustrated qualitatively by KANN can be accepted. A major criticism of ANNs has been that they are 'black box' models, which give no indication of the processes involved in the modelled system. The use of sensitivity analyses performed on the training data of RANNs rectifies this issue by revealing how changes in the input variables affect the output, thereby giving RANN an explanatory quality in addition to predictive capabilities. The so gained information can facilitate early warning and improve causal understanding of algal blooms. This novel approach has been demonstrated in several peer-reviewed publications including Welk et al. (2005), Recknagel et al. (2006c), Recknagel et al. (2006d) and Recknagel et al. (2006e).

7. DEVELOPMENT OF RULE-BASED AGENTS FORECASTING ALGAL DYNAMICS USING HEA

7.1 Introduction

Predictive agents for algal populations have the potential to be powerful tools for early warning and operational control of harmful algal blooms in lakes and drinking water reservoirs.
This research aims to develop rule-based agents that are predictive for specific algal populations and generic for a particular lake category. To achieve this, two concepts are applied for the development of predictive rule-based agents of algal populations: 1) rule discovery by means of HEA, and 2) rule generalisation by means of merged time series data of lakes belonging to the same lake category.

This chapter demonstrates the application of a framework, using HEA and *k*-fold cross validation, for the extraction and development of rule-based forecasting agents from merged limnological time series data from Myponga and Happy Valley reservoirs, both classified as warm monomictic and eutrophic. These rule-based agents are created with the intention of operational use within a real-time environment, therefore testing of the agents must be thorough. Previous work demonstrated that split-sample validation, whilst useful for some purposes, could only provide snapshot results and a more thorough approach would be necessary to produce an agent that could confidently be applied to changing conditions (see chapter 5). This research suggests *k*-fold cross validation as a suitable method for this purpose.

Three rule-based agents are validated and discussed, which demonstrate forecasting 7 days ahead of Chl-*a* concentrations and *Anabaena* abundance in the Myponga and Happy Valley reservoirs. The framework was applied to the following case studies: 10 years of data from the eutrophic drinking water reservoirs Myponga and Happy Valley were merged to discover a predictive rule for Chl-*a* concentration, and validated for each year of data from both lakes; 18 years of data from the reservoirs were merged to discover a predictive rule for the abundance of the blue-green algae *Anabaena*, and validated for each year of data from both lakes. Chl-*a* is important as it is considered representative of algal growth and biomass within water bodies, and

forecasting of *Anabaena* is important, as it is a toxic nuisance species that needs to be kept to low levels.

The resulting rule-based agents proved to be both predictive and explanatory. It has been demonstrated that the interpretation of these rules can be brought into the context of empirical and causal knowledge on Chl-*a* dynamics as well as population dynamics of *Anabaena* under specific water quality conditions. The rule-based agents for Chl-*a* and *Anabaena* proved to be reasonably valid for the lake datasets and years tested in each case study.

The Chl-*a* agent developed using Myponga and Happy Valley data sets was applied to real-time data from Hope Valley reservoir, a water body from within the ecosystem category but that was not used during the training process, to examine possible applications of agents two independent data sets and real-time situations.

7.2 Aims and Hypotheses

The aims of the research in this chapter include:

- to develop a process to build predictive rule based agents to accurately forecast the timing and magnitude of Chl-*a* concentration or *Anabaena* abundance for a particular lake ecosystem category, to take another step toward generic modelling and forecasting.
- to use a testing method that is more thorough than those previously used (split-sample);

and finally

- to test an agent using real-time data from a reservoir not used during the training process – to test not only the models applicability to independent water bodies, but also its applicability to real-time forecasting.

In the pursuit of these aims, the validity of the following hypotheses will be established:

Hypothesis 1 – predictive rule-based agents developed using HEA with *k*-fold cross validation will give good forecasting results, with regarding to timing and magnitude, when tested on all years of historical data from Myponga and Happy Valley reservoirs. Since algal population dynamics are more rapid, distinct and challenging to model than algal community dynamics represented by Chl-*a*, the forecasting agents for *Anabaena* will be less accurate than those forecasting Chl-*a*.

Hypothesis 2 – the use of only electronically measurable input variables means that the rule-based agent forecasting Chl-*a* is compatible with real time data acquisition and forecasting, and because of this can be successfully applied to real-time data from Hope Valley reservoir.

Hypothesis 3 - the suggested agent development process, including training by merged data sets and parameter optimisation, adds generality to the resulting rule-based agents, and therefore the predictive rule-based agent developed for Chl-*a* forecasting can be applied to a reservoir (from the same lake ecosystem category) that was not involved during the training process.

7.3 Materials and Methods

7.3.1 Data

The agents were developed for potential use in a real-time environment therefore they were developed to use only variables that could be electronically measured (i.e. could be obtained form a data logger), which reduced the possible input variables substantially (see Tab. 22). Data selection was also dependent upon what parameters were available for a reasonable duration. As the experiment used merged data, data was sometimes limited by the duration of availability of particular inputs in each reservoir. The maximum amount (duration) of data possible was used for each experiment.

Conductivity data was limited to only 5 years in the Myponga reservoir database, which is why only 1999-2003 data was used for experiments including that variable. However, this was useful in demonstrating that a reasonable rule-based agent can be developed using limited data. As discussed in 5.3.2.3, Chl-*a* was not made available as an input for the Chl-*a* forecasting agent, however it was offered for input selection for the *Anabaena* forecasting experiment. In order to examine the influence of the parameter on accurate forecasting of *Anabaena* abundance, one agent was developed using a rule that included Chl-*a* and another was developed which did not include it. The latter also provides an alternative for situations where real-time Chl-*a* measurement is not available.

As outlined in section 3.5.1, interpolated daily values were used for used for most modelling applications, including the training and testing of the rule-based agents developed throughout this chapter, allowing the production of daily forecast values. However, the final rule-based agent developed for Chl-*a* forecasting, was tested on data collected in real-time via online monitoring of Hope Valley reservoir (CRC for Water Quality and Treatment project no. 2.0.2.2.2.1). As the real-time monitoring records measurements every 60 minutes, the data set encompassed

measurements from overnight. As the model was developed using data which was sampled between 10am and midday, it was thought best to use an average of the measurements between these times giving one daily value for this application.

Table 22. Summary of experiment specific data used throughout the chapter

	Chl-a Agent (1999-2003)		Chl-a Agent (9/11/06 - 20/4/07)	Anabaena Agent (1996-2003)	
Variable	Myponga	Happy Valley	Hope Valley	Myponga	Happy Valley
	mean/min/max	mean/min/max	mean/min/max	mean/min/max	mean/min/max
Water temperature (ºC)	16.17 / 8 / 24	17.51 / 9 / 24	21.5 / 17.2 / 25.4	16.23 / 8 / 24	17.37 / 9 / 26
Turbidity (NTU)	3.37 / 0.74 / 8.5	10.81 / 2.8 / 30	4.2 /0.9 /10.3	3.33 / 0.72 / 10	10.7 / 1.7 / 41
Conductivity (μS/cm)	620.3 / 525.2 / 781	642.9 / 436 / 1030	595.5 / 504 / 709		
Dissolved oxygen (mg/L)	8.82 / 3 / 12.6	9.03 / 5.1 / 12.8	6.47 /4.3 / 9.92	8.81 / 3 / 13.16	9.16 / 2.2 / 21.4
Chl-*a* (μg/L)	7.72 /0.4 / 36	10.4 / 1.1 / 41	11.1 / 3.1 / 67.4	8.55 / 0.5 / 36	10.4 / 0.8 / 66.5
Nitrate (mg/L)				0.09 / 0 / 0.337	0.22 / 0.01/ 0.63
Anabaena (cells /mL)				146.4 / 1 / 13300	131.96 / 0 / 6600

7.3.1 Model Design

7.3.1.1 Framework for development of generic rule-based agents

A framework was established for the development of rule-based agents for predicting algal dynamics in lakes and reservoirs belonging to the same lake ecosystem category. The framework describes the method for the application of HEA (the HEA method is described in detail in 2.3.1 and 3.5.3, and is used as demonstrated in chapter 5) to extract predictive rules from merged time-series data from same category lakes and features the *k*-fold cross-validation technique. *k*-fold cross validation, according to Kohavi (1995) includes *k*-fold data partitioning and the consecutive use of each part of the data for both training and validation, where eventually all examples are used for training and validation.

Initially, HEA technology and *k*-fold cross validation were used to search for the best performing forecasting rule set using merged time-series water quality data. Subsequent steps involve the further optimization of numerical parameters in the rule set and the inclusion of functions of water temperature. The final rule-sets that are extracted from the process are referred to as rule-based agents.

Fig. 45 shows the framework devised for using HEA to develop rule-based agents from merged time series data of same category lakes by means of k-fold cross-validation.

The framework and process is summarised below:

(1) determine best performing predictive rule for lakes by iterative *k*-fold training and validation using merged time-series data. This step begins with a predetermined number of rules being discovered, and the best performing rule set is determined by visual analysis and r^2 values.

(2) freeze the structure of best performing rule and determine best performing parameter values within the rule by iterative *k*-fold training and validation. As with step one, this step produces a predetermined number of parameter values.

(3) determine mean and range of parameter values.

(4) determine best performing water temperature functions within the established range, by iterative *k*-fold training and validation.

(5) substitute constant parameter values in rule with best performing water temperature functions.

(6) freeze generic rule and validate for each year and lake.

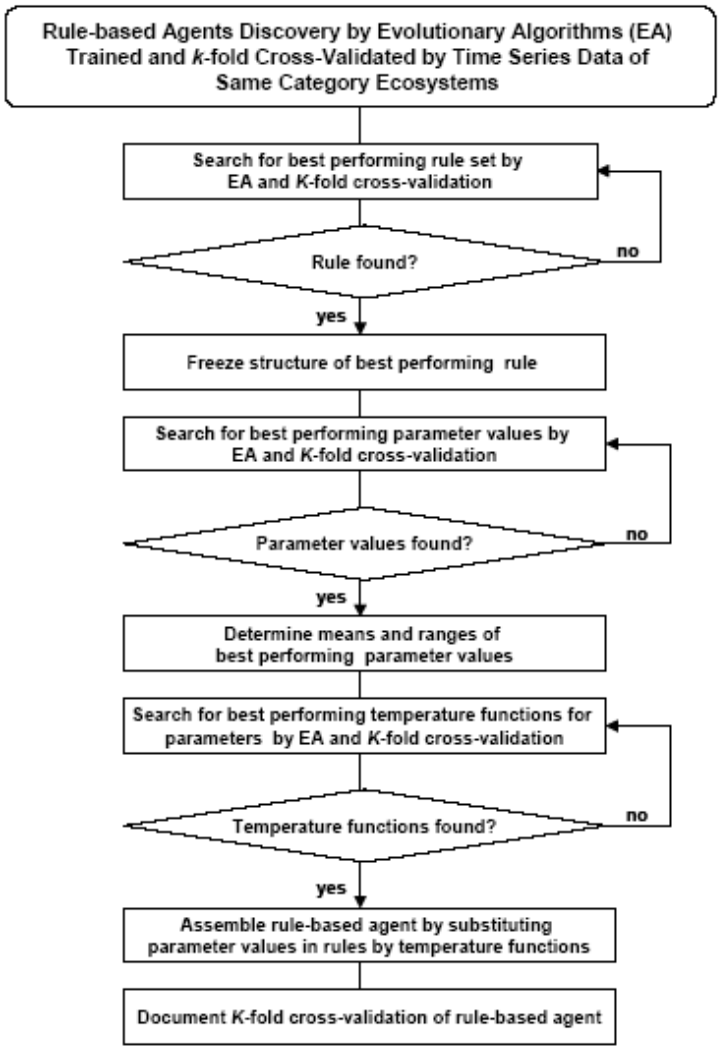


Figure 45. Development of rule-based agents by means of HEA and k-fold cross validation

To expand somewhat on the process summary, firstly HEA technology and *k*-fold cross validation were used to search for the best performing forecasting rule set using merged time-series water quality data. When the best performing rule (based on rule appropriateness and simplicity, RMSE, r^2 and visual assessment of measured vs predicted data) was found, the structure of that rule was frozen. However the numerical parameter values within the rule set remained variable, in order to determine the best performing parameter values via *k*-fold cross validation. The range and means of the parameter values were also derived from this step. The parameter values within the rule were then substituted with best performing water temperature functions within the pre-determined range (if the function gives rise to a value that falls outside the range – the mean value is used). The water temperature functions were derived by HEA and the best performing functions were determined by *k*-fold cross validation. The resulting form of the rule was frozen and is referred to as a rule-based agent, which could be applied to lakes of similar character and state. This concept is, in part, derived from the notion of adaptive agents for simulation of evolving species abundance and succession in aquatic environments (Recknagel 2003b), as they are agents that can respond to their environment, allowing the rule to adapt to suit the particular water body it is being applied to, and can therefore be successfully applied to more than one water body within the same ecosystem category.

The decision to replace constant values with temperature functions came after a series of preliminary experiments not shown here. Initially, the process discovered the mean parameter values, via k-fold cross validation and froze these values, leaving this as the rule based agent. Later it was postulated that the inclusion of water temperature functions, where constant mean values originally were, would make the rule more adaptable and better able to reflect conditions in the water body as they occur. Comparisons of results from the application of the rule using the mean values vs the results from the application of the rule set using temperature functions showed a slight improvement in accuracy. Although the improvement was only small in this instance, it was thought that when the rule set was applied to water bodies that were not included in the training process, this feature would increase the likelihood of successful application. Whilst functions of other variables could have potentially been used, it was decided that water temperature was the best option for a number of reasons. Importantly it is readily available for most water bodies and it captures seasonality and season-induced change. Water temperature is a very influential variable upon which many in lake processes are dependent or driven by, including algal growth.

7.3.1.2 Lake ecosystem categories

Ideally, the rule-based agents for algal population dynamics would be generic to some extent and applicable to more than one water body. As lake ecosystems exist in a broad spectrum of structures and functioning determined by climate, morphometry and trophic state, it is near to impossible to create a model that can be applied universally. Therefore it is suggested that rule-based agents be developed for application to water bodies that belong to the same ecosystem category, as suggested by Recknagel et al. (2007b). This is considered a good compromise and contribution toward generic modelling. Recknagel et al. (2007b) defined the lake categories by trophic state and circulation type, assuming that circulation type reflects climate conditions and morphometry to some extent, whilst the trophic state indicates habitat properties and community structures. Due to the similarities in key lake characteristics, it can be assumed that water bodies within a lake ecosystem category encounter similar ecological behaviours, health issues and management options. In this study only one lake ecosystem category (warm monomictic and eutrophic) is investigated, however other work including Recknagel et al. (2007b) and Welk et al. (2007) has looked at different categories. After successful validation of rule-based agents on Myponga and Happy Valley reservoirs, it was necessary to test the generality of the agent by applying it to real time data obtained from the Hope Valley reservoir, a water body that belongs to the same lake ecosystem category, but was not used during the training process.

7.4 Results

As a result of applying HEA within the *k*-fold cross-validation framework, according to Fig. 44, to time series data from two reservoirs belonging to the same lake ecosystem category, three rule-based agents have been developed. In accordance with the framework, the structure of the best performing rule-set was first discovered, and then the best performing temperature functions were determined (represented by P_1, P_2 etc. in the rule set and expressed in full below the equation). A rule-based Chl-*a* agent and two rule-based *Anabaena* agents were discovered for the two warm-monomictic and eutrophic reservoirs Myponga and Happy Valley. The agents were determined by the best performing IF-THEN-ELSE rule-sets with the best performing temperature function for parameters P_i, evaluated by the lowest RMSE, the highest r^2, and the visual closeness between measured and predicted data.

7.4.1 Rule-based Chl-*a* agent

A total of ten years of daily interpolated data from Myponga and Happy Valley reservoirs were merged to apply HEA with 5-fold cross-validation for developing a rule-based Chl-*a* forecasting agent (see Tab. 21). Only electronically measurable limnological variables of water temperature (WT), turbidity (TURB), conductivity (COND) and dissolved oxygen (DO) were used to develop the agent so that the extracted rule-set would be compatible with data obtained from a data logger, and therefore could be used to provide real-time forecasting. In order to design the agent for 7-day ahead forecasting, a time lag of 7 days was imposed between the input and output data. As a result, the rule-set for 7-day ahead forecasting of Chl-*a* documented in Fig. 45 was discovered and found to perform best.

The IF condition of the rule provides a threshold value for water temperature. The rule indicates that:

IF water temperature is less than 17.883°C,

THEN the Chl-*a* concentration is calculated by the equation:

$$\text{Chl-}a = \exp(\text{WT} / (P_1 / \text{DO} + P_2))$$

ELSE, the following equation is used:

$$\text{Chl-}a = \exp(\text{WT} / (P_3 / \text{DO} + \text{COND} / P_4))$$

Clearly, the rule distinguishes between conditions of slow algal growth characterised by cooler water temperatures and high oxygen concentrations typical for winter in South Australia, and conditions of fast algal growth in accordance with the higher water temperatures and low oxygen concentrations in summer. These relationships can also be demonstrated by the input sensitivity plots for the then and the else branches of the rule in Fig. 46. Fig. 46a and b both indicate that warmer water temperatures stimulate algal growth. However, Fig. 46a indicates that at water temperatures below 17.883°C Chl-*a* concentrations increase relatively slowly in response to increasing water temperatures, whereas at water temperatures above 17.883°C (Fig.45b), Chl-*a* concentration is shown to increase rapidly with increasing water temperatures. Fig 46a and b also show that Chl-*a* concentrations increase quite rapidly with growing oxygen concentrations across the entire spectrum of oxygen levels found in both branches of the rule (7.14-10.26mg/L). The oxygen levels during high algal growth (the ELSE branch, Fig. 46b) are shown to be lower than those during slow growth, owing to the fact that the larger algal community would have a higher biological oxygen demand. The sensitivity analysis revealed an interesting relationship between Chl-*a* concentration and conductivity, showing that algal biomass is highest when

conductivity is lowest. This could suggest that summer rainfall events bring an influx of nutrients, which triggers bloom events (increases Chl-*a* concentration), whilst also diluting conductivity levels, resulting in the described relationship.

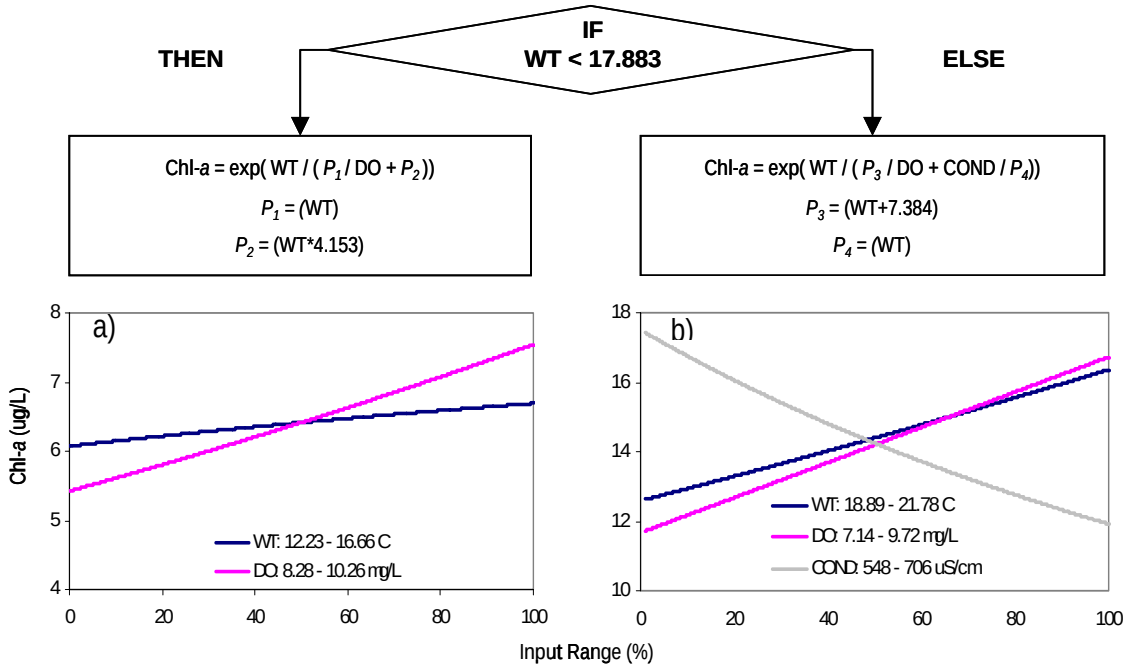


Figure 46. Structure of the rule-based Chl-*a* agent for Myponga and Happy Valley reservoirs. a) input sensitivity of the THEN branch, b) input sensitivity of the ELSE branch.

Fig. 47 provides the validation results of the rule-based agent for 7 days ahead forecasting of Chl-*a* concentrations in Myponga and Happy Valley reservoirs for the years 1999-2003. Generally the agent predicts the timing of peak events very well for both reservoirs. However on occasion the agent overestimates magnitudes of years with relatively low peak events (such as 2000 and 2003 in Myponga reservoir) and underestimates magnitudes of years with relatively high peaks (such as 2002 in Myponga reservoir and 2001 in Happy Valley reservoir). Despite these inaccuracies, the overall r^2 value of the linear regression between measured and calculated data of all 5 years for Myponga reservoir amounts to 0.42 and for Happy Valley reservoir 0.38. The validation results in Fig. 47 can be considered quite successful, as they have been achieved despite the highly stochastic environments in each of the study sites caused by periodic artificial mixing and occasional $CuSO_4$ dosing. These algal management strategies affect the ecology of both lakes, with the latter rapidly and drastically impacting upon the algal population abundances. The algicide treatment in particular can cause large errors in forecasting accuracy when validating the model against historical data. The agent may detect conditions that would normally

result in a bloom event and forecast algal populations accordingly, however if in reality $CuSO_4$ dosing has been carried out, there will obviously be a discrepancy between observed and predicted conditions.

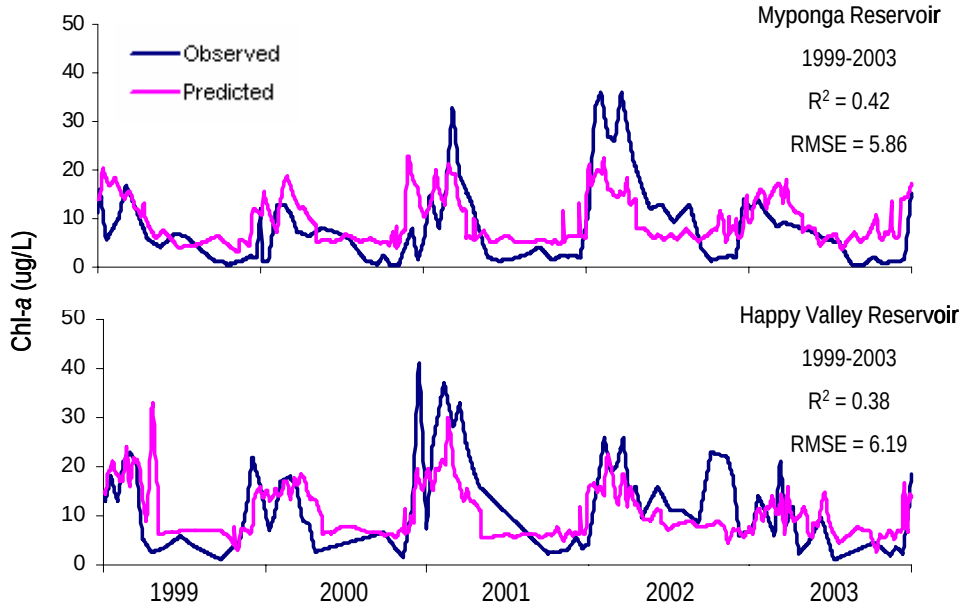


Figure 47. Validation results of the rule-based Chl-a agent for Myponga and Happy Valley reservoirs for all data (1999-2003)

7.4.1.1 Validation on Hope Valley real-time data

The rule based Chl-*a* agent was tested on 160 days of real-time data collected from Hope Valley reservoir. Hope Valley reservoir can be classified as a warm monomictic and eutrophic water body, along with Myponga and Happy Valley reservoirs. The purpose of testing the agent on this data set was two-fold: 1) to investigate whether the agent could be generic for lakes that were not involved in the training process but belong to the same lake ecosystem category as those used for model training and 2) to see how well an agent trained with interpolated data can be applied to real-time data.

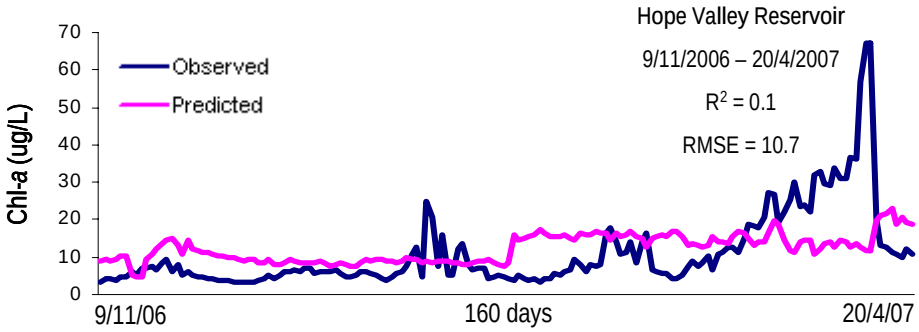


Figure 48. Validation results of the rule-based Chl-*a* agent applied 160 days of real-time data from Hope Valley

Fig. 48 provides the validation results of the rule-based agent for 7 days ahead forecasting of Chl-*a* concentrations in Hope Valley reservoir for 160 days (9/11/2006-20/4/2007). The overall r^2 value of the linear regression between measured and calculated data of all 160 days amounts to 0.1. The validation results show that the agent predicts the moderate Chl-*a* values well but fails to identify peak events of greater than 20μg/L. It is considered an achievement that a forecasting agent, derived from water quality data from other reservoirs, was able to predict Chl-*a* values within the correct range. It can be seen, particularly in the first 80 days, that the Chl-*a* values that are predicted are reasonably close to the measured values. In fact, with the exception of the two main peaks, the accuracy of the forecast Chl-*a* values would be acceptable. Several reasons can be proposed for the agents shortcomings. The largest peak event experienced in Hope Valley during the test data reaches a maximum Chl-*a* value of 67.4μg/L, however the maximum value from the training data set consisting of data from Myponga and Happy Valley reservoirs is 41μg/L. Therefore, the agent has had no exposure to or experience with the conditions and patterns coinciding with any event over approximately 40μg/L, which helps to explain why the major peak event is not well forecast. Obviously training the agent with data that encompassed the entire range of conditions experienced in the test lake would provide improved results and this may prove to be a problem that arises when an agent is applied to water bodies that were not involved during the training process.

Another challenge presented to the model was to match the real-time daily Chl-*a* concentrations, which are much more dynamic than the daily values derived from interpolated weekly or fortnightly data used to train the model. In order to understand the difference between the interpolated data used for training, and the real-time data from Hope Valley, Fig. 49 compares the 160 days of real-time data from Hope Valley with 160 days of interpolated data from the same period in Myponga reservoir. Obviously, a manual sampling regime giving weekly, fortnightly and monthly data points misses much variation within the water quality factors, as opposed to real-time monitoring. It can clearly be seen that the data used to train the agent is much smoother than the rapid rise and fall of the real time data. It was thought that this difference in dynamics between the training data and the testing data may have created some difficulty for the agent. Although the validation results in Fig.48 show that the agent can forecast day to day variations well, it is expected that results would improve if real-time data was included in the training data set.

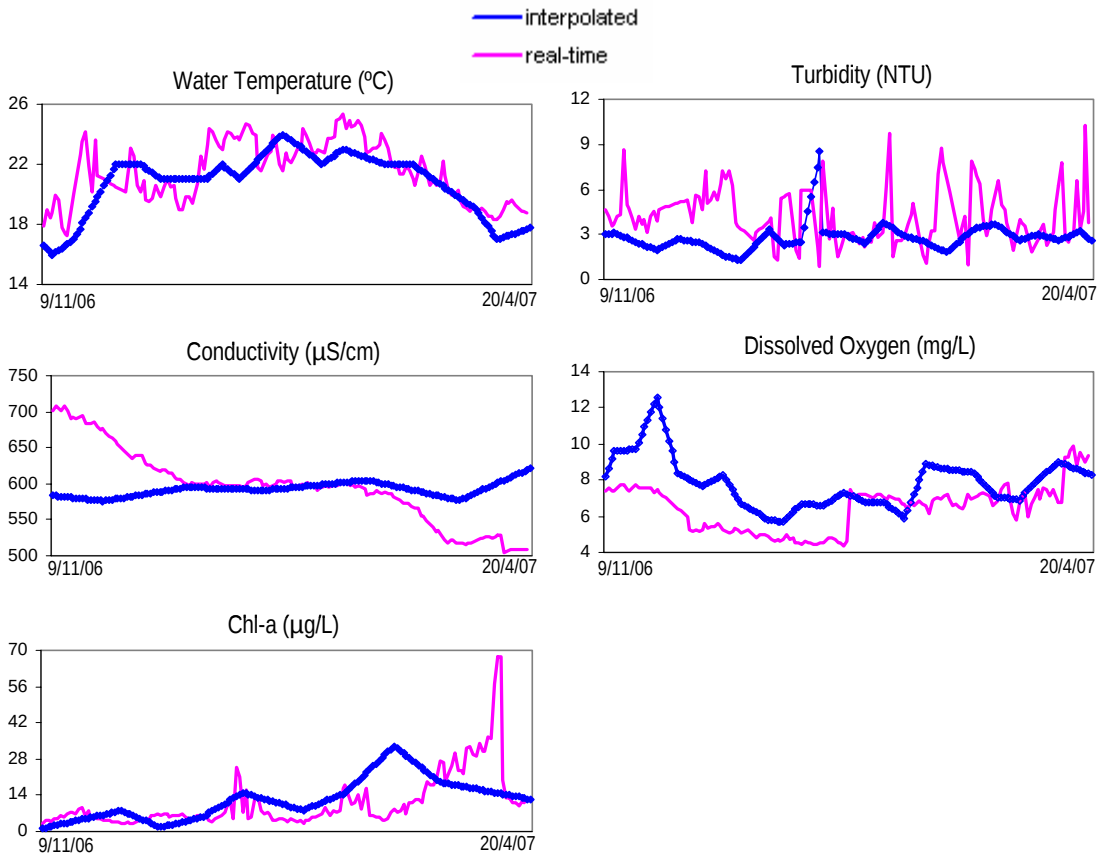


Figure 49. Comparison of real-time and interpolated data for the same period of the year

Although the Chl-*a* forecasting agent did not appear to achieve the same level of accuracy as some other examples of real-time forecasting discussed in section 2.6, the concept has some qualities that have the potential to provide a simple and improved alternative concept for real-time forecasting, with broader applicability. For example, Muttil and Lee (2005) used genetic program to discover a rule to forecast coastal algal blooms (using Chl-*a* fluorescence) with reasonable accuracy over a test period of approximately 5 months (RMSE of 165 and correlation coefficient of 0.85). However, this application used previous Chl-*a* as an input, produced a very complex rule (see Eq.3, section 2.6) and gave only forecasts only one day in advance. Subsequent testing of the rule for 3 days advanced forecasts showed a drop in accuracy to RMSE = 229 and correlation coefficient= 0.70 over the test period. Furthermore, the rule set has only been tested on the one location. The predictive Chl-a agent from this research has been much more stringently tested and achieved an average r^2 value of 0.4 for 7-day ahead forecasting of Chl-*a* when tested on ten years of data from the two reservoirs it was trained by, and was also capable of forecasting reasonable levels of Chl-*a* in a completely independent data set. With continued research and development into the suggested concept, the rule based agent for predicting Chl-*a*

concentrations in numerous sites belonging to a specific lake ecosystem category should become a viable option for real-time forecasting. It could provide one simple rule set that can be applied to a number of water bodies to give one weeks warning of impending bloom events with reasonable accuracy.

7.4.2 Rule-based *Anabaena* agent

A total of sixteen years of daily interpolated data from Myponga and Happy Valley reservoirs was merged used to develop rule-based *Anabaena* forecasting agents (see Tab. 22). Only electronically measurable limnological variables of water temperature (WT), turbidity (TURB), nitrate (NO_3), chlorophyll-*a* (Chl-*a*) and dissolved oxygen (DO) were used to develop the agent so that a rule-set would be extracted that would be compatible with data obtained from a data logger, and therefore could be used to provide real-time forecasting. However, unlike the forecasting agent for Chl-*a*, the *Anabaena* agents were not tested on real-time data from Hope Valley reservoir. Although NO_3 probes are now available for some data loggers, and therefore NO_3 can be considered an electronically measurable variable, real-time monitoring at Hope Valley did not include a NO_3 probe and therefore the agent was not able to be tested in a real-time situation. As outlined in 7.3.1, Chl-*a* was offered for input selection for the *Anabaena* forecasting experiment. In order to examine the influence of the parameter, one agent was developed using a rule that included Chl-*a* and another was developed which did not include it. The latter also provides an alternative for situations where real-time Chl-*a* measurement is not available. In order to apply the agent for 7-day ahead forecasting, a time lag of 7 days was imposed between the input and output data. As a result, 2 rule-sets for 7-day ahead forecasting of *Anabaena* (with and without Chl-*a* concentration as an input) were discovered, as documented in Figs. 50 and 52.

7.4.2.1 Rule-based *Anabaena* agent, including Chl-*a* as an input

The IF condition of the rule provides a threshold value determined by Chl-*a* concentration and DO. The rule indicates that:

IF (Chl-*a**DO)*DO is less than 369.24,

THEN the *Anabaena* abundance is calculated by the equation:

$$Anabaena = (DO*(((Chl\text{-}a * DO) / TURB)*DO))$$

Otherwise, by the ELSE equation:

$$Anabaena = \exp((\ WT + NO3\ ) / (TURB / NO3\ ))) * ((\ NO3\ *(P1)) + P2))$$

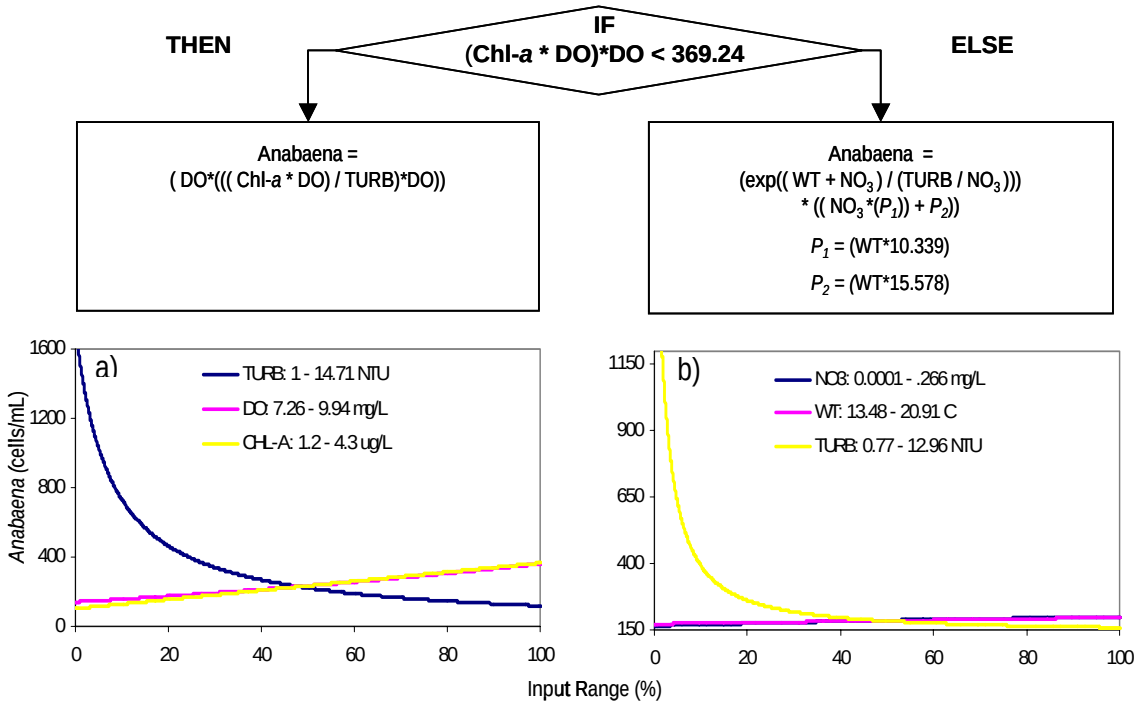


Figure 50. Structure of the rule-based *Anabaena* agent (inc. Chl-a) for Myponga and Happy Valley reservoirs a) input sensitivity of the THEN branch, b) input sensitivity of the ELSE branch

The rule suggests that Chl-*a*, DO, NO_3 concentrations as well as WT and TURB, are all important criteria to stimulate the growth of *Anabaena*. The causal relationships between these variables and *Anabaena* were further revealed by their sensitivity curves shown in Fig. 50a and b. The THEN branch of the rule encompasses *Anabaena* abundance of below approximately 1600cells/mL, whereas the ELSE branch covers abundances of 2200 cells/mL and below, but the graph was altered and the range reduced so that all variables could be better seen. The input sensitivity for the THEN branch of the rule (Fig. 50a) indicates that increasing DO and Chl-*a* concentrations coincide with increasing *Anabaena* abundance, presumably because the algae contributes to these variables through photosynthetic activity and presence respectively. Alternatively, the input sensitivity for the ELSE branch (Fig. 50b) shows positive relationships between NO_3 concentrations and WT and *Anabaena* abundance. *Anabaena* species are not reliant upon in-lake NO_3 concentrations, as they can fix atmospheric nitrogen, and therefore may have no negative impact on the levels of the nutrient in the water and, in fact, can contribute significantly to the nitrogen budget in the water body. Increasing WT is known to stimulate algal growth and cyanobacteria in particular are known to prefer warm waters (Reynolds 1989). Interestingly in both the THEN and ELSE branches, TURB is shown to have a negative

relationship with *Anabaena* abundance, with the higher *Anabaena* abundances being related to a slightly lower NTU range. From these results it would appear that increase turbidity is a factor that limits *Anabaena* growth, possibly through light limitation. The results of this sensitivity analysis between *Anabaena* and input variables concur with those discussed in section 5.4.2.

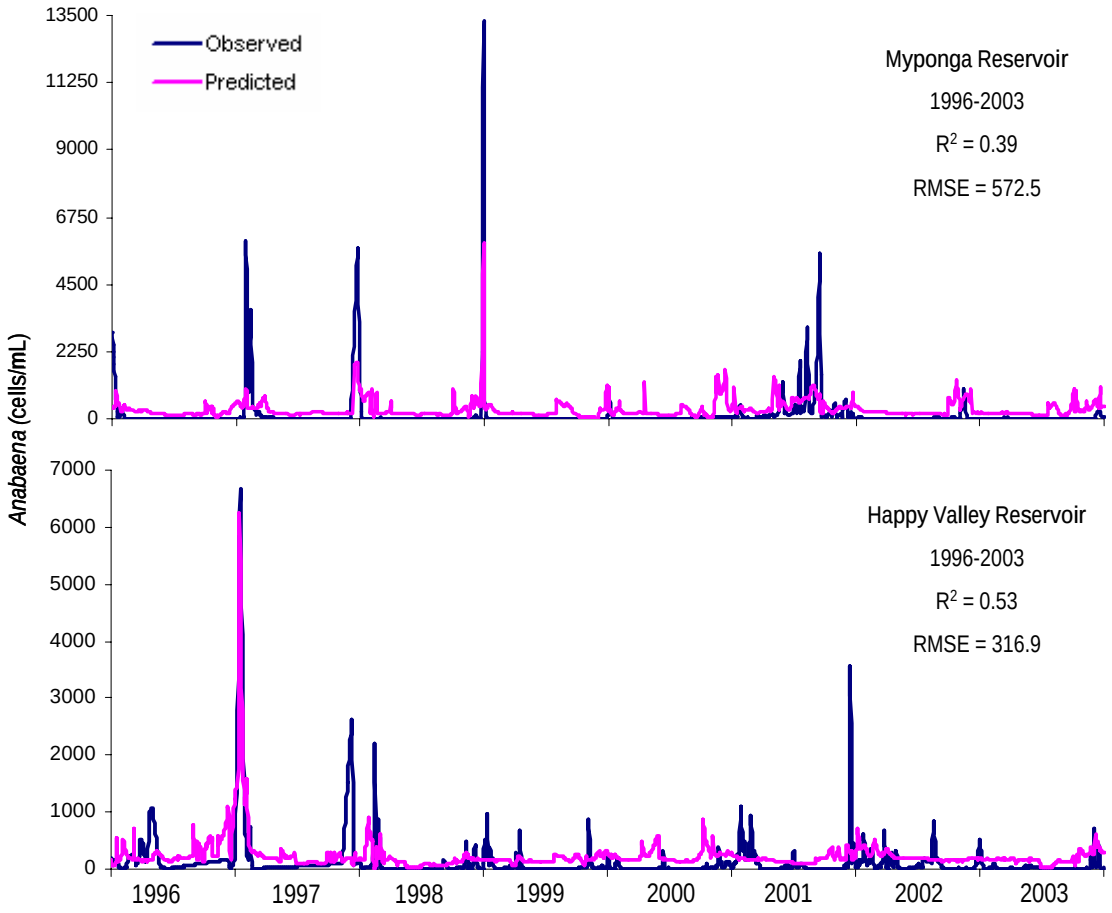


Figure 51. Validation results of the rule-based Anabaena agent (inc. Chl-a) for Myponga and Happy Valley reservoirs (1996-2003)

Fig. 51 provides the validation results of the rule-based agent for 7 days ahead forecasting of *Anabaena* abundances in Myponga and Happy Valley reservoirs for the years 1996-2003. Generally the model predicts the correct timing of peak events, particularly for Myponga reservoir but drastically underestimates magnitudes of events with extremely high peaks (such as 1997 and 1998/1999 summer peak in Myponga reservoir). For Happy Valley reservoir, the model forecasts the largest peak very well but underestimates all others, for example 2001. The overall r^2 value of the linear regression between measured and calculated data of all 8 years of Myponga reservoir data amounts to 0.39. For Happy Valley reservoir, the overall r^2 value of the linear regression between measured and calculated data of all 8 years amounts to 0.53. Whilst the r^2 values for the both reservoirs are quite good, particularly Happy Valley, the visual inspection and

RMSE shows that the results are not as useful as those obtained from the Chl-*a* rule-based agent (section 7.4.1), even though it had similar r^2 values. Again there is the issue of the impact of algal management strategies affecting the *Anabaena* dynamics in both lakes. Even though the historical data that was used to train the model includes management influenced patterns, to some extent it is an unpredictable factor, whilst it most often occurs in summer, there is no set pattern that can be learnt or forecast. But more relatively, *Anabaena* abundance has a huge range and is extremely dynamic. The species itself can behave in ways not necessarily in keeping with its usual pattern of occurrence in summer (ie if conditions are acceptable at other times).

7.4.2.2 Rule-based *Anabaena* agent, not including Chl-*a* as an input

The IF condition of the rule provides a threshold value determined by WT and DO. The rule indicates that:

IF WT*DO is greater than or equal to 177.013,

THEN the *Anabaena* abundance is calculated by the equation:

$$Anabaena = (((\ WT\ /\ (TURB\ /\ DO))\ /\ (TURB\ /\ (NO3\ *\ WT)))\ /\ (TURB\ *\ (TURB/\ (NO3\ *\ P1))))$$

Otherwise, by the ELSE equation:

$$Anabaena = ((\ P2\ /\ (TURB\ /\ (NO3*\ DO)))\ +\ P3)$$

The rule suggests that DO, NO_3 concentrations as well as WT and TURB, are important criteria to stimulate the growth of *Anabaena*. The causal relationships between these variables and *Anabaena* were further revealed by their sensitivity curves shown in Fig. 52a and b. The THEN branch of the rule examines input sensitivity at *Anabaena* abundances of 2000 cells/mL and below, although the scale of the graph was reduced so that all variables could be better seen. The ELSE branch covers conditions of much lower *Anabaena* populations of below 800 cells/mL. Input sensitivity for the THEN branch of the rule (Fig. 52a) indicates that WT, DO and NO_3 concentrations are positively related to *Anabaena* abundance. Similarly, the input sensitivity for the ELSE branch (Fig.52b) also shows positive relationships between DO and NO_3 concentrations and *Anabaena* growth. Again, in both the THEN and ELSE branches, TURB is shown to have a negative relationship with *Anabaena* abundance. These results support those discovered and discussed in sections 5.4.2 and 7.4.2.1.

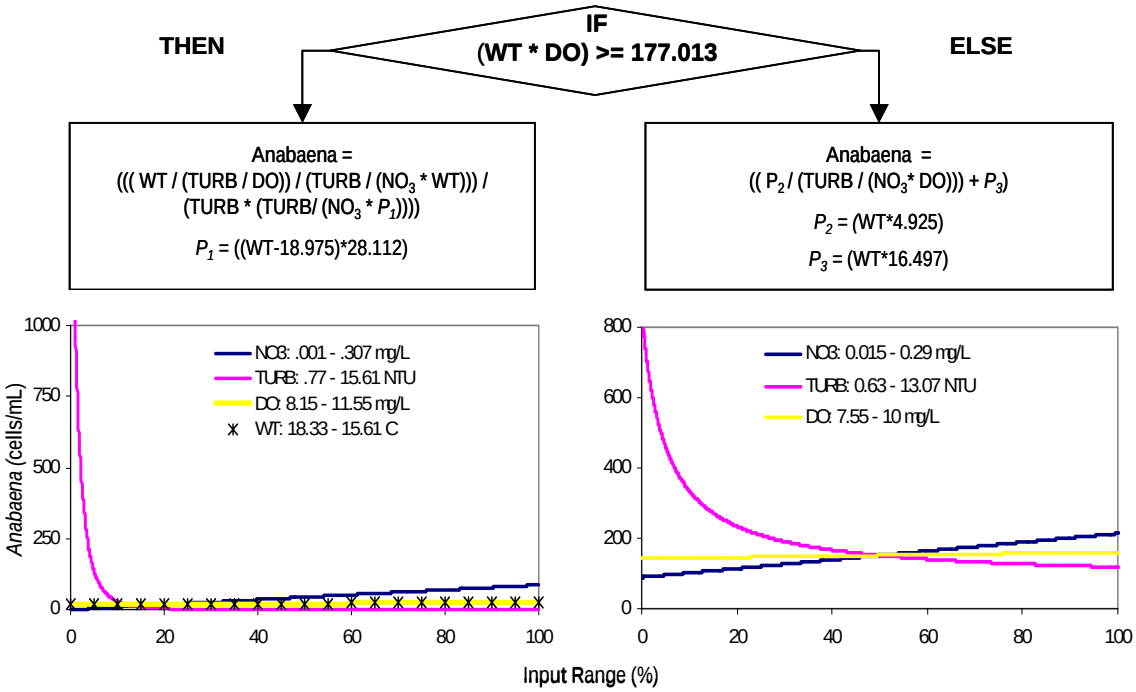


Figure 52. Structure of the rule-based *Anabaena* agent (not inc. Chl-*a*) for Myponga and Happy Valley reservoirs a) input sensitivity of the THEN branch, b) input sensitivity of the ELSE branch

Fig. 53 provides the validation results of the rule-based agent (not including Chl-*a* as an input) for 7 days ahead forecasting of *Anabaena* abundances in Myponga and Happy Valley reservoirs for the years 1996-2003. This rule-based agent that does not include Chl-*a* as an input, largely struggles to predict the correct timing of peak events. Although it does correctly predict the timing of the largest event in each reservoir, 1998-1999 summer bloom in Myponga reservoir and early 1997 peak in Happy Valley. The agent routinely underestimates magnitudes of peak events, however it slightly overestimates the magnitude for the largest peak in the data set (1998-1999 summer bloom in Myponga reservoir). Unlike the agent that included Chl-*a* in the rule set, this agent predicts less false peaks throughout the course of each year. All algal groups contribute to Chl-*a* concentration, and it may be the case that when a different algal genera causes Chl-*a* levels to rise, this results in false peaks in the *Anabaena* forecasts. The overall r^2 value of the linear regression between measured and calculated data of all 8 years amounts to 0.3 for Myponga reservoir and 0.49 for all 8 years of data from Happy Valley reservoir. Again the r^2 values for the both reservoirs are quite good, particularly Happy Valley, but the visual inspection and RMSE shows that the results are not as useful as those obtained by the Chl-*a* rule-based agent.

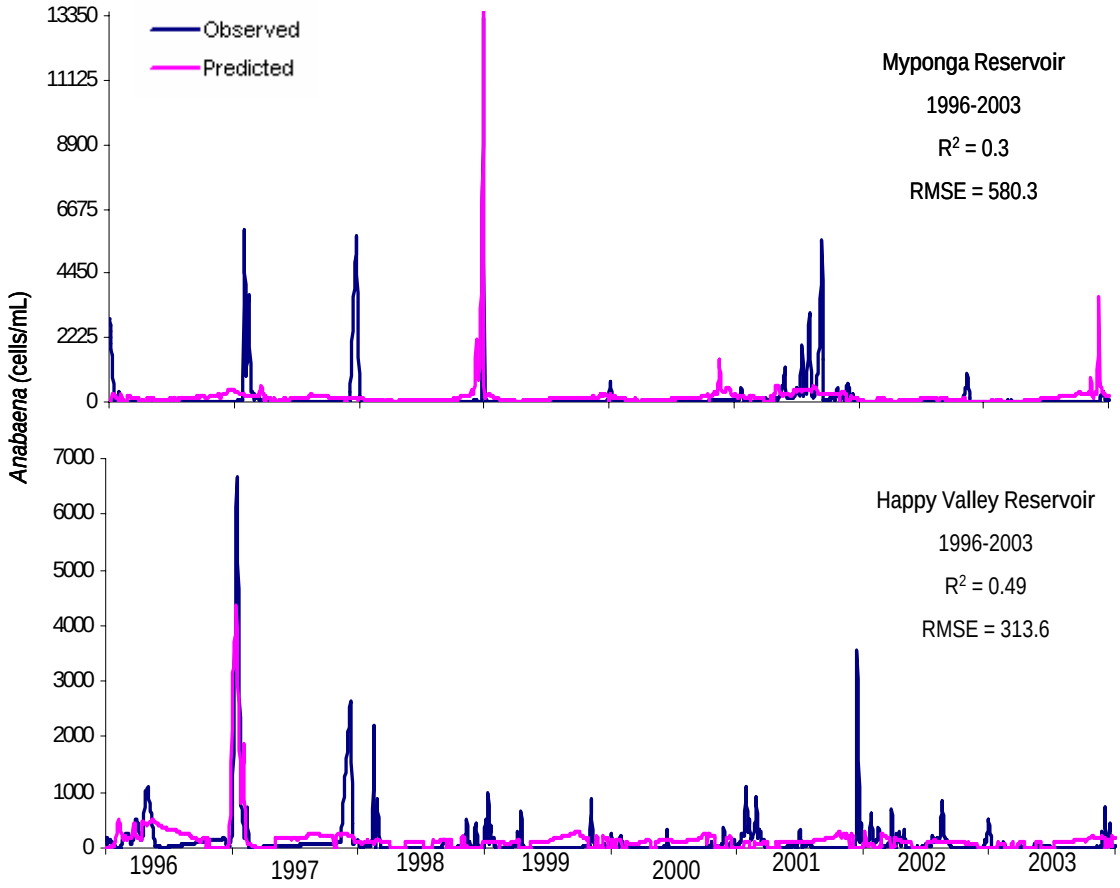


Figure 53. Validation results of the Anabaena rule-based agent (not inc. Chl-a) for Myponga and Happy Valley reservoirs (1996-2003)

7.5 Discussion

In the context of this study, two concepts have been applied for the development of rule based agents of algal populations: 1) rule discovery by means of HEA with *k*-fold cross validation, and 2) rule generalisation by means of merged time-series data of lakes belonging to the same lake category. This chapter aimed to develop a process, using HEA and *k*-fold cross validation, to discover predictive rule-sets in merged time-series water quality data. It aimed to investigate if further optimisation of the agent's parameters could result in predictive agents able to be applied to lakes from the same lake ecosystem category, and whether such agents, trained with interpolated data, could be applied to real-time situations. As a by-product of this investigation, through interpretation of the rule sets and sensitivity analyses, demonstration of causal relationships for algal abundance occurred.

Hypothesis 1 states that the agents developed for Chl-*a* and *Anabaena* forecasting would perform well when tested on historical data from Myponga and Happy Valley reservoirs. Fig. 47 shows

that the Chl-*a* forecasting agent performs well overall when tested on 5 years of data from each reservoir. The *Anabaena* forecasting agents performed to a lesser degree even though the r^2 values were similar if not better than the results from the predictive Chl-*a* agent, the visual analysis shows that the peak events were not as consistently predicted and even when the timing is correct, the magnitude is mostly underestimated. The RMSE values do reflect the difference in accuracy with the Chl-*a* agent having error levels of approximately 6, whilst the *Anabaena* agents gave values of approximately 575 for its application to Myponga reservoir and 315 for its application to Happy Valley. Consequently, Hypothesis 1 can be considered accepted as the agents provide acceptable results, with the Chl-*a* agent being consistently more accurate and useful than the *Anabaena* forecasts. The combination of reservoir management designed to interrupt the natural patterns of *Anabaena* dynamics and manipulate its occurrence, and the fact that algal population dynamics are faster and more difficult to capture than algal community dynamics explain this disparity of forecasting aptitude between the Chl-*a* and *Anabaena* agents.

The agents used only electronically measurable input variables to determine the algal abundance at any particular time so that they could be applied to real-time situations. Whilst the Chl-*a* forecasting agent was able to be tested using real-time data from Hope Valley, the *Anabaena* forecasting agents included NO_3 as an input which, though obtainable via some data loggers, was not available using the current data logger installed in Hope Valley and therefore was not tested on real-time data. The results gained from this 'electronically measurable inputs only' approach were better than expected, considering that nutrients have repeatedly been found to be important in accurate modeling of algal abundance (see Chapter 5). When tested on historical data the Chl-*a* agent, using only electronically measurable inputs, was successful enough to suggest that it could be applied to a real-time situation. To further test the applicability of the agent, the agent was applied to real-time data from the Hope Valley reservoir. Although the forecasting results (shown in section 7.4.1.1) failed to forecast peak Chl-*a* events, it was successful in calculating the concentrations outside of peak events with reasonable accuracy. For these reasons, it is shown that electronically measurable input variables can successfully forecast Chl-*a* concentrations, and the agent can be applied to real-time data, therefore Hypothesis 2 can be accepted.

Hypothesis 3 proposes that the Chl-*a* agent could be applied to another reservoir that belongs to the same lake ecosystem category as Myponga and Happy Valley reservoirs, yet was not used

during the training process. Real-time data from Hope Valley reservoir was used for the purpose of testing this assumption. Results suggested that the ecological conditions within Hope Valley for the 160 days tested were not similar enough to those experienced in Myponga and Happy Valley to enable highly accurate forecasting. Clearly the fact that Chl-*a* values in Hope Valley exceeded those found in Myponga and Happy Valley by a substantial amount meant that the conditions and patterns coinciding with such large peak events were not in the training data, and therefore the model was not capable of learning or accurately forecasting such events. Models are somewhat limited to patterns and events contained within the training data. There are several ways to potentially improve these results. To begin, only a maximum of 5 years from each reservoir was available for the development and training of this agent. Should this amount be increased, more patterns and conditions could be learnt and the likelihood of accurate prediction increases. Also training with more sites would improve the generality. Training with real-time data would also be particularly beneficial. Although the results were not as accurate as hoped, options exist to improve the agents generality and it is felt that the agent has demonstrated the potential to be applied to reservoirs that were not involved in the training process, and therefore Hypothesis 3 is acceptable.

The rule-based agents that were discovered as an outcome of this research have proved to be both predictive and explanatory. It has been demonstrated that the interpretation of the rules can be brought into the context of empirical and causal knowledge on Chl-*a* dynamics as well as population dynamics of *Anabaena* under specific water quality conditions. Sensitivity analyses were again useful in suggesting the nature of relationships between input and output variables and from that management options could be deduced. For example, a negative relationship was consistently shown between turbidity and *Anabaena* abundance, suggesting that turbidity is a limiting factor for *Anabaena* growth. The most probable explanation for this is that the turbid waters reduce the euphotic depth and this implies the potential for light limitation as a management method. The potential for knowledge discovery by this method was also highlighted by the revelation of the relationship between conductivity and Chl-*a* in 7.4.1, indicating that an interesting relationship exists between the variables, which can then be further researched.

The inclusion of Chl-*a* concentration as an input to the *Anabaena* forecasting agent appears to give slightly better results with regard to identification of growth events. Although management considers there to be no correlation between *Anabaena* abundance and Chl-*a* concentration in

Happy Valley reservoir (Daley & Ingleton 2006), results showed that this does not appear to be the case as forecasting was particularly improved in Happy Valley with the addition of Chl-*a* as an input. Overall, visually the results look slightly better than those achieved without Chl-*a* concentration (although it produces more small false peaks) and gives better r^2 values but there is little difference between RMSE. This discrepancy again highlights the need for different measures of fitness and accuracy as some are biased.

The use of *k*-fold cross validation for testing of the rule-based agents reveals their applicability to different conditions, and through this thorough analysis can give confidence in the agents suitability and likely performance in the case of operational implementation. Split-sample testing could not provide the same level of trust in the agents' capabilities as it only shows results from limited applications.

In conclusion, this chapter has demonstrated development of rule-based agents that can forecast algal abundance, be implemented in a real-time situation and be applied to water bodies of similar nature to those the model was trained by. The agents were demonstrated to be both explanatory and predictive with sensitivity analysis describing input-output relationships whilst the *k*-fold cross validation of the agents based on measured data of each year from two similar lakes revealed reasonable forecasting accuracy. The study has shown that using merged limnological data of lakes belonging to the same category is a promising direction for generalising rule-based agents induced by HEA, and the concept as a whole was concluded to be appropriate for real-time forecasting.

In the future, research could focus on the development of agent libraries for specific algal populations and lake categories, applicable for early warning and operational control of algal blooms.

8. CONCLUSION

The manner in which our water resources are managed has never been more important. In view of the deepening water crisis, globally and particularly in Australia, management must consider new approaches to combat water quality problems. The future of water management is certain to include online monitoring as a basic necessity as it offers higher frequency monitoring with reduced manpower. The extensive data that is collected can best be utilised by examining what has occurred in the past, but also what will happen in the future. Accurate forecasting of relevant target variables has the potential to reduce water treatment costs and prevent reservoirs from being taken offline.

This study has addressed the issue of developing computational models for forecasting and understanding of algal blooms in drinking water reservoirs. Working on the assumption that individuals using the models for forecasting and decision-making may not be experienced in computational modelling, the work has focussed on providing simplistic model representations and more generic models with the ability to be applied to multiple locations. Thus reducing the number and complexity of models that are required.

RANN and HEA were both demonstrated as capable of successfully building predictive models that enabled 7-day ahead forecasting of both Chl-*a* concentration and *Anabaena* abundance with good accuracy. The implementation of such models could allow the timely implementation of operational control measures to prevent or reduce the intensity of an algal bloom event.

The research also highlighted that models developed by KANN, RANN and HEA could be useful not just for operational management purposes but also provide information for ecological studies. KANN was shown to allow the examination of many different aspects of water quality within an ecological time-series data set. Both short and long term temporal patterns, such as seasonality and management related trends respectively, have been discovered and examined by KANN. It allowed comparative studies between the two reservoirs, investigation of the habitat preferences of different algal genera with regard to ranges of water temperature and nutrient concentrations, and the novel analysis of water quality conditions during different management regimes.

Sensitivity analyses associated with RANN and HEA provided constructive information regarding the relationships influencing the algal community. This information allowed the refinement of model inputs, the revelation of unknown relationships and patterns in the data and offered insight into manipulating the reservoir environment to promote desired algal populations.

The novel combination of KANN and sensitivity graphs derived from RANN was demonstrated to be capable of providing qualitative and quantitative analysis of the relationship between two variables. The technique could be suggested as an initial step in the analysis of large ecological databases, enabling rapid visualisation of non-linear relationships, highlighting patterns and areas of interest for further research.

The use of the three modelling approaches KANN, RANN and HEA throughout the research gave an arrangement that provided explanation and forecasting of algal dynamics. This approach should be considered for use even when the main goal is forecasting, as improved understanding of the system and the relationships that influence the output variable can ultimately only improve the management. An integrated approach such as this is capable of providing thorough data mining and analysis, and is necessary for the comprehensive investigation of complex and non-linear ecological processes within freshwaters.

The biggest advance from this study toward the goal of improved water management by real-time forecasting was the development of a rule-based Chl-*a* forecasting agent constructed using HEA and *k*-fold cross validation. The preliminary experiments showed promising results with regard to the models application to real-time data and its relevance to an independent water body. It can be concluded that the concept of rule-based agents will facilitate real-time forecasting of algal blooms in drinking water reservoirs provided on-line monitoring of relevant variables is implemented. The accurate prediction of algal dynamics as a result of real-time forecasting is key to successful management of nuisance algae species and will provide useful early warning for managers to consider and implement the most suitable management or mitigation option. Further to alerting water resource managers to an impending bloom event, there is the potential for this technology to be linked to operational controls. For example, when the branch of the predictive rule set that forecasts the higher levels of growth is activated – an alert is issued. This would mean that at any time when the threshold criteria (the IF branch) for higher growth is met, a warning of the potential for problems is released. In section 2.5.5, it was suggested that intermittent mixing linked to a forecasting agent should be trialled. This would go beyond simply alerting potential problems and could trigger the implementation of intermittent mixing when a pre-determined forecast level is predicted. This would provide discourage the establishment of large algal populations by having alternating stability conditions and would reduce energy costs dramatically. These are just some of the possibilities for the use of this new approach in the operational management of reservoirs.

8.1 Summary of findings

- KANN can reveal important patterns and relationships between physical and chemical water quality factors and algal dynamics by classification and clustering of complex temporal patterns, and allows for comparisons between reservoirs.
- KANN is a suitable tool for the analysis of water conditions corresponding with the presence of specific algal genera. Results showed the occurrence of the dominant blue-green, green algae and diatoms in relation to ranges of water temperature, PO_4 or NO_3 concentrations, further explaining seasonality and succession.
- KANN demonstrated that several periods of investigation could be examined and compared at once. In this study water quality patterns were linked to the different management regimes in place during four periods.
- Both RANN and HEA can be used to forecast algal dynamics at different resolutions from community (Chl-*a* concentration) to genera level (*Anabaena* abundance), 7-days in advance. These methods could provide both reservoir specific lake category generic model forecasting options.
- Both RANN and HEA could also provide sensitivity analyses to quantitatively describe relationships underlying algal dynamics. Two types of analyses, most influencing parameter (MIP) and sensitivity on wide ranged-disturbance (SWD) were found to be capable of comprehensively scrutinising the relationships between algal dynamics and water quality parameters.
- The combination of KANN and RANN was able to offer both quantitative and qualitative analysis and explanation of patterns in the data, and provided very useful visual representations of these patterns.
- HEA induced rule-sets extracted from merged data sets can be developed to use only electronically measurable inputs (available from online monitoring and therefore able to be linked to real-time forecasting) in order to produce algal population forecasts 7 days in advance, particularly for Chl-*a* concentrations. *K*-fold cross validation allowed the thorough assessment of the models usefulness under different conditions and results can be considered more wholly representative of the models capabilities compared with split-sample testing.
- Application of the Chl-*a* forecasting agent to all years of data from both Myponga and Happy Valley reservoirs showed that the agent was consistently successful at forecasting

the major trends in both reservoirs. If real-time data were available from these sites, it can be assumed that the agent would be equally useful in a real-time forecasting situation.

- Application of the Chl-*a* forecasting agent to real-time data from an independent reservoir, from which data was not used during the rule extraction process, demonstrated that the agent was able to be applied successfully to real-time data and forecast reasonable Chl-*a* values. Results demonstrated the potential for agents to be applied to other water bodies from within the same lake ecosystem category, with further work needed.
- Compared to RANN, HEA proved to be a more informative approach to forecasting and understanding of algal dynamics. Whilst the forecasting results were equal or superior to RANN; HEA also offers more information through interpretation of the rule-sets than neural networks can provide. This transparent model representation increases the likelihood of model acceptance from potential users.
- This study has used different modelling techniques and different validation techniques. Split-sample validation allowed comparison between two forecasting methods, but was not sufficient to provide a model that could reliably be used operationally. Ultimately it was found that *k*-fold cross validation and merged data from same category lakes was necessary for the development of operational agents generic for a certain lake ecosystem category and to thoroughly demonstrate their application to a wide-range of conditions.

8.2 Contributions of the study

The major contributions of this research include: (1) to offer insight into the capabilities of 3 kinds of computational modelling techniques applied to complex water quality data, (2) novel applications of KANN including the division of data into separate management periods for comparison of management efficiency, (3) to both qualitatively and quantitatively elucidate relationships between water quality parameters, (4) research toward the development of a forecasting tool for algal abundance 7 days in advance that could be generic for a particular lake ecosystem category and implemented in real-time, and (5) to suggest a thorough testing method for such models (*k*-fold cross validation).

Ultimately, the questions of most importance which were addressed by this research project were:

Is it possible to train a model to forecast algal dynamics using only electronically measurable data, so it could be compatible with a real-time monitoring/forecasting situation?

Nutrients have traditionally been considered as essential inputs for accurate modelling of phytoplankton dynamics. This was supported by sensitivity analyses results discussed in Chapter 5, which demonstrated that PO_4 and NO_3 concentrations were consistently the most influencing parameters when forecasting Chl-*a* concentration and *Anabaena* abundance. However, at this stage, many data loggers and online monitoring systems cannot measure nutrients in real-time, therefore they cannot be included in a model to be used for real-time forecasting.

In sections 5.4.1.6 and again in 7.4.1, it was shown that models trained with only electronically measurable inputs of water temperature, DO, turbidity and conductivity were able to successfully forecast Chl-*a* concentrations 7 days in advance. This produced models that are compatible with a real-time monitoring and forecasting situation, thus answering this research question.

Is it possible to develop forecasting rule-sets that can be generic for particular lake ecosystem categories?

Generic models have numerous benefits including reduced labour for model development and maintenance as well as increased comparability between and applicability to different sites. This study investigated whether it was possible to develop a single forecasting agent that could be applied to a cross-section of water bodies, specifically for a particular lake ecosystem category: warm monomictic and eutrophic. Merged data from both Myponga and Happy Valley reservoirs was used to discover and extract a forecasting rule using HEA. The numerical parameters in the rule were substituted with functions of water temperature to make the rule more adaptable and applicable to different data sets. Results of the application of the rule set to all years of data from Myponga and Happy Valley showed that it consistently provided useful forecasts in these locations. Results of the application of the agent to another water body from the same category, Hope Valley reservoir, showed that the notion of a generic forecasting agent for specific lake

ecosystem categories was a promising concept and that with some improvements a very successful outcome would be achievable.

Can this forecasting tool be implemented in a real-time environment to provide algal population forecasts up to a week in advance?

Whilst there is merit in the development and testing of models with historical data for research into model applications and underlying relationships driving algal dynamics, ultimately forecasting models are designed for practical use in a real-time environment, not a simulated one, and therefore must be tested in a real-time situation.

The Chl-*a* forecasting agent that was developed in Chapter 7 was applied to real-time data from Hope Valley reservoir. Ideally, the agent would have been tested on recent real-time data from Myponga or Happy Valley reservoirs, as well as real-time data from an independent reservoir, however neither Myponga nor Happy Valley reservoirs have real-time monitoring of any variables in the rule-set except water temperature. Results showed that the agent was able to successfully be applied to real-time data and was able to forecast realistic Chl-*a* concentrations by means of electronically measurable inputs only. Although the agent was unsuccessful in identifying the major algal bloom event, this is due to the application of the agent to data from an independent reservoir, and is not related to the real-time data. Therefore it can be concluded that forecasting agents developed in this manner are complimentary to online, real-time monitoring to achieve real-time forecasting.

8.3 Recommendations

Throughout the project, ideas emerged regarding how to improve or further the research that were not possible to carry out due to time or material constraints. Below are the major points of recommendation:

- Whilst RANN provided useful forecasts and information regarding algal dynamics, predictive rules extracted using HEA can match if not better these results as well as providing a simple model representation greater ease of use and understanding. For these reasons, it can be suggested that future work of a similar nature to this study need only use HEA for forecasting and sensitivity by wide ranged disturbance.

- With regard to the forecasting of algal dynamics, the inclusion of a measure of water stability as an input variable could greatly improve the forecasting of species that thrive in stable water columns, such as *Anabaena* during times of thermal stratification. It would indicate when artificial mixing was successfully achieving mixed conditions and thereby reducing the risk of *Anabaena* blooms, even if all other conditions are conducive for population growth. As it stands, the models forecast *Anabaena* blooms when nutrient and water temperature conditions suggest it would occur. Well-mixed conditions may actually prevent the impending bloom but as water column stability is not reflected in the data the model may forecast a false peak.
- The study sites involved in this project employ artificial destratification to deter harmful algal blooms over summer, but rely heavily on $CuSO_4$ dosing when this fails. In particular, analysis of Myponga reservoir water quality over time showed that, whilst management has successfully reduced some water quality issues, some variables are at higher levels in the latest period than prior to the implementation of mixing. This suggests that perhaps a new method of implementation could be trialled, as discussed in Chapter 2, to reduce the necessity for $CuSO_4$ dosing. This could include a forecasting model linked to intermittent mixing, which is triggered when significant algal population growth is forecast.
- There is much potential for improvement in both accuracy and applicability of rule-based forecasting agents. The method has clearly been shown to be appropriate for the task of real-time forecasting of algal dynamics in freshwaters, however further research is needed to ensure the model can be widely applied to other water bodies within a lake ecosystem category. The inclusion of more reservoir data sets from which the initial rule-set is extracted would be helpful in providing the model with a greater range of conditions and occurrences for it to consider and potentially incorporate into the rule. The inclusion of high frequency real-time data is also likely to improve the accuracy of the application to real-time forecasting.

8.4 The Future

Rule-based forecasting agents provide simple model representation in the form of an IF-THEN-ELSE rule which does not require costly software to run and can be simply be applied in a spreadsheet. The explicit nature of the rule-set means that it is reasonably understandable and

easy to apply, and the user does not have to be familiar with computational modelling, which is what makes it a viable tool for water managers in many situations. Rule-based forecasting agents for certain algal species could, of course, be developed for other lake ecosystem categories. An agent could be set up for real-time forecasting of specific algal dynamics in just one reservoir, using incoming real-time data. Or on a larger scale, a library of forecasting agents (as suggested by Recknagel 2003b) for numerous lake ecosystem categories could be provided online in an integrated intelligent data warehouse and modelling platform, which would enable the user to select the appropriate agent for the lake ecosystem category to be applied to incoming online water quality data, thus providing real-time forecasting. And, as earlier discussed, there is the potential for the agents themselves to be imbedded in the operational management of reservoirs.

Although there is further effort needed before rule-based forecasting agents for lake ecosystem categories are a practical option for water managers and decision makers, this research has been the necessary first step towards providing a generic model for warm monomictic and eutrophic water bodies that could be implemented for real-time forecasting and a contribution to the development of agent libraries for certain species and categories available for shared use.

APPENDIX A

Trophic state classifications

Three lake classification methods were used to identify the trophic state of Myponga and Happy Valley reservoirs.

Table 23. Carlson's trophic state index (TSI) (according to Carlson (1977))

Trophic State	Trophic state index (TSI)
Ultra-oligotrophic	≤ 20
Oligotrophic	≤ 40
Mesotrophic	≤ 50
Eutrophic	≤ 70
Hypertrophic	≥ 70

Table 24. OECD lake classification standard (according to Vellenweider and Kerekes (1982))

Trophic State	Mean total phosphorus	Mean Chl-*a*	Max. Chl-*a*	Mean secchi depth	Max. secchi depth
Ultra-oligotrophic	<4	<1	<2.5	>12	>6
Oligotrophic	<10	<2.5	<8	>6	>3
Mesotrophic	10-35	2.5-8	8-25	6-3	3-1.5
Eutrophic	35-100	8-25	25-75	3-1.5	1.5-0.7
Hypertrophic	>100	>25	>75	<1.5	<0.7

Table 25. German lake classification standard (according to Ryding and Rast (1989))

Criterion	Quality Class					
	1	2	3a	3b	4	5
PO_4 (mg/L)	0-0.002	0-0.005	0-0.1		>0.1	>0.5
Dissolved Inorganic Nitrogen (mg/L)	≤ 0.01	≤ 0.03	≤ 0.1		>0.1	>0.5
Chl-*a* (summer mean)	≤ 3	<10	10-20	20-40	40-60	>60
Secchi depth (m)	≥ 6	≥ 4	≥ 1		≥ 0.5	>0.5
pH	6.5-8	7-8.5	7-9	7-9.5	6.5-10	6-11
Trophic state	Oligo-trophic	Meso-trophic	Eutrophic stratified	Eutrophic unstratified	Poly-trophic	Hyper-trophic

Table 26. Observed water quality data used for reservoir trophic state classification

Variable	Myponga reservoir	Happy Valley reservoir
Chl-*a* (ug/L): mean	7.84	8.87
summer	9.14	14.79
$PO_{4\ (mg/L)}$	0.022	0.024
Total phosphorous	0.06	0.098
NO_3	0.11	0.226
Secchi depth	N/A	N/A
pH	N/A	N/A

Table 27. Reservoir trophic state classifications

Trophic state classification	Myponga reservoir	Happy Valley reservoir
Carlson's Trophic State Index (TSI)		
Secchi depth	N/A	N/A
Total phosphorus	Eutrophic	Eutrophic
Chl-a	Eutrophic	Eutrophic
OECD lake classification standard		
Total phosphorus	Eutrophic	Eutrophic
Chl-*a*	Mesotrophic	Eutrophic
Max Chl-*a*	Eutrophic	Eutrophic
Secchi depth	N/A	N/A
Min Secchi depth	N/A	N/A
German lake classification standard		
PO_4	Eutrophic	Eutrophic
DIN*	Polytrophic	Polytrophic
Summer Chl-*a*	Mesotrophic	Eutrophic, stratified
Secchi depth	N/A	N/A
pH	N/A	N/A

All variables are mean values unless otherwise stated.
* Classified according to NO_3 fraction.

Tab. 27 shows that in most cases both reservoirs were classified as eutrophic, although Happy Valley reservoir was indicated to be more eutrophic than Myponga reservoir.

APPENDIX B

The relationship between rainfall and turbidity

In section 4.4.1.1, a link between rainfall and turbidity is suggested. The graph below provides a snap shot, based on data availability, of the relationship between rainfall (and inflow) and turbidity levels in Myponga reservoir. Only limited data from Myponga reservoir was available for exploration of this concept and no rainfall data was available for Happy Valley reservoir, but it is assumed that the relationship between turbidity and rainfall demonstrated in the Myponga data would be relevant in Happy Valley reservoir also.

The graph (Fig. 54) shows that there appears to be a link between heavy rainfall and turbidity. It can be seen that turbidity experiences its annual peak during the largest rainfall event of the year, giving both variables maximum values during winter. Small rainfalls do not seem to impact turbidity levels as significant rainfall events do.

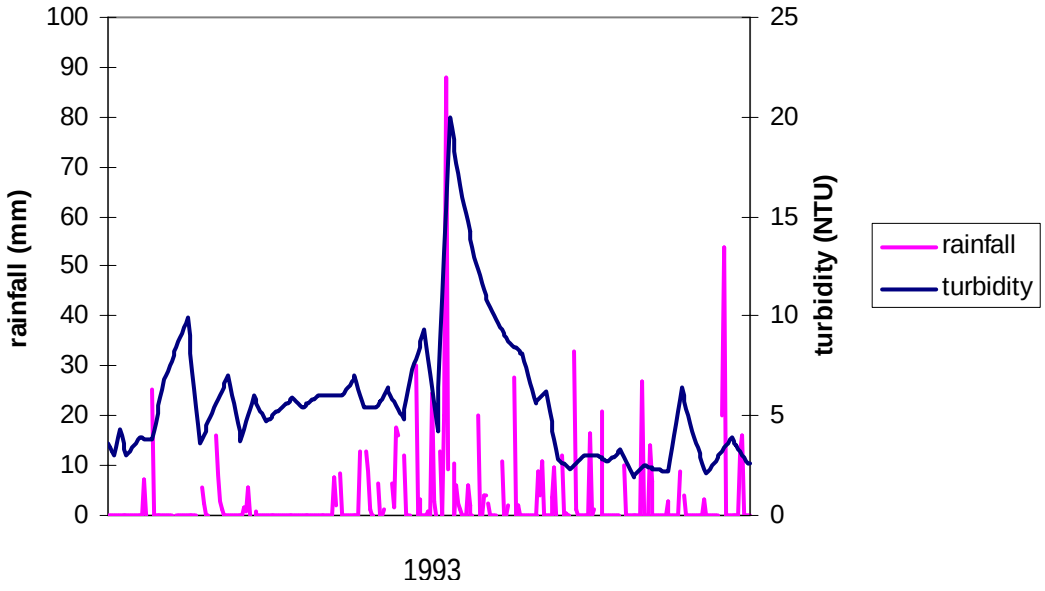


Figure 54. Rainfall and turbidity levels in Myponga reservoir 1993

APPENDIX C

The relationship between colour and dissolved organic carbon (DOC)

In Chapter 4 (see section 4.4.1.1), there is discussion regarding the difference in colour between the study sites, Myponga and Happy Valley reservoirs. It is suggested that DOC is linked to colour and that Myponga probably has higher levels of DOC in the water. These graphs provide a snap shot, based on data availability, of the relationship between colour and DOC in the reservoirs. The graphs below confirm that Myponga has higher levels of DOC (up to 14mg/L) compared with Happy Valley reservoir (up to 10mg/L). In the case of Myponga reservoir, the influence of DOC on colour can clearly be seen. Fig. 55 also illustrates the seasonality of the variables, with the peak values in each year occurring in early-mid September. The DOC build up from the winter would contribute to this early spring peak. In Happy Valley reservoir there does not appear to be such an obvious link between DOC and colour and there is no clear seasonality for the DOC (Fig. 56). The difference between clear seasonality patterns in Myponga reservoir and indistinct patterns in Happy Valley reservoir may be explained by the difference in origin and timing of inflows to the reservoirs. Myponga reservoir is entirely catchment fed and as such experiences most inflow in winter and little in other seasons. Happy Valley reservoir receives water from the river Murray throughout the year when reservoir levels are low and demand is high, which interrupts the seasonal patterns you might expect to find.

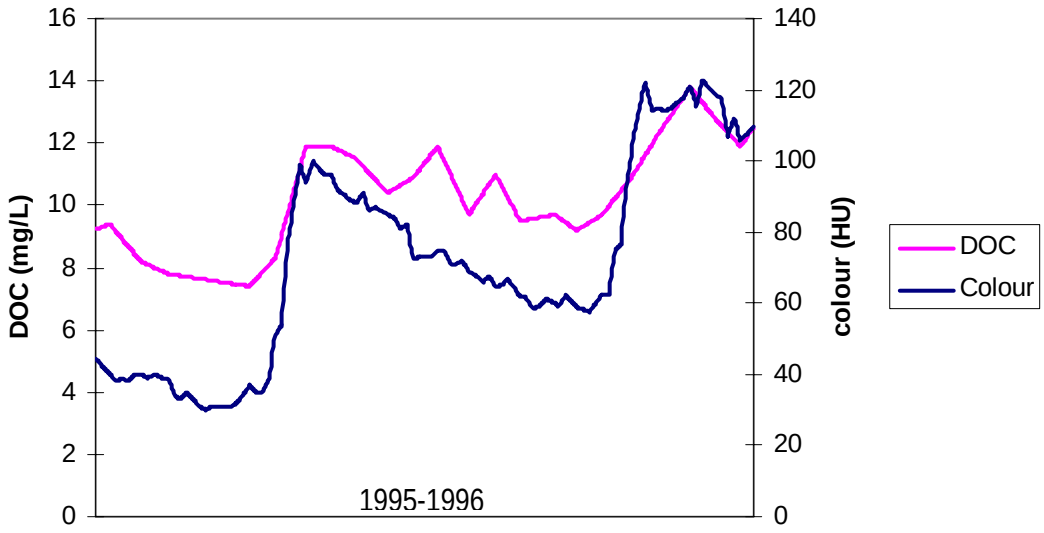


Figure 55. The relationship between colour and DOC in Myponga reservoir

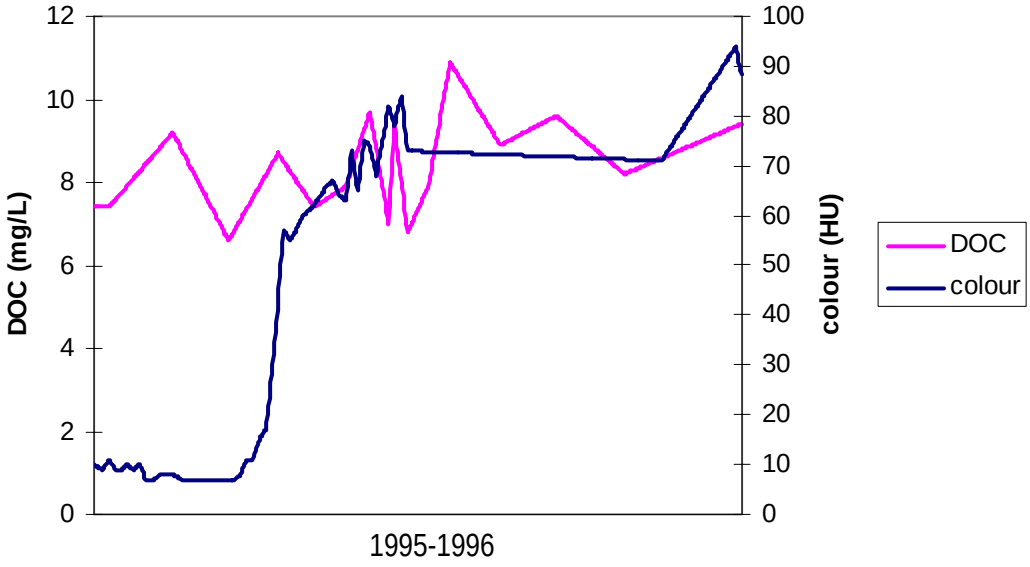


Figure 56. The relationship between colour and DOC in Happy Valley reservoir

APPENDIX D

Chl-a as an input to HEA

Whether to include previous Chl-*a* concentrations as an input to predict present or future Chl-*a* concentrations is a contentious issue, as discussed in Chapter 5. Experiments were carried out to test HEA forecasting rule extraction if past Chl-*a* values were included as input to forecast future Chl-*a* levels. It was found that the algorithm largely ignored other input values and relied to heavily on the past Chl-*a* concentrations (see rule below for an example). Fig. 57 shows that whilst the result is excellent with regard to magnitude, the timing is always approximately one week out as the forecasting rule, at any given time, is simply forecasting the Chl-*a* values from one week prior, so the forecasts are always one week behind. It was decided, for this reason among others, that past Chl-*a* concentrations would not be used as input to predict current or future Chl-*a* concentrations in this project.

IF (Chl-*a*7day lag<=35.244)
THEN Observed Chl-*a*=((Chla7daylag*Chla7daylag)/Chla7daylag)
ELSE
Observed Chl-*a*=(DO+23.154)

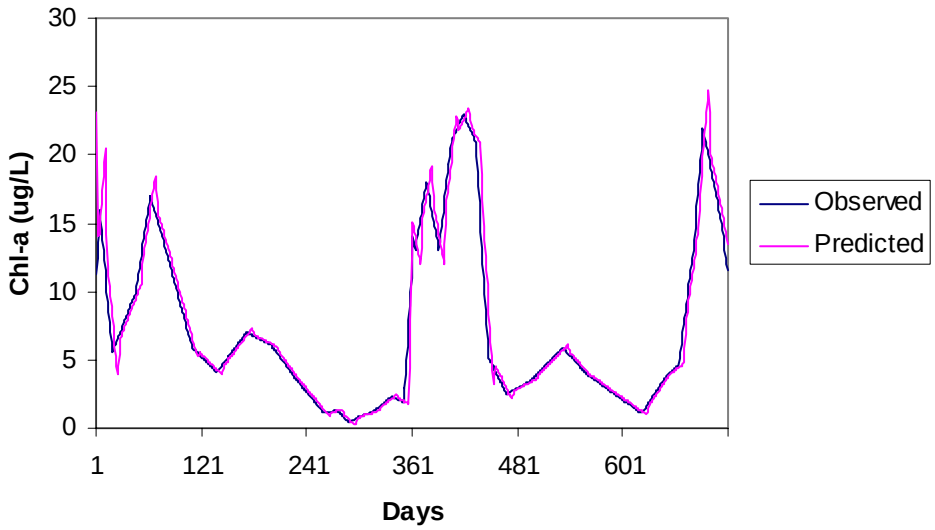


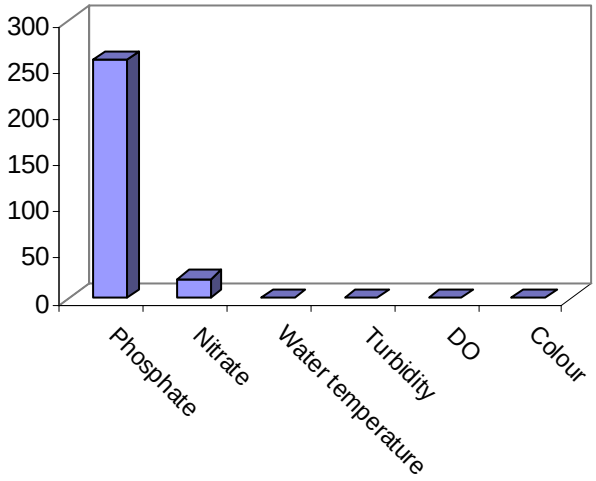
Figure 57. Forecasting results from HEA rule using past Chl-a values as input to predict current Chl-a levels.

APPENDIX E

Most Influencing Parameter graphs

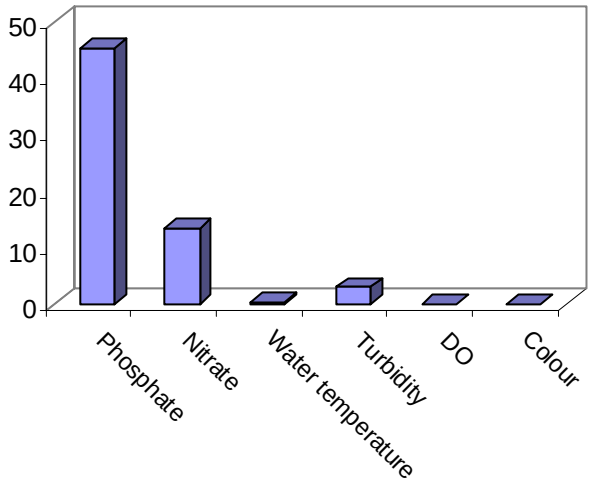
The column graphs below are the 'most influencing parameter' (MIP) sensitivity analyses produced in association with RANN models from Chapter 5.

Chl-*a* forecasting in Myponga reservoir validated for 2001



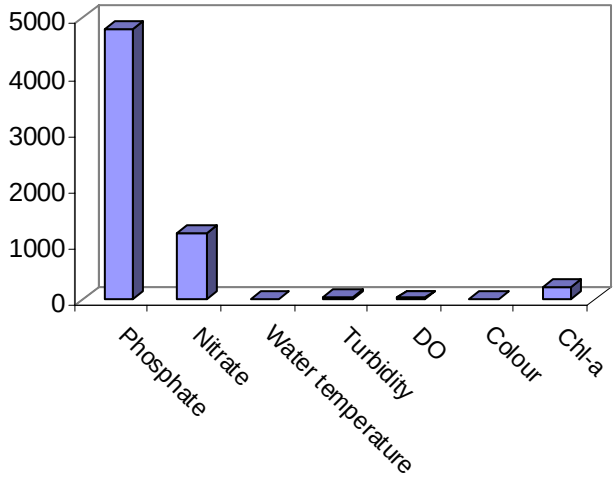
Sensitivity	***Chla***
Phosphate	258.0392
Nitrate	21.12864
Water temperature	0.151754
Turbidity	0.184198
DO	0.149959
Colour	0.030213

Chl-*a* forecasting in Happy Valley reservoir validated for 1999



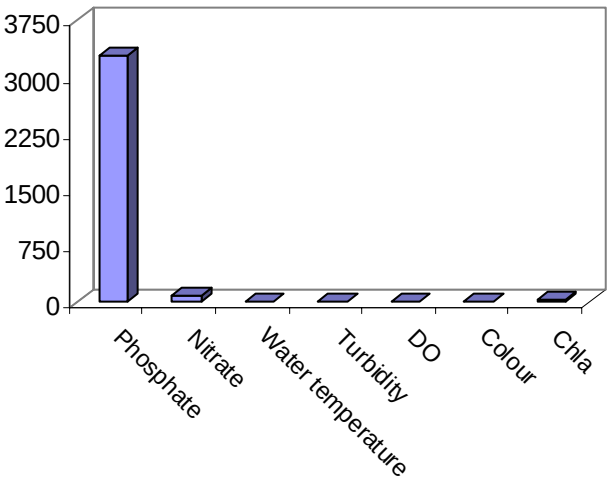
Sensitivity	***Chl-a***
Phosphate	45.37164
Nitrate	13.48344
Water temperature	0.354704
Turbidity	2.948941
DO	0.035365
Colour	0.000522

Anabaena forecasting in Myponga reservoir validated for 2002



Sensitivity	*Anabaena*
Phosphate	4822.9766
Nitrate	1171.9799
Water temperature	18.487995
Turbidity	48.652557
DO	25.487
Colour	3.6750431
Chl-a	214.987

Anabaena forecasting in Happy Valley reservoir validated for 1997



Sensitivity	*Anabaena*
Phosphate	3286.0769
Nitrate	79.275391
Water temperature	9.7696857
Turbidity	3.5685189
DO	3.7291214
Colour	2.35887
Chla	28.143349

These results show that in all cases phosphate was found to be the most influencing parameter followed distantly by nitrate. In all cases colour was found to be the least influencing parameter.

REFERENCES

Back T., Hammel U. & Schwefel H.-P. (1997) Evolutionary computation: comments on the history and current state. *IEEE Transactions on Evolutionary Computation* 1: 5-16.

Ball G. R., Palmer-Brown D. & Mills G. E. (2000) A Comparison of Artificial Neuronal Network and Conventional Statistical Techniques for Analysing Environmental Data. In: *Artificial Neuronal Networks: Applications to Ecology and Evolution* (eds. S. Lek & J.-F. Guegan) pp. 165-182. Springer-Verlag, Berlin.

Barker J. L. (1976) Effects of air injection at Prompton Lake Wayne Co, Penn. *J. Res. USGS* 4: 19-25.

Bernhardt H. (1967) Aeration of Wahnbach Reservoir without changing the temperature profile. *J. Am. Water Wastewater Assoc.* 59: 266-275.

Bobbin J. (2002) Self-Adaptive Evolution of Model Structures and Parameters. The University of Adelaide, Adelaide.

Bobbin J. & Recknagel F. (1999) Mining Water Quality Time Series for Predictive Rules of Algal Blooms by Genetic Algorithms. In: *MODSIM '99 - Modelling the Dynamics of Natural, Agricultural, Tourism and Socio-economic Systems* (eds. L. Oxley, F. Scrimgeour & A. Jakeman) pp. 679-684, Hamilton, New Zealand.

Bobbin J. & Recknagel F. (2003) Predictive Rules for Phytoplankton Dynamics in Freshwater Lakes Discovered by Evolutionary Algorithms. In: *Ecological Informatics: Understanding Ecology by Biologically-Inspired Computation* (ed. F. Recknagel) pp. 291-311. Springer-Verlag, Berlin.

Boddy L. & Morris C. W. (1999) Artificial neural networks for pattern recognition. In: *Machine Learning Methods for Ecological Applications* (ed. F. A) pp. 37-87. Kluwer Academic Publishers, Boston, Dordrecht, London.

Boney A. D. (1989) *Phytoplankton*. Edward Arnold, London.

Boulton A. J. & Brock M. A. (1999) *Australian Freshwater Ecology: processes and management*. Gleneagles Publishing, Mt Osmond, South Australia.

Bowden G. J. (2003) Forecasting Water Resources Variables Using Artificial Neural Networks. In: *School of Civil and Environmental Engineering* pp. 566. University of Adelaide, Adelaide.

Bowden G. J., Dandy G. & Maier H. (2006) An Evaluation of Methods for the Selection of Inputs for an Artificial Neural Network Based River Model. In: *Ecological Informatics* (ed. F. Recknagel) pp. 275-292. Springer-Verlag, Berlin.

Brey T., Jarre-Teichmann A. & Borlich O. (1996) Artificial neural network versus multiple linear regression: predicting P/B ratios from empirical data. *Marine ecology Press Series*: 251-256.

Brookes J. D. & Antenucci J. (2006) Artificial Destratification for Control of Cyanobacteria. In: *Cyanobacteria: Management and Implications for Water Quality*. CRC for Water Quality and Treatment, Adelaide.

Brookes J. D., Baker P. D. & Burch M. D. (2002a) Ecology and Management of Cyanobacteria in Rivers and Reservoirs pp. 33-42. CRC for Water Quality and Treatment, Salisbury, SA.

Brookes J. D. & Burch M. (2006) The Ecology of Cyanobacteria. In: *Cyanobacteria: Management and Implications for Water Quality*. CRC for Water Quality and Treatment, Adelaide.

Brookes J. D., Burch M. & Tarrant P. (2000) Artificial Destratification: Evidence for Improved Water Quality. *Water* 27: 18-22.

Brookes J. D., Lewis D. M., Linden L. G. & Burch M. (2002b) On-line Monitoring of Reservoirs for Risk Management. *Water* 29: 20-27.

Brosse S., Giraudel J. L. & Lek S. (2001) Utilisation of non-supervised neural networks and principal component analysis to study fish assemblages. *Ecological Modelling* 146: 159-166.

Burch M. (1987) Limnology of Happy Valley Reservoir pp. 43. Engineering and Water Supply Department, Adelaide.

Burch M. (2005a) Project Meeting, Adelaide.

Burch M. (2005b) Supervisor meeting, Adelaide.

Burch M., Chow C. W. K. & Hobson P. (2002) Algicides for control of toxic cyanobacteria pp. 23-32. CRC for Water Quality and Treatment, Salisbury, SA.

Burke L. I. (1991) Introduction to artificial neural systems for pattern recognition. *Computers and Operations Research* 18: 211-220.

Cao H., Recknagel F., Joo G.-J. & Kim D.-K. (2004) Rule-Set Discovery for Prediction and Explanation of Chlorophyll a Dynamics in the Nakdong River (Korea) by Using a Hybrid Evolutionary Algorithm. In: *4th Conference of the ISEI*. In Press, Pusan, South Korea.

Cao H., Recknagel F., Joo G.-J. & Kim D.-K. (2006a) Discovery of predictive rule sets for chlorophyll-*a* dynamics in the Nakdong River (Korea) by means of the hybrid evolutionary algorithm HEA. *Ecological Informatics* 1: 43-53.

Cao H., Recknagel F., Kim B. & Takamura N. (2005) Hybrid Evolutionary Algorithm for Rule Set Discovery in Time -Series Data to Forecast and Explain Algal Population Dynamics in Two Lakes Different in Morphometry and Eutrophication. In: *Ecological Informatics* (ed. F. Recknagel) pp. 347-367. Springer-Verlag, Berlin.

Cao H., Recknagel F., Kim B. & Takamura N. (2006) Hybrid Evolutionary Algorithm for Rule Set Discovery in Time -Series Data to Forecast and Explain Algal Population Dynamics in Two Lakes Different in Morphometry and Eutrophication. In: *Ecological Informatics* (ed. F. Recknagel) pp. 347-367. Springer-Verlag, Berlin.

Cao H., Recknagel F., Kim B. & Takamura N. (2006b) Hybrid Evolutionary Algorithm for Rule Set Discovery in Time -Series Data to Forecast and Explain Algal Population Dynamics in Two Lakes Different in Morphometry and Eutrophication. In: *Ecological Informatics* (ed. F. Recknagel) pp. 347-367. Springer-Verlag, Berlin.

Capblanq J. & Catalan J. (1994) Phytoplankton: which and how much? In: *Limnology Now: A Paradigm of Planetary Problems* (ed. R. Margalef) pp. 9-36. Elsevier Science.

Chakraborty K., Mehrotra K., Mohan C. K. & Ranka S. (1992) Forecasting the Behaviour of Multivariate Time Series Using Neural Networks. *Neural Networks* 5: 961-970.

Chan W.-S., Recknagel F., Cao H. & Park H.-D. (2007) Elucidation and short-term forecasting of microcystin concentrations in Lake Suwa (Japan) by means of artificial neural networks and evolutionary algorithms. *Water Research* 41: 2247.

Chon T.-S., Park Y. S., Kwak I.-S. & Cha E. Y. (2006) Non-linear Approach to Grouping, Dynamics and Organizational Informatics of Benthic Macroinvertebrate Communities in Streams by Artificial Neural Networks. In: *Ecological Informatics* (ed. F. Recknagel). Springer-Verlag, Berlin.

Chon T.-S., Park Y. S., Moon K. H. & Cha E. Y. (1996) Patternising communities by using artificial neural networks. *Ecological Modelling* 90: 69-78.

Chon T.-S., Park Y. S. & Park J.-H. (2000) Determining temporal patterns of community dynamics by using unsupervised learning algorithms. *Ecological Modelling* 132: 151-166.

Coad P., Cathers B. & Van Senden D. (2005) Predicting Estuarine Algal Blooms Utilising Neural Network Modelling - A Preliminary Investigation. In: *MODSIM05*, Melbourne.

CRCWQT (2003) Annual Review 2002 / 2003. CRC for Water Quality and Treatment, Adelaide.

CRCWQT (2005) Drought and Water Quality pp. 20. CRC for Water Quality and Treatment, Brisbane.

Creagh C. (1992) What can be done about toxic algal blooms? *Ecos* 72: 14-19.

Daley R. & Ingleton G. (2006) Data warehousing meeting, Adelaide.

Drury D. D., Porcella D. B. & Gearheart R. A. (1975) The effects of artificial destratification on the water quality and microbial populations in the Hyrum Reservoir. Utah Water Resources Lab., Utah State University, Logan, Utah.

Ewing T., Romero J. R., Imberger J., Antenucci J. & Deen A. (2004) A real-time reservoir decision support system. In: *6th International Conference on Hydroinformatics* (ed. P. B. Liong). World Scientific Publishing Company, Singapore.

Ferguson A. D. (1997) The role of modelling in the control of toxic blue-green algae. *Hydrobiologia* 349: 1-4.

Fielding A. (1999a) How should accuracy be measured? In: *Machine Learning Methods for Ecological Applications* (ed. F. A) pp. 209-223. Kluwer Academic Publishers, Boston, Dordrecht, London.

Fielding A. (1999b) An introduction to machine learning methods. In: *Machine Learning Methods for Ecological Applications* (ed. F. A) pp. 1-36. Kluwer Academic Publishers, Boston, Dordrecht, London.

Flood I. & Kartam N. (1997) Systems. In: *Artificial Neural Networks for Civil Engineers: Fundamentals and Applications* (eds. N. Kartam, I. Flood & J. H. G. Jr.) pp. 19-43. American Society of Civil Engineering, New York.

Fogel D. (1998) *Evolutionary Computation: The Fossil Record*. IEEE Press, Piscataway, New Jersey.

Fogel D. (2000) *Evolutionary Computation: Toward a New Philosophy of Machine Intelligence*. IEEE Press, Piscataway, New Jersey.

Fogelman S. (2004) Development of a Rapid and Non-destructive Drinking Water Monitoring System. In: *Fourth Postgraduate Students Conference: CRC for Water Quality and Treatment* pp. 111-116, Noosa Lakes, Queensland.

Fogelman S. (2006) Development of the Universal Calibration System for the On-line Analysis of Drinking Water Quality. In: *Fifth Postgraduate Student Conference of the CRC for Water Quality and Treatment*, Melbourne.

Foody G. M. (1999) Applications of the self-organizing feature map neural-network in community data-analysis. *Ecological Modelling* 2-3: 97-107.

Forsberg C. & Ryding S. O. (1980) Eutrophication parameters and trophic stateindices in 30 Swedish waste-receiving lakes. *Arch. Hydrobiologia* 89: 189-207.

Freeman K. (2000) PSYCHIC NETWORKS: Training Computers to Predict Algal Blooms. *Environmental Health Perspectives* 108: 464-467.

French M. (1996) Environmental modelling using artificial neural networks: Implications for real-time forecasting systems. In: *Distinguished Lecturer Series*. University of Adelaide, Adelaide.

Gibbs M. (2004) The Application of Evolutionary Algorithms to Water Distribution Systems. In: *Fourth Postgraduate Student Conference: CRC for Water Quality and Treatment* pp. 189-193. CRC for Water Quality and Treatment, Noosa Lakes, Queensland.

Giraudel J. L. & Lek S. (2006) Ecological Applications of Non-supervised Artificial Neural Networks. In: *Ecological Informatics* (ed. F. Recknagel) pp. 49-67. Springer-Verlag, Berlin.

Goethals P., Dedecker A., Gabriels W. & De Pauw N. (2006) Development and Application of Predictive River Ecosystem Models Based on Classification Trees and Artificial Neural Networks. In: *Ecological Informatics* (ed. F. Recknagel). Springer-Verlag, Berlin.

Government A. (2004) National Water Quality Management Strategy.

Government S. A. (1962) *Myponga Reservoir and Pipeline*. Adelaide Government Printer, Adelaide.

Gower A. M. (1980) *Water Quality in Catchment Ecosystems*. John Wiley and Sons, New York.

Hanson M. J. & Stefan H. G. (1984) Side effects of 58 years of copper sulphate treatment of the Fairmont Lakes, Minnesota. *Water Resources Bulletin* 20: 889-900.

Happey-Wood C. M. (1988) Ecology of Freshwater Planktonic Green Algae. In: *Growth and Reproduction Strategies of Freshwater Phytoplankton* (ed. C. D. Sandgren). Cambridge University Press.

Harris G. P. (1986) *Phytoplankton Ecology: Structure, Function and Fluctuation*. Cambridge University Press.

Harris G. P. (1996) Catchments and Aquatic Ecosystems: Nutrient ratios, flow regulation and ecosystem impacts in rivers like the Hawkesbury-Nepean. pp. 57. CRC for Freshwater Ecology; CSIRO Australia, Canberra.

Herath G. (1997) Freshwater Algal Blooms and Their Control: Comparison of the European and Australian Experience. *Journal of Environmental Management* 51: 217-227.

Hoang H., Recknagel F., Marshall J. & Choy S. (2003) Elucidation of Hypothetical Relationships between Habitat Conditions and Macroinvertebrate Assemblages in Freshwater Streams by Artificial Neural Networks. In: *Ecological Informatics: Understanding Ecology by Biologically-Inspired Computation* (ed. F. Recknagel) pp. 179-194. Springer-Verlag, Berlin.

Holland J. H. (1975) *Adaptation in Natural and Artificial Systems*. Addison-Wesley, New York.

House J. & Burch M. (2002) Using Algicides for the Control of Algae in Australia - Registered Products for Use Against Algae and Cyanobacteria in Dams, Potable Water and Irrigation Water Supply Systems, in Australia. CRC for Water Quality and Treatment, Salisbury, SA.

Howard A. (1997) Computer simulation modelling of buoyancy change in Microcystis. *Hydrobiologia* 349: 111-117.

Hrudey S., Burch M., Drikas M. & Gregory R. (1999) Remedial Measures. In: *Toxic Cyanobacteria in Water: A guide to the public health consequences, monitoring and management.* (eds. I. Chorus & J. Bartram). E & FN Spon, London, New York.

Humpage A. & Froscio S. (2006) The Cyanobacterial Toxins. In: *Cyanobacteria: Management and Implications for Water Quality*. CRC for Water Quality and Treatment, Adelaide.

Imberger J., Patterson J. C., Hebbert B. & Loh J. (1978) Dynamics of Reservoir of Medium Size. *J. Hydraulic Div Proc. Am. Soc. ir. Eng* 104: 725-743.

Imteaz M. A. & Asaeda T. (2000) Artificial Mixing of Lake Water by Bubble Plume and Effects of Bubbling Operations on Algal Bloom. *Water Resources* 34: 1919-1929.

Ingleton G. (2003a) Nutrient Loads, Cyanobacteria Growth, and Algal Mitigation Options for Happy Valley Reservoir pp. 80. SA Water, Adelaide.

Ingleton G. (2003b) Scenarios for Growth Potential of Cyanobacteria and the Generation of Water Quality Hazards in Happy Valley Reservoir pp. 22. SA Water and AWQC, Adelaide.

Ingleton G. (2005) Personal Communication, Adelaide.
ITT-Industries (1990) No more pong from Myponga: Improving the reservoir's water quality. ITT Flygt Limited.

Jeffers J. N. R. (1999) Genetic Algorithms I. In: *Machine Learning Methods for Ecological Applications* (ed. F. A) pp. 107-122. Kluwer Academic Publishers, Boston, Dordrecht, London.

Jeong K.-S. & Joo G.-J. (2003) Modelling the succession of blue-green algae species in a flow regulated river (lower Nakdong River, S.Korea) by means of a Self-Organizing Map (SOM). *Unpublished manuscript*.

Jeong K.-S., Joo G.-J., Kim H.-W., Ha K. & Recknagel F. (2001) Prediction and elucidation of algal dynamics in the Nakdong River (Korea) by means of a recurrent artificial neural network. *Ecological Modelling* 146: 115-129.

Jeong K.-S., Kim D.-K., Jung J.-M., Kim M.-C. & Joo G.-J. (2007) Non-linear autoregressive modelling by Temporal Recurrent Neural Networks for the prediction of freshwater phytoplankton dynamics. *Ecological Modelling* (in press).

Jeong K.-S., Recknagel F. & Joo G.-J. (2003) Prediction and Elucidation of Population Dynamics of the Blue-green Algae *Microcystis aeruginosa* and the Diatom *Stephanodiscus hantzchii* in the Nakdong River-Reservoir System (South Korea) by a Recurrent Artificial Neural Network. In: *Ecological Informatics: Understanding Ecology by Biologically-Inspired Computation* (ed. F. Recknagel). Springer-Verlag, Berlin.

Johnstone P. C. (1994) Algal Bloom Research in Australia: a report of the current status of issues and the development of national research priorities. Agricultural and Resource Management Council of Australia and New Zealand, Water Resource Management Committee, Parramatta, NSW.

Jones G. & Orr P. (1994) Release and degradation of microcystin following algicide treatment of a *Microcystis aeruginosa* bloom in a recreational lake, as determined by HPLC and protein phosphatase inhibition assay. *Water Research* 28: 871-876.

Jongman R. H. G., ter Braak C. J. F. & O.F.R. v. T. (1987) *Data analysis in community and landscape ecology.* Cambridge University Press, Cambridge.

Jorgensen S. E. (1994) Models as instruments for combination of environmental theory and environmental practice. *Ecological Modelling* 75/76: 5-20.

Kalff J. (2002) *Limnology: Inland Water Ecosystems*. Prentice Hall Inc., New Jersey.

Kartam N., Flood I. & (Eds.) J. G. J. (1997) *Artificial Neural Networks For Civil Engineers: Fundamentals and Applications*. American Society of Civil Engineers, New York.

Karul C. & Soyupak S. (2006) A Comparison between Neural Network Based and Multiple Regression Models for Chlorophyll-*a* Estimation. In: *Ecological Informatics* (ed. F. Recknagel). Springer-Verlag, Berlin.

Kelly L. (1998) The diversity and abundance of algae in Myponga Reservoir over the 1997-1998 summer season. CRC for Water Quality and Treatment, Adelaide.

Kim D.-K., Cao H., Jeong K.-S., Recknagel F. & Joo G.-J. (2007) Predictive function and rules for population dynamics of *Microcystis aeruginosa* in the regulated Nakdong River (South Korea), discovered by evolutionary algorithms. *Ecological Modelling* in press.

Kim D.-K., Jeong K.-S., Whigham P. & Joo G.-J. (2007a) Winter diatom blooms in a regulated river in South Korea: explanations based on evolutionary computation. *Freshwater Biology* 52: 2021-2041.

Kirke B. K. (2000) Circulation, Destratification, Mixing and Aeration: Why and How? *Water* 27: 24-30.

Klapper H. (1991) *Control of Eutrophication of Inland Waters*. Ellis Horowood, New York.

Knoppert P. L., Rook J. J., Hofker T. & Osaka G. (1970) Destratification experiments at Rotterdam. *J. Am. Water Wastewater Assoc.* 62: 448-454.

Kohavi R. (1995) A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. In: *International Joint Conference on Artificial Intelligence*, Montreal, Quebec, Canada.

Kohonen T. (1982) Self-Organized Formation of Topologically Correct Feature Maps. *Biological Cybernetics* 43: 59-69.

Kohonen T. (1984) *Self-organisation and associative memory*. Springer-Verlag, Berlin, New York.

Kruk A., Lek S., Park Y. S. & Penczak T. (2007) Fish assemblages in the large lowland Narew River system (Poland): Application of the self-organizing map algorithm. *Ecological Modelling* 203: 45-61.

Lee J., Inn-Sil K., Lee E. & Kim K. A. (2007) Classification of breeding bird communities along an urbanization gradient using an unsupervised artificial neural network. *Ecological Modelling* 203: 62-71.

Lee J. H. W., Fernando T. M. K. G. & Wong K. T. M. (2004) Real Time Prediction of Coastal Algal Blooms Using Artificial Neural Networks. In: *6th International Conference on Hydroinformatics*, Singapore.

Lee J. H. W., Huang Y., Dickman M. & Jayawardena A. W. (2003) Neural network modelling of coastal algal blooms. *Ecological Modelling* 159: 179-201.

Lek S., Delacoste M., Baran P., Dimopoulos I., Lauga J. & Aulagnier S. (1996) Application of neural networks to modelling nonlinear relationships in ecology. *Ecological Modelling* 90: 39-52.

Lek S., Giraudel J. L. & Guegan J.-F. (2000) Neuronal Networks: Algorithms and Architectures for Ecologists and Evolutionary Ecologists. In: *Artificial Neuronal Networks* (eds. S. Lek & J. L. Guegan) pp. 3-25. Springer-Verlag, Berlin.

Lewis D. M. (2004) Surface Mixers for Destratification and Management of *Anabaena circinalis*. In: *School of Civil and Environmental Engineering* pp. 249. University of Adelaide, Adelaide.

Lewis D. M., Elliot J. A., Brookes J. D., Irish A. E., Lambert M. F. & Reynolds C. S. (2003) Modelling the effects of artificial mixing and copper sulphate dosing on phytoplankton in an Australian reservoir. *Lakes & Reservoirs: Research and Management* 8: 31-40.

Lewis D. M., Elliot J. A., Lambert M. F. & Reynolds C. S. (2002) The simulation of an Australian reservoir using a phytoplankton community model: PROTECH. *Ecological Modelling*: 107-116.

Maier H. (1995) Use of Artificial Neural Networks for Modelling Multivariate Water Quality Time Series. In: *Faculty of Engineering* pp. 559. University of Adelaide, Adelaide.

Maier H. (2005) Personal Communication, Adelaide.

Maier H., Dandy G. & Burch M. (1998) Use of artificial neural networks for modelling cyanobacteria *Anabaena* spp. in the River Murray, South Australia. *Ecological Modelling* 105: 257-272.

Mayer D. G. & Butler D. G. (1993) Statistical validation. *Ecological Modelling* 68: 21-32.

McAuliffe T. F. & Rosich R. S. (1990) The Triumphs and Tribulations of Artificial Mixing in Australian Water Bodies. *Water* 17: 22-23.

McGraw & Hill (2005) McGraw-Hill Dictionary of Science and Technology. The McGraw-Hill Companies, Inc.

McKay R. I., Hao H. T., Mori N. & Hoai N. X. (2006) Model-building with interpolated temporal data. *Ecological Informatics* 1: 259-268.

MDBC (1993) Technical Advisory Group Outcomes. Murray-Darling Basin Commission, Canberra.

Minski M. L. & Pappert S. (1969) *Perceptrons*. MIT Cambridge.

Morrall D. (2003) Ecological Applications of Genetic Algorithms. In: *Ecological Informatics: Understanding Ecology by Biologically-Inspired Computation* (ed. F. Recknagel) pp. 35-48. Springer-Verlag, Berlin.

Moss B. (1998) *Ecology of Freshwaters: Man and Medium, Past to Future*. Blackwell Science, United Kingdom.

Muttil N. & Lee J. H. W. (2005) Genetic Programming for Analysis and Real-Time Prediction of Coastal Algal Blooms. *Ecological Modelling* 189: 363-376.

National-Climate-Centre (2007) Drought Statement - Statement on Drought for the 5 and 12 month periods ending 31st December 2006. Australian Bureau of Meterology.

NeuralWare (1993) *Neural Computing: A Technology Handbook for Professional II and NeuralWorks Explorer*. NeuralWare Inc., Pittsburgh.

NeuroDimension (2003) *NeuroSolutions: The Neural Network Simulation Environment (Version 4.24 Developers) and NeuroSolutions for Excel (Version 4.21)*.
NRA (1990) Toxic blue-green algae. National Rivers Authority, London.

Oh H.-M., Ahn C.-Y., Lee J.-W., Chon T.-S., Choi K. H. & Park Y. S. (2007) Community patterning and identification of predominant factors in algal bloom in Daechung Reservoir (Korea) using artificial neural networks. *Ecological Modelling* 203: 109-118.

Paerl H. W. (1988) Growth and Reproductive Strategies of Freshwater Blue-Green Algae (Cyanobacteria). In: *Growth and Reproductive Strategies of Freshwater Phytoplankton* (ed. C. D. Sandgren). Cambridge University Press.

Park Y. S., Lek S., Scardi M., Verdonschot P. F. M. & Jorgensen S. E. (2006) Patterning exergy of benthic macroinvertebrate communities using self-organizing maps. *Ecological Modelling* 195: 105-113.

Parker R. A. (1968) Simulation of an aquatic ecosystem. *Biometrics* 24: 803-821.

Paruelo J. M. & Tomasel F. (1997) Prediction of functional characteristics of ecosystems: a comparison of artificial neural networks and regression models. *Ecological Modelling*: 173-186.

Pineda F. (1987) Generalisation of backpropagation to recurrent neural networks. *Phys. Rev. Lett.* 19: 2229-2232.

Pitois S., Jackson M. H. & Wood B. J. B. (2000) Problems associated with the presence of cyanobacteria in recreational and drinking waters. *International Journal of Environmental Health Research* 10: 203-218.

Power M. (1993) The predictive validation of ecological and environmental models. *Ecological Modelling* 68: 33-50.

Prepas E. E. & Murphy T. P. (1988) Sediment-water interactions in farm dugouts previously treated with copper sulphate. *Lakes & Reservoirs: Research and Management* 4: 161-168.

Ragab R. & Prudhomme C. (2002) Climate Change and Water Resources Management in Arid and Semi-arid Regions: Prospective and Challenges for the 21st Century. *Biosystems Engineering* 81: 3-43.

Recknagel F. (1989) *Applied systems ecology: approach and case studies*. Akademie-Verlag, Berlin.

Recknagel F. (1997) ANNA- artificial neural network model for predicting species abundance and succession of blue-green algae. *Hydrobiologia* 349: 47-57.

Recknagel F. (2001) Applications of machine learning to ecological modelling. *Ecological Modelling* 146: 303-310.

Recknagel F. (2002) *Ecology and Management of Freshwater Systems III - Subject Booklet*. University of Adelaide, Adelaide.

Recknagel F. (2003) Preface. In: *Ecological Informatics: Understanding Ecology by Biologically Inspired Computation* (ed. F. Recknagel). Springer-Verlag, Berlin.

Recknagel F. (2003b) Ecological Applications of Adaptive Agents. In: *Ecological Informatics: Understanding Ecology by Biologically-Inspired Computation* (ed. F. Recknagel) pp. 73-88. Springer-Verlag, Berlin.

Recknagel F. & Benndorf J. (1982) Validation of the Ecological Simulation Model "SALMO". *Int. Revue ges. Hydrobiologia* 67: 113-125.

Recknagel F. & Cao H. (2007 (in press)) Ecological informatics by means of neural, evolutionary and object-oriented computation. In: *Handbook for Ecological Modeling and Informatics* (eds. S. E. Jorgensen, F. Recknagel & T.-S. Chon). WIT Press, Southampton.

Recknagel F., Cao H., Kim B., Takamura N. & Welk A. (2006c) Unravelling and forecasting algal population dynamics in two lakes different in morphometry and eutrophication by neural and evolutionary computation. *Ecological Informatics*: 133-151.

Recknagel F., French M., Harkonen P. & Yabunaka K. (1997) Artificial neural network approach for modelling and prediction of algal blooms. *Ecological Modelling* 96: 11-28.

Recknagel F., Fukushima T., Hanazato T., Takamura N. & Wilson H. (1998) Modelling and prediction of phyto- and zooplankton dynamics in Lake Kasumingaura by artificial neural networks. *Lakes & Reservoirs: Research and Management* 3: 123-133.

Recknagel F., Kim B. & Welk A. (2004) Elucidation of Ecosystem Behaviour of Lake Soyang (South Korea) in Response to Seasons and Changed Management by means of Artificial Neural Networks. In: *International Association of Theoretical and Applied Limnology - 29th Congress*, Lahti, Finland.

Recknagel F., Talib A. & van der Molen D. T. (2006b) Phytoplankton community dynamics of two adjacent Dutch lakes in response to seasons and eutrophication control unravelled by non-supervised artificial neural networks. *Ecological Informatics* 1: 277-285.

Recknagel F., van Ginkel C., Cao H., Cetin L. & Zhang B. (2007b) Generic Limnological Models on the Touchstone: Testing the Lake Simulation Library SALMO-OO and the Rule-based *Microcystis* Agent for Warm-monomictic Hypertrophic Lakes in South Africs. *Ecological Modelling (under review)*.

Recknagel F., Welk A., Kim B. & Takamura N. (2005) Artificial Neural Network Approach to Unravel and Forecast Algal Population Dynamics of Two Lakes Different in Morphometry and Eutrophication. In: *Ecological Informatics* (ed. F. Recknagel). Springer-Verlag, New York.

Reynolds C. S. (1984) *The Ecology of Freshwater Phytoplankton*. Cambridge University Press, Cambridge.

Reynolds C. S. (1989) Physical determinants of phytoplankton succession. In: *Plankton Ecology: Succession in Plankton Communities* (ed. U. Sommer). Springer-Verlag, Berlin.

Reynolds C. S., Wiseman S. W. & Clarke M. J. O. (1984) Growth- and Loss-Rate Responses of Phytoplankton to Intermittent Artificial Mixing and Their Potential Application to the Control of Planktonic Algal Biomass. *Journal of Applied Ecology* 21: 11-39.

Robinson E. L., Irwin W. H. & Symons J. M. (1969) Influence of artificial destratification on plankton populations in impoundments. *Trans. Kint. Acad. Sci.* 30: 1-18.

Rogers R. D. & Vemuri V. (1994) Introduction - Time Series and the Forecasting Problem. In: *Artificial Neural Networks: Forecasting Time Series* (eds. R. D. Rogers & V. Vemuri). IEEE, Computer Society Press, California.

Romero J. R., Imberger J., Ewing T., Antenucci J., Deen A. & Craig R. (2003) A Real-time Decision Support System for Reservoir Management. In: *OzWater*, Perth.

Rumelhardt D. E., Hinton G. E. & Williams R. J. (1986) Learning representations by backpropagating error. *Nature* 323: 533-536.

Rykiel Jr E. J. (1996) Testing ecological models: the meaning of validation. *Ecological Modelling* 90: 229-244.

SAWater (2002) Happy Valley Reservoir Dam Wall Upgrade (ed. G. o. S. Australia), Adelaide.

SAWater (2003) Happy Valley Reservoir.

Scardi M. (1996) Artificial neural networks as empirical models for estimating phytoplankton production. *Mar.Ecol.Prog*: 289-299.

Scardi M. (2000) Neuronal network models of phytoplankton primary production. In: *Artificial Neuronal Networks: Application to Ecology and Evolution* (eds. S. Lek & J.-F. Guegan) pp. 116-129. Springer-Verlag, Berlin.

Scardi M. & Harding Jr L. W. (1999) Developing an empirical model of phytoplankton primary production: a neural network case study. *Ecological Modelling* 120: 213-223.

Schleiter I. M., Obach M., Wagner R., Werner H., Schmidt H.-H. & Borchardt D. (2006) Modelling Ecological Interrelations in Running Water Ecosystems with Artificial Neural Networks. In: *Ecological Informatics* (ed. F. Recknagel). Springer-Verlag, Berlin.

Sellner K. G., Doucette G. J. & Kirkpatrick G. J. (2003) Harmful algal blooms: causes, impacts and detection. *Journal of Industrial Microbiology and Biotechnology* 30: 383-406.

Shapiro J. (1990) Current beliefs regarding the dominance by blue-greens: the case for the importance of CO_2 and pH. *Verh.Int.Verein.Limnol.*: 38-54.

Silvert W. & Baptist M. (2000) Can Neuronal Networks be Used in Data-Poor Situations? In: *Artificial Neuronal Networks: Application to Ecology and Evolution* (eds. S. Lek & J.-F. Guegan) pp. 241-248. Springer-Verlag, Berlin.

Smalley T. M. (1996) Preliminary Study (for scenario analysis) on Eutrophication Control in the Myponga Reservoir using the Lake Ecosystem Model SALMO pp. 48. University of Adelaide, Adelaide.

Smalley T. M. (1998) Efficiency of Artificial Destratification fo Control of Blue Green Algae in the Myponga Reservoir. In: *Environmental Science and Management* pp. 73. University of Adelaide, Adelaide.

Sommer U. (1989) Toward a Darwinian Ecology of Plankton. In: *Plankton Ecology: Succession in Plankton Communities* (ed. U. Sommer). Springer-Verlag, Berlin, Heidelberg, New York, London, Paris, Tokyo.

Song M.-Y., Park Y. S., Inn-Sil K., Woo H. & Chon T.-S. (2006) Characterization of benthic macroinvertebrate communities in a restored stream by using self-organizing map. *Ecological Informatics* 1: 295-305.

South M. (1994) The application of genetic algorithms to rule finding in data analysis. In: *Department of Chemical and Process Engineering*. University of Newcastle upon Tyne.

Sparks D. L. & Schreurs B. G. (2003) Trace amounts of copper in water induce (beta)-amyloid plaques and learning deficits in a rabbit model of Alzheimer's disease. *Proc Natl Acad Sci USA* August 14.

Steel J. A. (1997) Scope and limitation in algal modelling - an example from the Thames Valley Reservoirs. *Hydrobiologia* 349: 27-37.

Stockwell D. R. B. (1992) Machine learning and the problem of prediction and explanation in ecological modelling. Australian National University, Australia.

Sung A. H. (1998) Ranking importance of input parameters of neural networks. *Expert Systems with Applications* 15: 405-411.

Talib A., Recknagel F., Cao H. & van der Molen D. T. (2005) Use of Recurrent ANN and Hybrid EA for the Prediction of Phytoplankton Abundance and Succession Before and After Eutrophication Control of Two Shallow Lakes. In: *MODSIM05: International Congress on Modelling and Simulation* (eds. A. Zerger & R. M. Argent) pp. 98-105. Modelling and Simulation Society of Australia and New Zealand Inc., Melbourne, Australia.

Thomas D., Kotz S. & Rixon S. (1999) Watercourse survey and management recommendations for the Myponga River Catchment. Environmental Protection Agency, Adelaide.

United-Water (2005) The Happy Valley Water Treatment Plant. United Water.

Van Hullebusch E., Chatenet P., Deluchat V., Chazal P. M., Froissard D., Botineau M., Ghestem A. & Baudu M. (2003) Copper Accumulation in a reservoir ecosystem following copper sulfate treatment (St. Germain Les Belles, France). *Water, Air, and Soil Pollution* 150: 3-22.

Van Tongren O. F. R., Van Liere L., Gulati R. D., Postema G. & Boesewinkel-De Bruyn P. J. (1992) Multivariate analysis of the plankton communities in the Loosdrecht lakes: relationship with the chemical and physical environment. *Hydrobiologia* 233: 105-117.

Varis O., Sirvia H. & Kettunen J. (1989) Multivariate analysis of lake phytoplankton and environmental factors. *Arch. Hydrobiol.* 117: 163-175.

Velzeboer R. M. A., Cugley J. A. & Patterson J. C. (1991) Modelling optimum conditions for reservoir destratification using mechanical mixers. Urban Water Research Association of Australia, Melbourne.

Visser P. M., Ibelings B. W., Van der Veer B., Koedood J. & Mur L. R. (1996) Artificial mixing prevents nuisance blooms of the cyanobacterium *Microcystis* in Lake Niewe Meer, the Netherlands. *Freshwater Biology* 36: 435-450.

Vollenweider R. A. (1968) The scientific basis of lake eutrophication, with particular reference to phosphrous and nitrogen as eutrophication factors. Technical Report DAS/DSI/68.27. OCED, Paris.

Walter M., Recknagel F., Carpenter C. & Bormans M. (2001) Predicting eutrophication effects in the Burrinjuck Reservoir (Australia) by means of the deterministic model SALMO and the recurrent neural network model ANNA. *Ecological Modelling* 146: 97-113.

Weiss M. & Kulikowski C. (1991) *Computer systems that learn: Classification and predictioon methods from statistics, neural networks, machine learning and expert systems*. Morgan Kaufman.

Weiss S. M. & Indurkhya N. (1998) *Predictive Data Mining: A Practical Guide*. Morgan Kaufmann Publishers Inc., San Francisco.

Welk A. (2003) Explanation and Prediction of Changes in Plankton Communities and Water Quality of a Temperate Stratified Lake by Artificial Neural Networks. In: *Discipline of Environmental Biology, School of Earth and Environmental Science*. University of Adelaide, Adelaide.

Welk A., Recknagel F. & Burch M. (2005) Ordination, clustering and forecasting of phytoplankton dynamics in the Myponga drinking water reservoir by means of supervised and non-supervised artificial neural networks. In: *International Congress on Modelling and Simulation, MODSIM 2005*, Melbourne, Australia.

Welk A., Recknagel F., Cao H., Chan W.-S. & Talib A. (2007) Rule-based agents for forecasting algal population dynamics in freshwater lakes discovered by hybrid evolutionary algorithms. *Ecological Informatics (under review)*.

Wetzel R. G. (1983) *Limnology*. Saunders College Publishing, USA.

Whigham P. (2005) Local Modelling by SOM partitioning and linear regression for Ecological Modelling. In: *MODSIM05* (eds. A. Zerger & R. M. Argent). The Modelling and Simulation Society of Australia and New Zealand Inc., Melbourne, Australia.

Whigham P. & Fogel G. B. (2003) Ecological Applications of Evolutionary Computation. In: *Ecological Informatics: Understanding Ecology by Biologically-Inspired Computation* (ed. F. Recknagel) pp. 49-66. Springer-Verlag, Berlin.

Whigham P. & Fogel G. B. (2006) Ecological Applications of Evolutionary Computation. In: *Ecological Informatics* (ed. F. Recknagel) pp. 85-107. Springer-Verlag, Berlin.

Whigham P. & Keukelaar J. (2001) Evolving structure-optimizing content. In: *Congress on Evolutionary Computation (CEC 2001)* pp. 1228-1235, Seoul, Korea.

Whigham P. & Recknagel F. (1999) Predictive Modelling of Plankton Dynamics in Freshwater Lakes using Genetic Programming. In: *MODSIM '99 International Congress on Modelling and Simulation*. The Modelling and Simulations Society of Australia and New Zealand Inc, Hamilton, New Zealand.

Whigham P. & Recknagel F. (2001) Predicting chlorophyll-a in freshwater lakes by hybridising process-based models and genetic algorithms. *Ecological Modelling* 146: 243-251.

Whitehead P. G., Howard A. & Arulmani C. (1997) Modelling algal growth and transport in rivers: a comparison of time series analysis, dynamic mass balance and neural network techniques. *Hydrobiologia* 349: 39-46.

Wilson H. (2004) Short Term Forecasting of Algal Blooms in Drinking Water Reservoirs using Artificial Neural Networks. In: *Discipline of Environmental Biology, School of Earth and Environmental Science* pp. 292. University of Adelaide, Adelaide.

Wilson H. & Recknagel F. (2003) A Generic Artificial Neural Network for Short Term Predictions of Algal Blooms in Lakes and Reservoirs. In: *Ecological Informatics: Understanding Ecology by Biologically-Inspired Computation* (ed. F. Recknagel) pp. 265-290. Springer-Verlag, Berlin.

Wong K. T. M. (2004) Red Tides and Algal Blooms in Subtropical Hong Kong Waters: Field Observations and Lagrangian Modelling. In: *Department of Civil Engineering*. University of Hong Kong, Hong Kong.

Zar J. (1984) *Biostatistical Analysis*. Prentice-Hall, New Jersey.