Short Term Forecasting of Algal Blooms in Drinking Water Reservoirs using Artificial Neural Networks

A thesis submitted for the award of Doctor of Philosophy

Hugh Edward Campbell Wilson
Discipline of Environmental Biology
School of Earth and Environmental Sciences
The University of Adelaide

April 2004

Abstract

Artificial neural networks (ANNs), trained to make short term forecasts of algal blooms in lakes and rivers, are potentially useful decision making tools for the operational management of eutrophication. This thesis addresses the question of whether a standardised, generic ANN model representation can be developed to achieve this goal. It is argued that four requirements need to be addressed; i) compatibility of models with existing water quality monitoring regimes, ii) stability and repeatability of training outcomes, iii) realistic and meaningful estimates of model performance and iv) explanation of predictions.

ANN model inputs were represented as *summary statistics* of *sliding time windows*. This approach was shown to increase the compatibility of typical time-series ANN model structures with datasets compromised by missing values and uneven sampling intervals. To improve stability, models were represented as an ensemble of ANNs trained on bootstrap samples of data (ie *bagging* (Breiman, 1994)). It was shown that the average prediction of the bagging ensemble was relatively unaffected by variance of the individual member models. Validation set representation was maximised by use of leave-k-out methods. Comparative error measures were devised to illustrate model performance characteristics relative to "naive" controls. A *sensitivity analysis through time* approach was utilised to explain the relative importance of input variables and to account complex interactions between variables.

Training data was available from six sites including Lake Biwa (Japan), Burrinjuck Dam (NSW, Australia), Darling River (NSW, Australia), Lake Kasumigaura (Japan), Myponga Reservoir (SA, Australia) and Lake Soyang (South Korea). These datasets were found to differ significantly from each other in terms of environmental characteristics and data availability. Models were developed to make one and two week forecasts. Predicted variables included chlorophyll *a* concentration and cell counts of the three most abundant algal species for each dataset. Experimental results showed that site/output specific input layers lead to better performance than site/output generic models. Furthermore, it is evident that ANNs capable of non-linear processing generalise better over local (short term) time scales, whereas perceptron models constrained to linear decision boundaries perform better over global (long term) scales.

Statement of Originality

This work contains no material which 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 Libraries, being available for photocopying and loan.

12-7-09

Hugh Wilson

Ackowledgements

I would like to thank my supervisors, Friedrich Recknagel and Hugh Possingham. Friedrich, my principal supervisor, has been instrumental in the development of this thesis with his enthusiasm, encouragement, ideas, advice and support. Likewise, Hugh, my co-superviser, has provided invaluable help with advice, support and physical exercise in the form of lunchtime volleyball sessions at Roseworthy Campus.

This work would not have been possible without the financial and material support from both SA Water and the Cooperative Research Centre for Water Quality and Treatment. Also, I have been fortunate to have excellent technical, administrative and moral support from the staff at Roseworthy, Waite and North Terrace. So thank you to Rob Murray, Brian Glaetzer, John Willoughby, Stan Eckert, Emiel Storken, John Davey, Keith Cowley, Richard Norrish and many other individuals who create the environment at Adelaide University for postgraduate students to conduct their research.

None of this would have been possible without something for the ANNs to learn. Painstaking work of countless scientists and technicians over many years has created the water quality datasets used for training the models in this thesis. I am indebted to all these individuals who, unlike me, braved the elements in all seasons to collect the data that I have used.

During these years of research work, I have been fortunate to enjoy the company, advice and diversion of wonderful fellow students and office mates. Their insights have given me many leads (intentional or otherwise) on issues instrumental to my ideas. But most importantly, they have kept me sane by reminding me that life gets pretty boring if we take it seriously all the time. So thank you to Jason Bobbin, Tanja Jankovic, Tumi Bjornsson, Mardi van der Wielan, Anita Talib, Lydia Cetin, Nelli Horrigan, Alla Baklan, Huong Huang, Amber Welk, Li Wen, Drew Tyre, Hongqing Cao and others I may have regrettably forgotten to mention.

Also I must thank my family and friends who, I must say, were starting to avoid the question of the thesis. In particular, I would like to thank my parents, Ann and Jim, my sisters, Alice and Ruth and my parents in law, Kit, Rod and Chris, for their love, support, babysitting and meals.

Finally and most importantly I must thank my beautiful wife Jacqui and my wonderful daughter Olivia. Their love, patience and understanding is part of everything I do.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Accelerated eutrophication of freshwater lakes and rivers, as a result of human activity, leads to increased frequency and severity of algal blooms and a succession in algal dominance towards potentially toxic cyanobacteria (Young et al., 1996). Algal blooms adversely affect the value of freshwaters as a natural resource by increasing treatment costs for drinking water production, reducing recreational amenity, causing adverse environmental effects such as reduced biodiversity and causing economic loss to aquacultural and agricultural activities as a result of toxin release by blue-green algae (Senate Standing Committee on Environment Recreation and the Arts (Aust.), 1993). Furthermore, cyanotoxins have been identified as a direct cause of human mortality (Azevedo et al., 2002) and longer term health risks (Freitas de Magalhẽs et al., 2001; Ueno et al., 1996; Ueno and Nagata, 1997). The growing awareness of the potential dangers to public health posed by cyanotoxins has prompted calls to relieve the emphasis on water treatment facilities by development of effective in-lake management tactics for control of algal blooms (Burch and Nicholson, 2000).

Models that predict variables associated with eutrophication are useful decision making tools for the development of management responses. They may be used to set goals for strategies to limit nutrient loading (for example, Vollenweider (1970)), or to carry out scenario analyses by which lake responses to competing proposals are compared (Ferguson, 1997). Also, time-series models can provide real-time forecasts of relevant variables to ensure the correct timing of a variety of tactical responses, such as those listed in appendix A.

French and Recknagel (1994), Recknagel et al. (1997), Maier et al. (1998) and others suggested that *artificial neural networks* (ANNs) be used as an alternative to classical empirical and deterministic approaches for modelling eutrophication variables. ANNs have captured the interest of ecologists because of their properties as "universal approximators" (Hornik, 1993) – that is, their ability to "learn" models without the *a–priori* assumptions or simplifications of existing empirical and deterministic approaches (Lae et al., 1999; Lek and Guégan, 1999). This

property promises a purely empirical modelling method that captures the *realism* of complex deterministic models without the headaches of formalising and parameterising process equations.

In practice, ANNs have been shown to promise equivalent or superior performance to traditional modelling approaches for a wide range of modelling problems (Lek and Guégan, 1999). Specifically, Recknagel et al. (1997) and Recknagel and Wilson (2000) showed that ANNs applied to time-series modelling of eutrophication variables, such as algal abundance, can outperform existing empirical and deterministic models. Thus, it is clear that the power of ANNs as model approximators is sufficient to meet the requirements of operational or strategic decision making tools. However, Maier and Dandy (2000) points out that, in practice, ANN models and methods are being applied in an *ad-hoc* manner leading to sub-optimal performance, difficulties in making reasonable comparisons and, most importantly, confusion amongst potential users.

The general thrust of this thesis is to attempt to answer the question; is it possible to develop standardised or *generic* ANN model representations and methodologies for forecasting phytoplankton abundance that guarantee optimum predictive performance, repeatability and ease of use? Since the ANN development *process model* consists of a number of (possible interacting) steps (Maier and Dandy, 2000), the answer to such a question depends on identification and resolution of not one, but a range of issues. The principle issues identified are as follows;

- **Database Compatibility**

  Being empirical in nature, ANN modelling requires long (5–20 years) time-series of relevant variables for training purposes. Database compatibility refers to the problem of selecting appropriate input and output variables for the ANN model from these time-series. It is argued that two issues need to be addressed. Firstly, there is the problem of selecting a subset of input variables that have causative and/or correlative links with the output variables. A review of six datasets (chapter 3) shows that each study site has a unique set of monitored variables, which means that the task of input selection must be addressed for each new model application. This thesis compares and contrasts three approaches to this task; a *generic model* comprised of the set of variables common to all datasets (chapter 5), a forward selection approach and a backwards elimination approach based on *data strip mining* (Embrechts et al., 2001) (chapter 6).

  Secondly, there is the problem of modelling links between past, present and future states of the system. To be useful in a tactical decision support role ANN models should make *forecasts* rather than *same-day* predictions (Lee et al., 2003), meaning that they must define links between present and future states of the system by using *time delay connections*. However, it is shown in the literature review (chapter 2) and the analysis of six datasets (chapter 3) that typical monitoring data are rarely well ordered sequences of

observations – sampling is usually irregular and observations are frequently missing. The usual approach to dealing with this issue is to interpolate a large number of synthetic observations between the actual sample dates for each variable.

It is proposed that modelled variables be represented not as values at discrete dates, but as the summary statistic of a sliding window in time. It is shown that such a representation is capable of approximating forecasting models in the context of uninterpolated raw datasets. It is argued that, as a means of ensuring compatibility between time-series ANN models and typical datasets, the *input-window* representation has many advantages over interpolation of data. Also, such a representation provides scope for further exploration of alternative window summary approaches.

- **Model Stability**

  While reasonable *model approximation* by ANNs on training sets is shown to be a straightforward task, ensuring optimum *generalisation* to independent population data is somewhat more difficult. Usually, this is carried out by tuning some determinant of ANN *fitting power*, such as the hidden layer size or training time, by means of cross-validation. However, Breiman (1996b) points out that ANNs belong to a class of inference methods that exhibit significant instability even when regularised in this way.

  It is proposed that, as suggested by Breiman (1994), stabilisation is achieved by representing the model as an ensemble of many ANNs trained on bootstrap samples of data (ie bootstrap aggregation or *bagging*). Such a representation is hypothesised to have the effect of "cancelling out" uncorrelated, erroneous predictions and emphasising the correlated, correct predictions by member models of the ensemble. It is shown in chapter 4 that bagging significantly improves modelling outcomes by reducing prediction error and reducing sensitivity of performance to *overfitting*.

- **Performance Estimation**

  Accurate performance estimation is required to measure the effect of changes to the model representation and/or methodology and to determine the suitability of the model for its intended purpose. However, it is shown (chapter 2) that, in practice, performance estimation may be compromised in three ways; poor representation of independent validation sets, the use of validation data for model selection and contamination of time-series causing models to have unauthorised access to information from the future relative to the current forecast period.

  It is proposed that validation methods based on *resampling* (so called *rotation estimators*) allow use of the entire sample for model performance estimation without seriously limiting training set representation. Furthermore, it is proposed that, when used in combination with bagging, there is no need

to use validation data in the process of model selection. Grouping validation samples into discrete time periods and imposing a hold-out period following each validation block, is suggested as a means to avoid contamination of time-series information between training and validation sets. Experiments with two different types of rotation estimators – the leave-one-out bootstrap (L1OB) and 20-fold blocked cross-validation (CV) – show that, in the context of a time-series model, the time lag between training and validation data has a significant affect on performance estimation outcomes.

- **Transparency of Predictions**

  There is concern amongst ecologists regarding the *black–box* nature of ANNs – that is, the lack of accompanying explanations with predictions. *Transparency* of the model, as well as allowing knowledge discovery, enables further validation that the model is making reasonable inferences from the data. Sensitivity analysis is most commonly used to determine the relative importance of input variables. However, it is argued (chapter 2) that existing approaches to sensitivity analysis do not account for complex interactions between input variables and do not consider the model's characteristics relative to the entire model input–space.

  To this end, a *sensitivity analysis through time* technique is proposed that takes account of the following assumptions with regards to learned models;

  - Inputs are likely to have non–linear relationships with output variables.
  - Inputs may have complex interactions with other input variables with respect to relationships with output variables.
  - ANNs are inherently unreliable when asked to make extrapolations (Geman et al., 1992).

While none of the ideas proposed are entirely original on their own, together they represent a new approach to computational modelling based on typical environmental time-series. This thesis presents a validation of this approach for a broad range of data due to the kindness of a number of scientists and water resource managers in donating many years of water quality monitoring data and algal cell counts. Models are developed for a total of six sites including Burrinjuck Dam, Myponga Reservoir and the Darling River in Australia, Lakes Biwa and Kasumigaura in Japan and Lake Soyang in South Korea.

Organisation of Thesis

Chapter 2 introduces principles of ANN knowledge representation, supervised learning and model development. A selection of published applications of ANNs to time-series modelling of algal abundance is reviewed. Key issues affecting

the development of a generic model representations are identified. Resolutions to these issues are proposed.

Chapter 3 reviews the six datasets available for this study in terms of situation, climate, morphometry, water quality and data availability. Two models are proposed as starting points for the modelling work conducted for this thesis – a *generic* model comprised of commonly available variables and a *site specific* model unique to each dataset.

Chapter 4 presents results regarding the the effect of a number of ANN "meta–parameters" on model inference properties. Specifically, this chapter describes the effect of training algorithm, hidden layer configuration, stopping error of training and model aggregation through *bagging* (Breiman, 1994) on approximation and generalisation characteristics.

Chapter 5 comprehensively validates the generic ANN model identified in chapter 3 for all six datasets at predicting a total of 21 output variables taking into account the findings of chapter 4. A number of standardised and comparative error measures are introduced for performing meaningful analysis of the predictive performance of models. Finally, results of a sensitivity analysis are presented.

Chapter 6 investigates two approaches to identification of optimum site specific models. Also, an investigation is carried out to determine the interaction between the validation method, the input layer type and the non–linear processing capacity of the ANN on model performance outcomes.

Chapter 7 summarises the achievements of the thesis.

Chapter 2

Development of ANN Models of Phytoplankton Abundance

2.1 Introduction

This chapter introduces the principles of model representation and supervised learning by ANNs. It also reviews how ANNs are being applied to the problem of modelling the dynamics of phytoplankton abundance in lakes, rivers and marine ecosystems. Additionally, a number of proposals are outlined for improving the *compatibility* of ANN models with typical datasets, increasing the *stability* of model inference and improving the *accuracy* of model performance estimations and knowledge discovery.

2.2 Knowledge Representation and Inference by Artificial Neural Networks

2.2.1 ANN Structure and Information Processing

ANNs are derived from theories of brain structure and function and are intended to model:

- the ability of the brain to apply principles of parallel processing in solving difficult problems such as image recognition faster than conventional serial computing devices.
- the brain's property of self adaptation – that is, the ability to learn from experience.

(Cheng and Titterington, 1994)

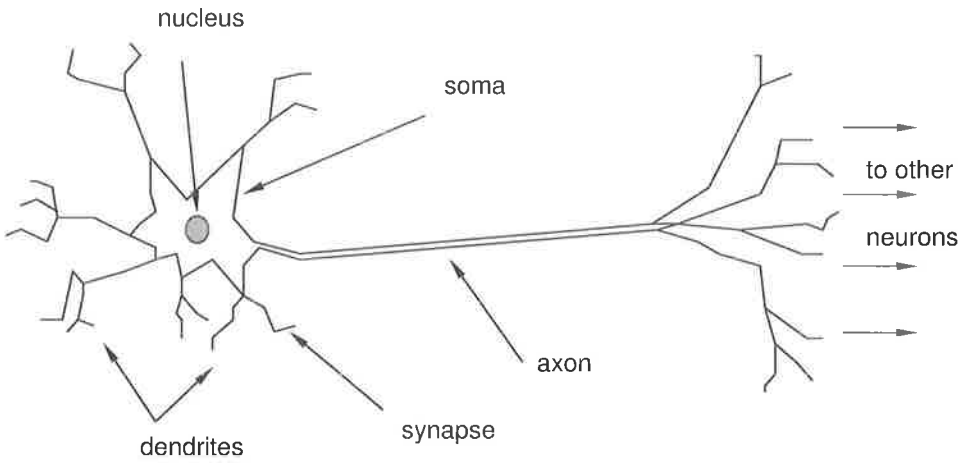

Figure 2.1: Schematic of a neuron (after Cheng and Titterington (1994))

Figure 2.1 shows a diagrammatic representation of the fundamental processing unit of the brain – a neuron. Neurons are comprised of a cell body, or *soma* and two types of radiating branching structures called *dendrites* and *axons*. In the brain, neurons are highly networked with the axons and dendrites of many neighbouring cells being interconnected by means of electro-chemical interfaces called *synapses*.

The sequence of events in neural processing commences when an electrical discharge through an axon (ie an *action potential*), causes the release of neurotransmitters across the synaptic cleft separating the transmitting axon from the receiving dendrite of a connected neuron. The neurotransmitters are bound on the receiving side of the synapse causing the induction of a small electric charge, called a *post-synaptic potential* or PSP, in the relevant dendrite. Incoming PSPs from a number of receiving synapses diffuse from the dendrites into the soma where they have either an excitory or inhibitory effect on the total charge of the cell. When the somatic potential reaches some threshold level, an action potential is discharged through the axon stimulating the release of neurotransmitters in "downstream" synapses. Learning is thought to be caused by the adaptability, or *plasticity* of certain aspects of the brain's structure thus allowing the alteration of neural processing as a result of experience. A key element of the brain's plasticity is the efficiency of the synaptic interfaces between interconnected neurons (Amit, 1989).

McCulloch and Pitts (1943) formalised the theory regarding the behaviour of a single neuron into a simple mathematical model (ie the McCulloch-Pitts neuron)

$$y = \text{sgn}\left(w_0 + \sum_{i=1}(x_i w_i)\right) \tag{2.1}$$

In this model, x_i is a boolean representing the firing of a neuron connected "upstream", w_i is the synaptic efficiency (generally called a *weight*) with respect to x_i, w_0 represents the threshold of the soma at which an action potential is fired and y is a boolean indicating whether or not the neuron has fired at a given time step. The input-output form of the model neuron may be generalised as;

$$y = f(\phi(x, w)) \tag{2.2}$$

Including both ϕ and f in the model is useful for identifying the combination and the activation components respectively (Cheng and Titterington, 1994). ϕ (the *combination function*) is a vector to scalar function calculating the total input activation (or the total post synaptic charge, to follow the neuron analogy). f, (the *activation function*), calculates the neuron output from the activation (the so-called firing rate). In practice, ϕ is generally a summation, whereas f is commonly substituted for one of any number of arbitrary linear, non-linear or step functions including;

- $f(a) = \text{sgn}(a)$, producing binary (± 1) output.
- $f(a) = (sgn(a) + 1)/2$ producing binary $(0/1)$ output.
- $f(a) = (1 + e^{-a})^{-1}$ producing continuous non-linear output between 0 and 1.
- $f(a) = \tanh(a)$ producing continuous non-linear output between -1 and 1.
- $f(a) = a$ producing linear output (the identity function).
- $f(a) = |a|$ producing non-negative output.

A key element of the McCulloch-Pitts model is inclusion of time in the form of an arbitrary delay between presentation of the inputs and the processing of the output. McCulloch and Pitts (1943) (cited by Amit (1989)) proposed that when the model neurons are combined into temporal sequences, with the outputs of one neuron feeding the inputs of another, the activity of the output neuron will be the truth value of any binary logic operation represented at the input neurons. More recently it has been shown that ANNs consisting of 3 or more layers of neurons (ie an input layer, an output layer and an arbitrary number of interceding hidden layers), where the activation functions of the hidden layer neurons are continuous and non-linear, are capable, given sufficient hidden layer neurons, of mapping any continuous function between inputs and outputs (Hornik, 1993). This form of ANN, called a feedforward multi-layer perceptron (MLP), is represented diagrammatically in figure 2.2.

Cannon and Whitfield (2002) mathematically represent an MLP thus;

$$y = \sum_j \tanh\left(\sum_i x_i {}^1w_{ij} + {}^1b_j\right) {}^2w_j + {}^2b \tag{2.3}$$

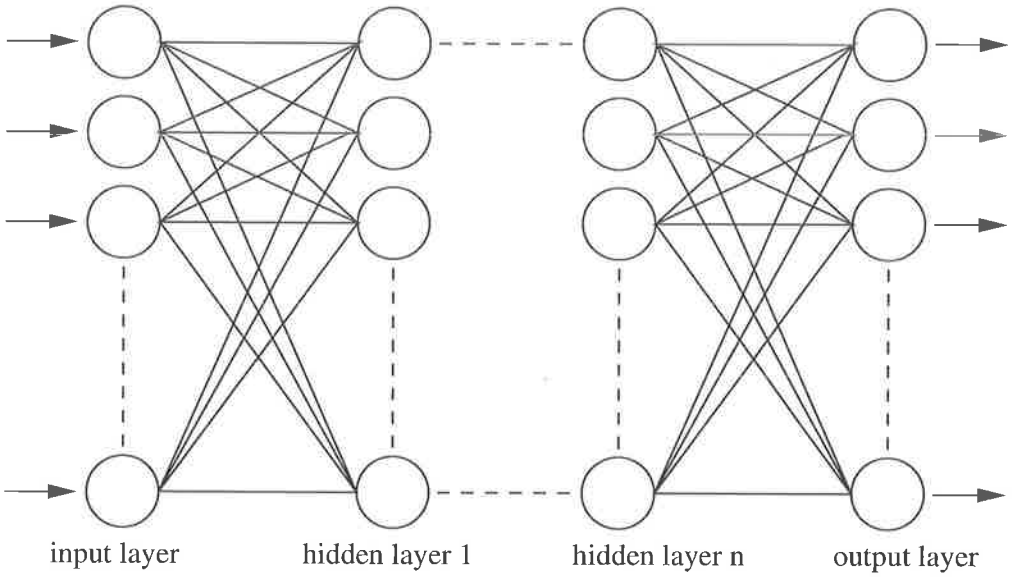

Figure 2.2: Feedforward multilayer perceptron (MLP)

In this model, x_i represents the input variable, ${}^1w_{ij}$ and ${}^2w_j$ are the input-hidden and the hidden-output layer weights and ${}^1b_j$ and ${}^2b$ are the input-hidden and hidden-output layer biases. It is assumed that tanh is the activation function of the hidden layer neurons and that the output layer uses the identity function.

In terms of information processing properties, a single artificial neuron (ie, a perceptron) is functionally equivalent to a multiple linear regression equation (Cheng and Titterington, 1994). Regardless of the activation function used, a perceptron with n inputs is able to define a single decision boundary, or hyperplane, of $n-1$ dimensions with respect to the n-dimensional input space (Wasserman, 1989). The weights w_0 to w_i each represent the slope of the decision boundary with respect to the respective inputs. The threshold (otherwise known as *bias*) represents the intercept.

Figure 2.3 shows that the type and complexity of decision boundaries mapped by a MLP depends on the number and configuration of hidden layer neurons. A perceptron is limited to classification of linearly separable functions. Two units in a single hidden layer allow the MLP to map open convex decision boundaries. Three or more units in a single hidden layer allow the MLP to map closed convex decision boundaries, where the upper limit to complexity depends on the number of hidden units used. When hidden layer units are arranged into two or more layers, the MLP can approximate concave decision boundaries. Thus, it can be concluded that while the processing and memory of each neural processor in an ANN is very simple, when arranged to allow parallel processing they are potentially powerful computational devices capable of mapping non-linearly separable classifications (Wasserman, 1989; Hinton, 1992)

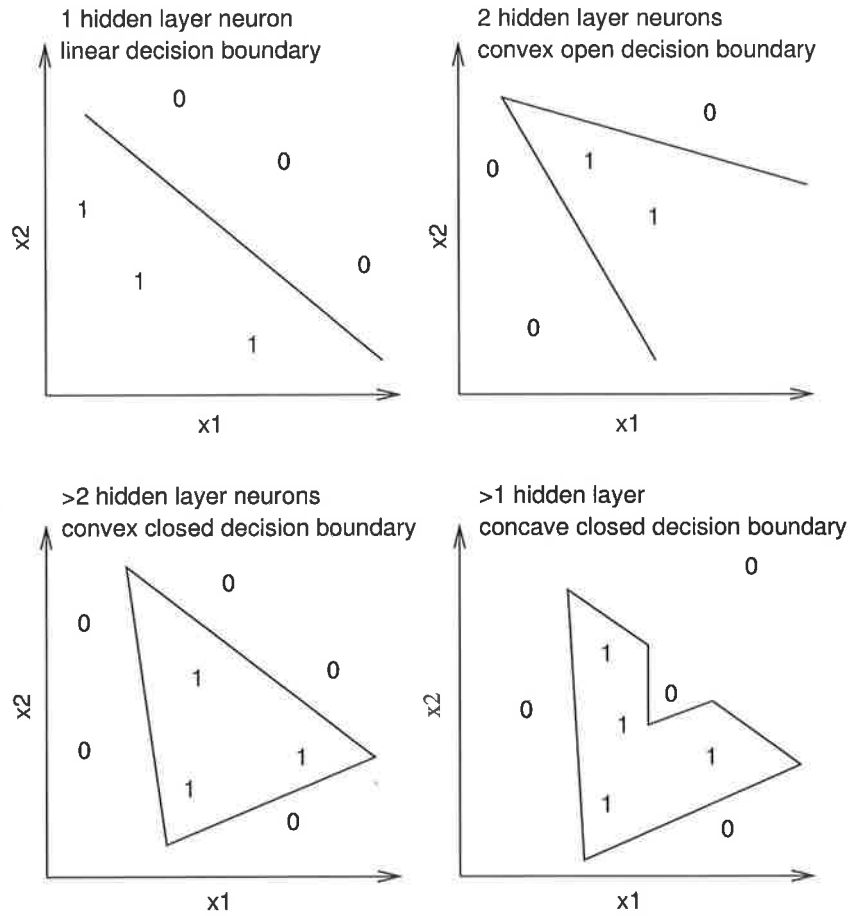

Figure 2.3: Effect of MLP architecture on decision boundaries for a two input model (after Wasserman (1989))

2.2.2 Supervised Learning by ANNs

The learning, or *approximation*, problem describes the task of mapping an m-space to any n-space within a given distortion criteria and time limit (Takahashi, 1993). Moody (1991) defines the approximation problem as follows; given a set of real input/output pairs $\xi = \{\xi^i = (x^i, y^i); i = 1, \ldots, n\}$ generated by the "signal plus noise model" outlined in equation 2.4, the task of the approximation exercise is to estimate a model $\widehat{\mu}(x)$ of $\mu(x)$ on the basis of training set ξ.

$$y^i = \mu(x^i) + \varepsilon^i \tag{2.4}$$

where y^i = dependent variable
x^i = independent variable sampled with probability density $\Omega(x)$
ε = independent noise sampled with density Ψ
$\mu(x)$ = an unknown function

2.2.2.1 Historical Context

Rosenblatt (1962) outlined the *single-unit perceptron convergence theorem* showing that if a training set is linearly separable by a single hyperplane into two distinct classes, application of the *generalised delta rule* (Widrow and Hoff, 1960) to updating connection weights allows a perceptron to approximate the hyperplane in a finite number of steps. However, Minsky and Papert (1969) pointed out the inability of perceptrons at mapping non-linearly separable functions, such as the exclusive and/or (XOR) (see figure 2.4), places a limitation on their usefulness. It was proposed that such a mapping is possible with multi-layered perceptrons (MLPs), but at the time no suitable training algorithm had been devised for such architectures.

This limitation was overcome by the development of an adapted version of the generalised delta rule that was capable of addressing the approximation problem for MLPs. This discovery was made independently by Werbos (1974), Parker (1982) and Rumelhart et al. (1986). The latter authors succeeded in introducing the approach, which they called *backpropagation*, to a wide audience leading to a resurgence of interest in ANNs since the late 1980s (Hecht-Nielsen, 1990).

2.2.2.2 Backpropagation

Backpropagation is a recursive technique based on the principle of gradient descent. As the name suggests, gradient descent involves finding the slope of the error with respect to the network weights and modifying the weights by an amount in negative proportion to the slope. This process is iterated until the minimum of the goal function with respect to the network weights is reached (ie *convergence*).

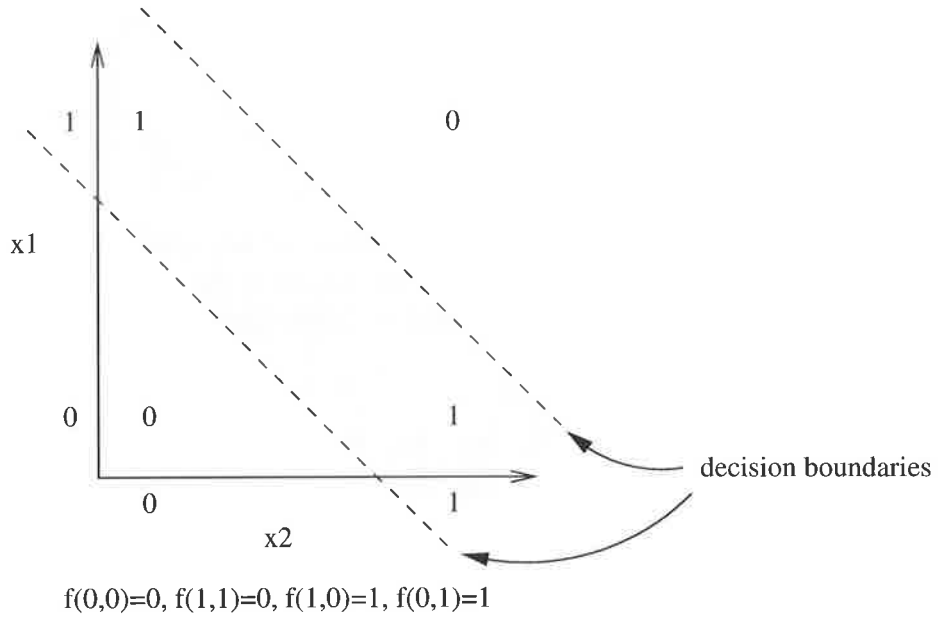

Figure 2.4: Classifying the exclusive OR (XOR)

Thus, if the prediction error with respect to the weights is $E(w)$, gradient descent can be expressed according to the recursive rule;

$$w_i(t+1) = w_i(t) - \rho \frac{\partial E(w_i(t))}{\partial w_i} \tag{2.5}$$

where t is a positive integer representing the training data iteration number, ρ is a small positive constant called the *learning rate* and $\partial E(w_i(t))/\partial w_i$ is the partial derivative of E with respect to the weight on input i. Now, given a random starting point $w(0)$, gradient descent implements a search strategy whereby a sequence of weight vectors;

$$w(0), w(1), w(2), \ldots, w(t), \ldots$$

is generated such that;

$$E[w(0)] \geq E[w(1)] \geq E[w(2)] \geq \ldots \geq E[w(t)] \geq \ldots$$

(Hassoun, 1995)

Thus, as t tends to infinity and ρ tends to 0, gradient descent is guaranteed to converge on a local minimum on the error surface (Hinton, 1992). Backpropagation applies the chain rule of differential calculus to find the slope of the goal function with respect to the network weights. In an ANN with i inputs, j hidden neurons and k output neurons, $\partial E/\partial w_{jk}$ (ie the slope of the error with respect to

the weight of the connection between hidden unit j and output unit k), is calculated as follows;

$$\frac{\partial E}{\partial w_{jk}} = \sum_{\text{cases}} \frac{\partial E}{\partial y_k} \cdot \frac{\partial y_k}{\partial a_k} \cdot \frac{\partial a_k}{\partial y_{jk}} \tag{2.6}$$

where a is the total input and y is the total activation of any given unit. Further expansion to find $\partial E/\partial w_{ij}$ (ie the slope of E with respect to the weight connecting input unit i to hidden unit j) can be achieved by further application of the chain rule;

$$\frac{\partial E}{\partial w_{ij}} = \sum_{k} \frac{\partial E}{\partial y_k} \cdot \frac{\partial y_k}{\partial a_k} \cdot \frac{\partial a_k}{\partial y_j} \cdot \frac{\partial y_j}{\partial a_j} \cdot \frac{\partial a_j}{\partial w_{ij}} \tag{2.7}$$

Calculation of partial derivatives of E with respect to connection weights, using the chain rule, requires that the error term and the activation function are both differentiable. Hence a continuous activation function, such as the logistic function $f(x) = (1 + e^{-}x)^{-}1$, is commonly used. Error is usually calculated thus;

$$E = \sum_{cases} \frac{1}{2}(y - y')^2 \tag{2.8}$$

where y is the desired output and y' is the actual output.

Each presentation of training examples undergoes two phases. Firstly, unit activations, given a vector of input examples and the current network weights, are propagated forwards from the input layer to the output layer. Secondly, the error derivatives, given the target values and activations, are propagated from the output layer back to the first hidden layer (Weiss and Kulikowski, 1991). At each backward propagation step, connection weights are updated according to equation 2.5.

2.2.2.3 Alternatives to Backpropagation

There are now many learning algorithms for feedforward MLPs that are claimed to be considerably more efficient than conventional backpropagation. In particular, methods that utilise second order information of the error surface (ie the curvature) are theoretically able to calculate step size and direction more accurately leading to faster training (Alpsan et al., 1995). Pandya and Macy (1996) point out that such second order methods eliminate some of the persistent drawbacks of backpropagation that cause slow learning (eg local minima and shallow plateaus on the error surface).

There are many approaches to utilising second order information. One method is based on Newton's method;

$$w_{t+1} = w_t - H^{-1} \nabla_{w^t} E \tag{2.9}$$

where H^{-1} is the inverse of the Hessian matrix of second derivatives. A problem with this approach is that the memory and time requirements needed to calculate H^{-1} rise exponentially with the number of connection weights in the ANN (Maier and Dandy, 2000). This processing bottleneck can be overcome by so-called "quasi-Newton" methods that estimate H^{-1} in a more computationally efficient manner (Alpsan et al., 1995). Alternatively, conjugate gradient methods, such as Møller (1993), utilise line-search methods to determine the step size and analytical techniques to determine the optimum momentum.

Whilst second order methods are theoretically more efficient than first order approaches such as backpropagation, Saarinen et al. (1993) argues that they have two disadvantages that may lead to slow training. Firstly, they have a higher processing and memory overhead per iteration than backpropagation. Secondly, poor "numerical conditioning" of many datasets may lead to bad training performance.

2.2.3 Unsupervised Learning

In contrast to supervised training methods, unsupervised training is not guided by known output or "answers" in the training data. This approach stems from the ideas of Hebb (1949), who proposed that repeated firing of a neuron would affect the efficiency of firing of neighbouring neurons and that the connection between two neurons strengthened with simultaneous firing. Examples of ANNs that perform unsupervised learning include the *Self Organising Map* (SOM) (Kohonen, 1982), the *Hopfield Net* (Hopfield, 1982) and the *Boltzmann machine* (Ackley et al., 1985). In practice, unsupervised learning is often used to cluster objects on the basis of perceived closeness in n-dimensional hyperspace (Lek et al., 2000). Examples of applications of unsupervised ANNs include Chon et al. (1996), Foody (1999) and Brosse et al. (2001). Since the present study is focused on time-series-models for prediction of phytoplankton abundance, the discussion will be limited to supervised ANNs.

2.3 Developing Predictive Models – An ANN Model Development Process-Model

It is proposed that the basic supervised ANN model development procedure, as carried out by many practitioners, can be reduced to a series of hierarchical steps illustrated in figure 2.5. This process is relatively straightforward, with a number of decisions or outcomes at each step. The starting conditions for most applications are a database of observations and assumptions, or hypotheses, regarding

input-output relationships between variables. In general, the desired end point of the process is a model that may be used to generate predictions or to elucidate knowledge. The intervening steps comprise the model inference process, which, for the purposes of this discussion, have been decomposed into approximation and generalisation sub-tasks. This section reviews each of the steps in the procedure outlining the issues and problems to be dealt with.

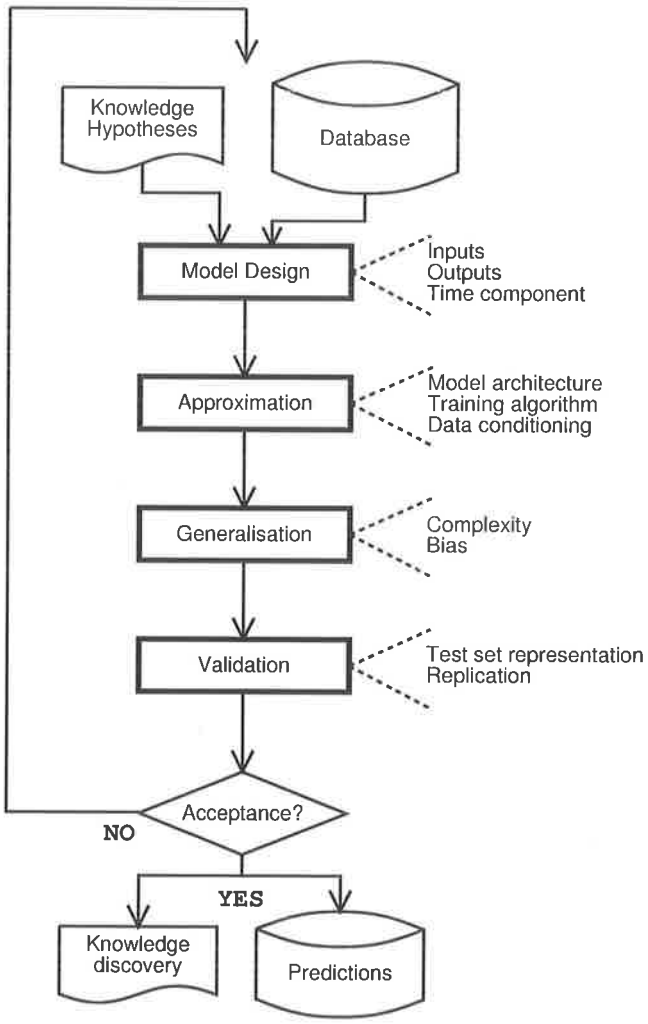

Figure 2.5: A process model for ANN model development.

2.3.1 Step 1 – Model Design

This step involves selection of independent and dependent variables, from a database, that will represent the input and output layers respectively of the ANN model. In general, practitioners choose inputs that are known to have some deterministic link with the outputs (Maier and Dandy, 2001). Scardi (2001) proposed that inputs

known to have a correlative rather than deterministic relationship (ie so-called "co-predictors") may also benefit modelling outcomes. Where the database used is a time-series of observations and it is hypothesised that relationships exist between past and present states in the data, some component representing these links needs to be built into the model design. A commonly used approach is the "Time Delay Neural Network" or TDNN (Waibel, 1989) whereby the inputs were represented with varying delays in time relative to the output variable. This enables the ANN to learn the dynamic properties of a set of moving inputs. Figures 2.6 and 2.7 compare the structures of ANN models with and without time-delay connections.

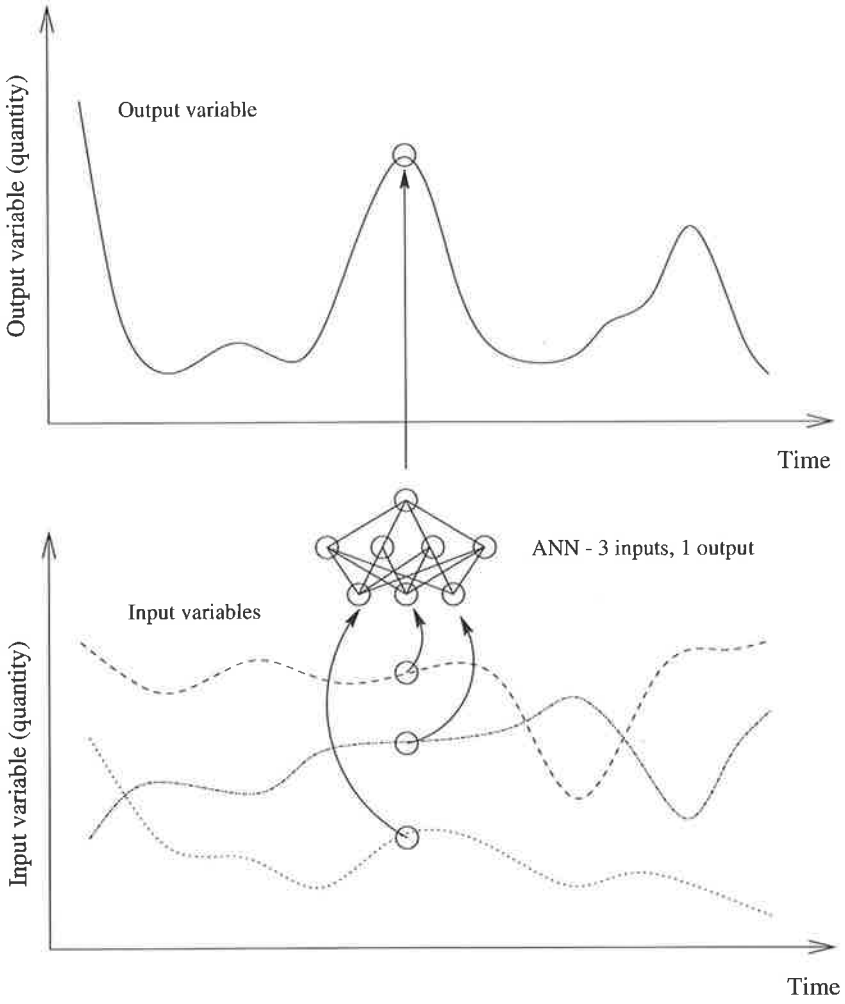

Figure 2.6: The "same day predictor" neural network structure.

One of the problems facing practitioners when designing ANN models is the so-called "curse of dimensionality", whereby the state-space increases exponentially with an increase in the dimensionality of the problem (Bellman, 1961). According to Sarle (2001), excessive input dimensionality causes 2 problems;

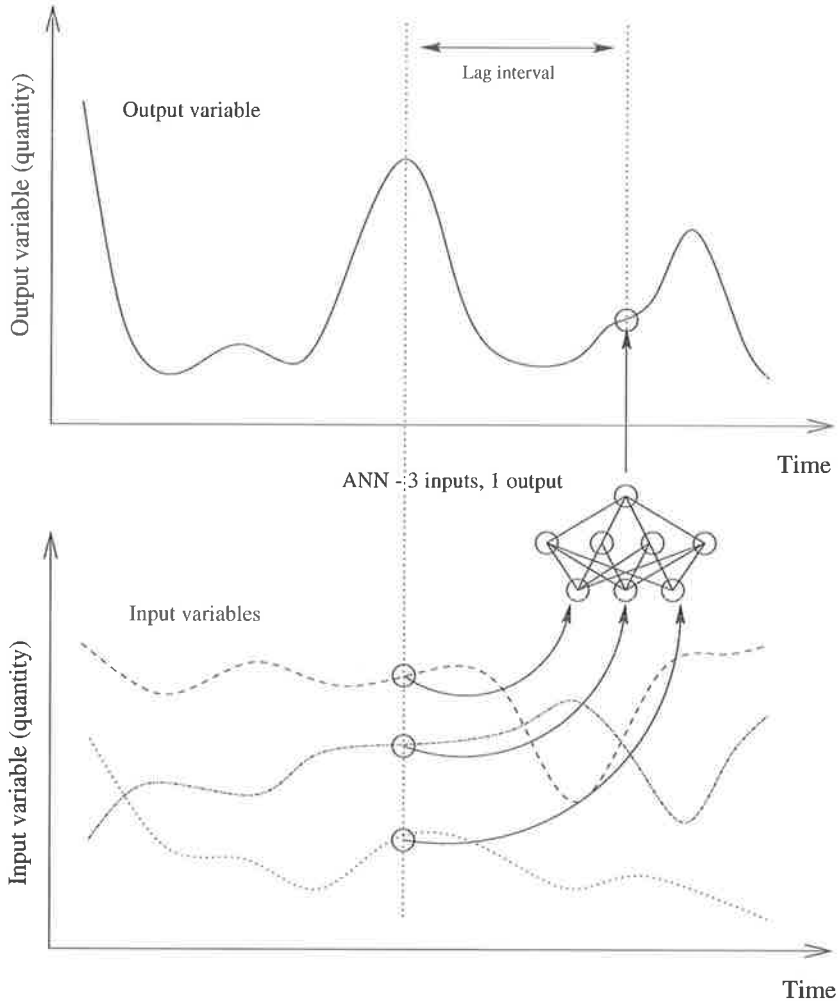

Figure 2.7: The TDNN structure with a single lag for all inputs.

1. Poor performance of ANN models as a result of being focused on irrelevant regions of the state space.
2. Prohibitive data requirements for discovery of the relevant regions of state space by the ANN.

According to Maier and Dandy (1997), these problems are amplified when training TDNNs since there is essentially no upper limit to the number of lags that may be included for each input variable. These authors demonstrated empirically that improvement in modelling outcomes can be achieved by reducing the problem dimensionality. Suggested approaches included guiding input selection by relevant domain knowledge, adding inputs in a stepwise fashion and implementing analytical techniques to discover input relevance *a-priori* (such as Haugh and Box (1977)).

Embrechts et al. (2001) devised a feature selection method called "data strip

mining". This approach compares the bootstrapped sensitivity of the candidate inputs with a dummy input drawn from a random distribution. Inputs that are less sensitive than the dummy inputs are deemed to be irrelevant to the model and are discarded. The process is iterated until only relevant inputs remain. These authors showed that data strip mining is capable of identifying 35 relevant inputs from a feature set of 300-1000 variables, leading to significantly improved model performance.

Olden and Jackson (2000) reviewed a variety of model selection methods commonly used for the purposes of identifying variables for multiple regression studies. These included forward selection, backwards elimination, stepwise selection, exhaustive search and bootstrapping. They used Monte-Carlo simulation to show that all model selection methods are biased in that they include irrelevant variables or exclude relevant variables. They found that the nature of bias varies according to the quantity of data available for model inference. It was found that where 60 or more records were available, a bootstrapping approach was the least biased approach to model selection in the context of multiple regression.

Recurrent neural networks (RNNs) (Pineda, 1987; Elman, 1990; Connors et al., 1994) have been proposed as a better approach to modelling sequences of data such as time-series. RNNs implement a feedback loop that uses hidden to output layer activations as additional inputs to the model. This provides the RNN with a hidden "temporal context" that is supposed to improve performance in the context of time-series data. Examples of time-series modelling applications using RNNs to predict phytoplankton dynamics include Jeong et al. (2001), Walter et al. (2001) and Jeong et al. (2003).

2.3.2 Step 2 – Model Approximation (Training)

This step is concerned with the issue of ANN learning, which is referred to as the *approximation problem* (Moody, 1991). The general principles of ANN learning by gradient descent based methods are described above in section 2.2.2. This section reviews some of the key decisions to be made when configuring learning algorithms and some of the causes for learning failure.

2.3.2.1 Numerical Conditioning

Analytic results by Saarinen et al. (1993) showed that learning by ANN training algorithms is inhibited by numerical ill-conditioning of the Jacobian matrix (ie the matrix of second derivatives of the prediction error with respect to the network weights and biases). It was concluded that "rank-deficiency" of the Jacobian causes the training algorithm to retrieve incomplete search information resulting in prolonged training. An outcome of this conclusion is that network configurations that do not solve a problem exactly are not necessarily parsimonious in

terms of their parameterisation. Thus they may exhibit redundancy as a result of the ill-conditioning of the problem. According to Van Der Smagt and Hirzinger (1998), the ill-conditioning is reflected by many saddle points and flat areas on the error landscapes.

Van Der Smagt and Hirzinger (1998) suggests that stochastic learning methods can overcome the problem of ill-conditioning. However, they concede that such methods are not well suited to problems where rapid training is required. Haykin (1994) made the following suggestions for overcoming ill-conditioning where gradient descent based training methods are used;

1. Every adjustable network parameter (ie weight) should have its own learning rate parameter.
2. Every learning rate parameter should be allowed to vary from one iteration from the next.
3. Similar δw signs on consecutive iterations should cause the learning rate to be increased.
4. Dissimilar δw signs on consecutive iterations should cause the learning rate to decrease.

Sarle (2001) states that preprocessing of input data may have a positive influence on the conditioning of Hessian and Jacobian matrices leading to superior learning performance – particularly where gradient descent based training algorithms such as backpropagation are utilised. It is suggested that standardising inputs, so that they have a mean of 0 and a standard deviation of 1, is beneficial to ANN learning. This is because, providing ANN connection weights are initialised to small random values, the hyperplanes described by the hidden layer units will effectively pass through the data cloud thus avoiding areas on the error surface that are not conducive to learning. Sarle (2001) concludes that Saarinen et al. (1993) overstated the effect of ill-conditioning on ANN learning because in their empirical investigations, they do not standardise input data.

2.3.2.2 Incremental vs Batch-Mode Training

Learning is termed as being either incremental or batch-mode, depending on the frequency of backward propagation of error derivatives. In incremental or on-line training, back propagation steps (and thus weight updates) occur with each presentation of a random vector from the training sample. On the other hand, batch or off-line training conducts the backward propagation step after presentation of the entire database of training vectors.

Haykin (1994) and Bertsekas and Tsitsiklis (1996) argue that incremental mode training has a number of advantages over batch mode training. The stochasticity introduced by random selection of training records makes the trajectory taken

through weight space variable. This has the benefit of making entrapment of training at local optima less likely. Also, incremental learning potentially has a much lower computational overhead on large training sets than batch-mode training leading to faster training (eg see results of Alpsan et al. (1995)). Despite these advantages, the stochasticity of incremental training means that, unlike batch mode methods, it is not conditionally guaranteed to converge at a minimum in the cost function (Gori and Tesi, 1992).

2.3.2.3 Training Meta-Parameters: Learning Rate and Momentum

The approximation characteristics of backpropagation are highly sensitive to the learning rate (ρ) and momentum (α) meta-parameters (Weiss and Kulikowski, 1991). Large ρ leads to rapid progression of learning towards the minimum of the cost function (Adeli and Hung, 1995). However, it also causes convergence on a "limit cycle" of oscillations around but not on the minimum (Bertsekas and Tsitsiklis, 1996). Small ρ on the other hand leads to a smoother more accurate trajectory through weight space and convergence on a limit cycle closer to the minimum. However if ρ is too small, training time may be excessively long (Haykin, 1994). Clearly there is a need to find a compromise value that enables reasonable training times, but is not subject to unstable, oscillatory behaviour. Most authors agree that a good value of ρ tends to be problem specific.

The performance of backpropagation is significantly improved by momentum. In momentum training, the current weight update is the sum of the calculated weight update and a proportion of the previous weight update as follows;

$$\Delta w_t = -\rho \frac{\partial E}{\partial w} + \alpha \Delta w_{t-1} \tag{2.10}$$

where α is the momentum coefficient. This approach (sometimes called the "heavy ball" method (Bertsekas and Tsitsiklis, 1996)) considers second order information about the error surface using a one step memory. This makes it computationally far cheaper than more complex algorithms, such as Newton's method, that need to compute matrices of second order derivatives (Alpsan et al., 1995). Momentum endows backpropagation with the following properties;

- Filtering of higher frequency variations of error surface thus reinforcing overall training direction (Gallant, 1993; McClelland and Rumelhart, 1988; Bertsekas and Tsitsiklis, 1996).
- Acceleration of learning through regions of the error surface that have similar slope.
- Stabilisation of oscillatory behaviour in learning (Haykin, 1994).
- Avoidance of shallow local optimum on the error surface (Haykin, 1994).

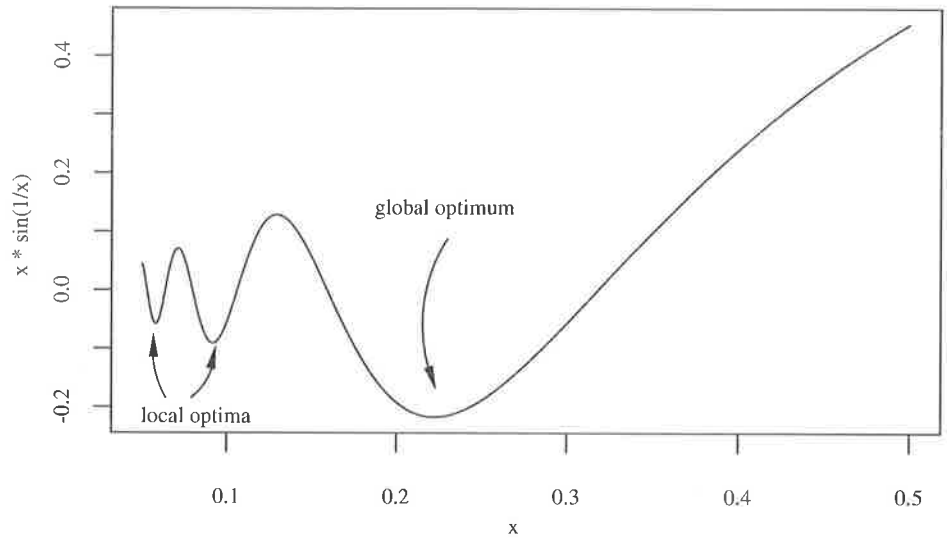

Figure 2.8: Example of global and local minima of $f(x) = \sin(1/x)$.

2.3.2.4 Local Minima

A potentially major source of training error is caused by the fact that gradient descent does not guarantee global optimisation of the cost function. Global optimisation in the context of ANN approximation can be defined as the problem of finding a value of the weights w^* such that the cost function $E(w^*)$ takes on the extreme minimum value. The problem is that the surface of the cost function $E(w)$ may contain local optima. w^* is defined as being a local optimum of E if $E(w^*) < E(w)$ for all w such that $||w^* - w|| \leq \varepsilon$ for some $\varepsilon > 0$ (Hassoun, 1995). Figure 2.8 displays an example where the aim is to find the value x which minimises $f(x) = \sin(1/x)$ where $x \in [0.05, 0.5]$. Assuming that a gradient descent based method is used to find the minimum, if $x(0) \in [0.05, 0.13]$, the minimum converged on will be one of the two local minima on the left hand side of figure 2.8 and not the global minimum in the centre.

According to Gallant (1993) and McClelland and Rumelhart (1988), local optima may present a greater problem to learning in low dimensional problems, since higher dimensional problems permit a greater chance of escape. Furthermore, Gori and Tesi (1992) point out that the linearity of the function underlying the data may also affect the propensity for entrapment at local optima, since linear functions do not have this problem.

Several approaches have been suggested for dealing with local optimum. Many authors point out that a certain level of stochasticity in the optimisation procedure can help "shake" the weight vector out of shallow "hollows" in the objective function. This can be achieved by the use of "global" optimisation procedures such as genetic algorithms or simulated annealing, in combination with "local" methods such as gradient descent (eg Masters (1994)). Haykin (1994) points

out that stochasticity introduced by incremental weight updates instead of batch updates during gradient descent can be helpful. Another approach is to train with many random starting conditions and allow the user to choose the best approximation. Momentum in training is able to carry training through shallow local optima (Haykin, 1994). Gallant (1993) suggests that sufficiently high dimensionality of the problem definition will make entrapment at bad optima unlikely.

2.3.3 Step 3 – Generalisation

The generalisation problem describes the task of finding a model estimation that achieves reasonable prediction accuracy on population data outside of the training sample. Moody (1991) defines the generalisation error as the expected error $E[y(x),\widehat{\mu}(x)]$ on new data taken over all possible training sets of size n and all possible test sets, where $\widehat{\mu}(x)$ is the model estimation. This author points out that all possible training sets should be considered in estimating E because $\widehat{\mu}$ is estimated from a finite, noisy, training sample, meaning that it is an implicit function of the random variables $\{\varepsilon^i; i = 1,\ldots,n\}$.

The generalisation performance of ANNs is known to be influenced by the relationship between the complexity of the mapping achieved and the number of training set records available. Theoretical studies suggest that generalisation is most likely when the quantity of training data, relative to the size or *complexity* of the ANN model, is large (Abu-Mostafa, 1989; Kung, 1993; Nejad and Gedeon, 1995). If this condition is not met, the ANN may exhibit a condition characterised by accurate mapping of training set data, but poor performance on independent population data. This behaviour, known as *overfitting*, is a result of high model variance caused by the large number of parameters available to fit the data (Geman et al., 1992). On the other hand, if there is insufficient ANN complexity for the approximation task at hand, the model will be *underfitted*, leading to consistent errors on training and test set data characteristic of prediction bias. Therefore, the task of maximising the generalisation ability of ANN models is one of finding the right compromise between too little or too much model complexity to optimise the bias/variance tradeoff. This problem, described by Moody (1991) as the *minimisation of prediction risk*, is illustrated in figure 2.9.

Traditionally the key to achieving this minimisation task has been to bias the ANN by limiting the model complexity somehow (Cannon and Whitfield, 2002). According to Prechelt (1998), there are two widely applied approaches to limiting model complexity to prevent overfitting;

1. Reduce the number of dimensions in the parameter space (ie reduce the number of connection weights).
2. Reduce the size of the dimensions in the parameter space (ie reduce the magnitudes of connection weights).

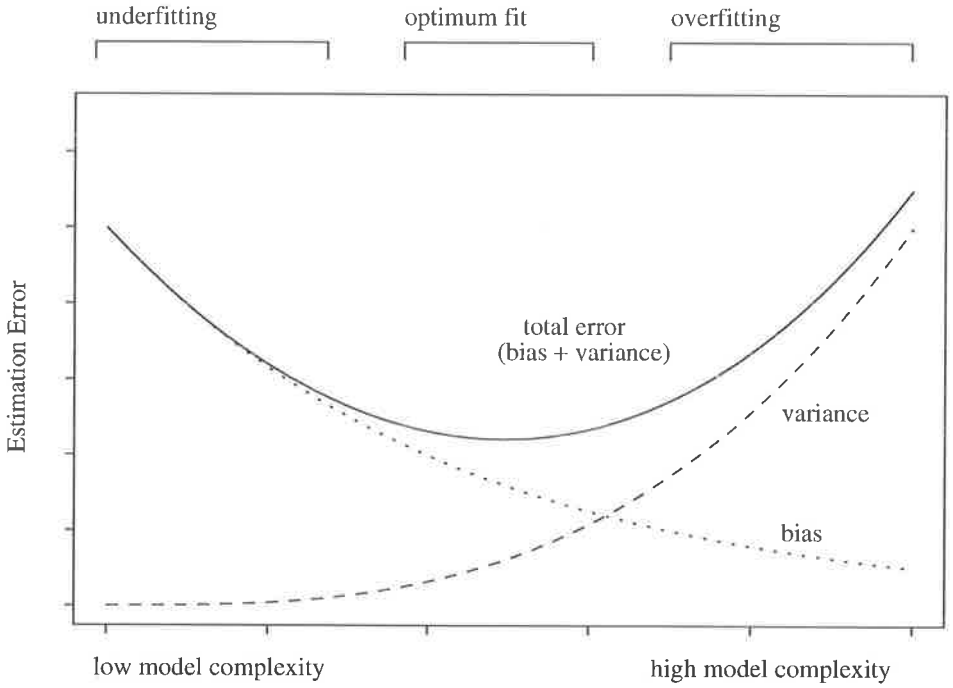

Figure 2.9: Minimisation of prediction risk (after Moody (1991)).

A number of approaches to achieving the first of these options have been applied. According to Prechelt (1998), these include greedy constructive learning (eg Fahlman and Lebiere (1990)), pruning (eg LeCun et al. (1990); Hassibi and Stork (1993); Levin et al. (1994)) and weight sharing (eg Nowlan and Hinton (1992)). Techniques for achieving the second option include regularisation by weight decay (eg Krogh and Hertz (1992); Weigend et al. (1991)) and early stopping of training (eg Morgan and Bourlard (1990)). Prechelt (1998) states that out of these techniques, early stopping is the most widely applied because it is well understood and relatively easy to implement. Finnoff et al. (1993) showed that early stopping produces superior outcomes in terms of model performance to other methods.

A key problem with these techniques is how to determine the optimum complexity for a given dataset. Analytical work that considers quantities such as the Vapnik-Chervonenkis (VC) dimension (Vapnik and Chervonenkis, 1971), the Akaike Information Criterion (AIC) (Akaike, 1969) or Bayesian prior probabilities, have been shown to be potentially useful. However, Moody (1991) points out that, in practice, their utility is limited by two problems; firstly, the calculations involved may be very difficult and secondly, they frequently give worst case bounds of network complexity that may be inappropriate for many applications. Practitioners are therefore generally left to determine complexity empirically by cross-validation; either by manual "off-line" determination or by some automatic "on-line" procedure (de la Maza, 1991).

The criteria for empirically determining optimum parameterisation is generally *ad-hoc* (Prechelt, 1998). Breiman (1996b) points out that in the context of limited data, selection of the optimum *regulariser* (ie, a model smoothed by limiting complexity) is subject to "predictive loss" arising from inaccuracies of cross-validation. Predictive loss is defined by these authors as the difference in error rate between the modeller's selection based on cross-validation and the "crystal ball" selection (ie where unlimited test data is available). It is highest for unstable inference methods, such as ANNs, where small changes in training sets and/or learning processes can cause large changes in models (Breiman, 1996b). To combat predictive loss and model instability in general, Breiman (1994) proposed bootstrap aggregation, or bagging (from ***b***ootstrap ***agg***regation). In bagging, the model representation is defined as an ensemble of perturbed predictors, where the aggregation is conducted by voting or averaging the outputs of the members of the ensemble. This proposal is summarised as follows;

> Given a learning set $\mathcal{L}$ consisting of data $\{(y_n, x_n), n = 1 \ldots N\}$ where y is a response variate to input x, we can form a predictor $\varphi(x, \mathcal{L})$. If we are given a sequence of learning sets $\{\mathcal{L}_k\}$, where each set is N independent observations form the same underlying distribution as $\mathcal{L}$, it is possible to obtain a better predictor by taking the average of $\varphi(x, \mathcal{L}_k)$ over k – ie $\varphi_A(x)$ where A denotes aggregation. Since most applications only have a single learning set $\mathcal{L}$ available without replication, a simulated aggregation can be made by taking repeated bootstrap samples $\{\mathcal{L}^B\}$ and forming $\{\varphi(x, \mathcal{L}^B)\}$ [1].

Breiman (1994) showed that the aggregated model φ_A is always better than φ in theory, although the amount of improvement depends on the extent of variation between individual models. For unstable classifiers, aggregation is shown to make large improvements, whereas the improvements for relatively stable inference methods are generally slight. According to Andersen et al. (2001), bagging works best when errors between variables and predictions are uncorrelated, which is most likely when using complex predictors. Cannon and Whitfield (2002) and Wilson and Recknagel (2001) showed that the stabilising effect of bagging on ANN learning reduces or eliminates the problem of increasing variance caused by overfitting. Thus, it can be hypothesised that bagging may even eliminate the need for a model selection procedure on the basis of complexity.

Cannon and Whitfield (2002) proposed that bagging promotes greater data efficiency by reducing the importance of cross-validation to determine stopping error or hidden layer configurations. Furthermore, it complements the "leave-one-out bootstrap" error estimator (Efron and Tibshirani, 1997), where the "out of bag" records not included in each bootstrap sample (on average, 37% of $\mathcal{L}$) are used as validation set data. Embrechts et al. (2001) points out that another advantage of

[1] It is assumed that the bootstrap distribution, taken by randomly sampling N samples from $\mathcal{L}$ *with replacement* approximates the distribution underlying $\mathcal{L}$. See Efron and Tibshirani (1993) for more information about bootstrap samples.

bagging over early stopping is that it overcomes the effects of model instability and thus variance resulting from different initialisation of network weights between runs. However, Breiman (1994) points out that the downside to the gains in model performance and data efficiency is the loss of a simple interpretable structure.

Empirical evidence from Lawrence and Giles (2000) and Alpsan et al. (1995) suggests that the training algorithm used may also have a significant impact on the generalisation performance achieved by an ANN. Lawrence and Giles (2000) compared the performance of backpropagation, conjugate gradients and a polynomial approximation method at undertaking a simple curve fitting exercise. They found that backpropagation was the only algorithm to achieve good performance over the entire function without overfitting. It was reasoned that backpropagation is biased towards smoother solutions because of difficulties in learning the larger connection weights required by relatively complex, discontinuous mappings. Alpsan et al. (1995) compared many different learning algorithms at learning a real-world medical problem and found that backpropagation trained in batch mode produced better generalising ANNs than any more advanced algorithm such as the "modified backpropagation algorithms" or second order training methods.

2.3.4 Step 4 – Model Validation

The empirical check of model performance is an important stage in the overall ANN model development process. It enables the modeller to estimate performance on newly sampled data and thus determine how well the ANN is generalised. It is emphasised by Weiss and Kulikowski (1991), Flexer (1995) and others that a *resubstitution* approach, whereby the model error rate is estimated as the performance of the ANN on the training set, is not an acceptable means of validation because it leads to an optimistically biased expectation of model performance. Therefore, the generally recommended approach is *cross-validation* (Stone, 1974), where the model error rate reported is calculated on an independent validation set sampled from the same distribution as the training set but held out from training.

According to Weiss and Kulikowski (1991) and Flexer (1995), when data is limited, splitting the sample into training and validation subsamples may result in suboptimal performance. This is because the data requirements for a statistically significant validation set may deplete training set representation to the point where model inference is impaired. These authors suggest the use of a so-called *rotation estimator*. This approach (often referred to as *leave-k-out* cross-validation) divides the data into k equally sized subsamples. Each subsample is used in turn as a validation set while the remainder are pooled as a training set.

Frequently, there is a need to empirically tune meta-parameters related to ANN learning to obtain optimum model performance. This is particularly the case

Table 2.1: Minimum requirements for evaluation of ANN model (after Flexer (1995))

1. Separate train and test set.
2. Computation of multiple runs to avoid random influences in training set composition, trajectory through weight space and weight initialisation.
3. A third independent validation set where parameter tuning is performed.
4. Reporting of mean, variance and confidence intervals of performance measures.
5. Computation of statistical tests (t-test) for performance comparison.

where "early stopping" of training is utilised to bias the ANN model, since there is a need in this case to empirically determine the stopping criteria Prechelt (1998). Flexer (1995) points out that such tuning necessitates the use of a third independent dataset to assess the effect of the tuning parameters, since the use of the validation set to choose between candidate models will make performance estimations optimistically biased. Mosteller and Tukey (1977) refer to this approach as *double cross-validation*, although many ANN practitioners refer to it as simply "cross-validation". Flexer (1995) further suggests that where meta-parameter tuning is carried out, statistical inference is necessary to verify that the observed effects on model performance are real and not due to random chance. Thus, conclusions about meta-parameter effects must be inferred statistically from a distribution of models.

In summary, Flexer (1995) proposes a list of minimum requirements for evaluation of ANN models, which are outlined in table 2.1.

2.3.5 Step 5 – Knowledge Discovery

Explanation of model predictions is generally seen as a priority by ecologists. A frequent criticism of ANNs is that they are a "black-box" approach to modelling, since the internal state of the model is generally considered to be hidden. However, a review by Olden and Jackson (2002) showed that a number of approaches have been developed to elucidate knowledge from trained ANNs in an ecological context. These authors state that since the contribution of each input variable depends on the connection weights within the ANN, analysis of these weights, either directly or indirectly, is the key to knowledge discovery from ANNs.

Sensitivity analysis of trained ANNs to quantitatively determine the effect of each input variable on the output is the most widely employed approach to knowledge discovery. Each input is varied in turn to determine its effect on the output variable while the remaining input variables are held or "blocked" at set values. According to Olden and Jackson (2002) a commonly employed form of sensitivity analysis is Lek's algorithm (Lek et al., 1996) whereby "response curves" are determined by varying each input across a number of intervals of that input's range. These authors implemented this approach while holding the blocked values at 20th, 40th, 60th and 80th percentiles in order to illustrate the interactions between input values.

Other variations of the sensitivity analysis procedure employed by ecologists include an approach by Sivonen and Jones (1999) where small quantities of white noise were added to the input of interest, while the remaining inputs where swept across the entire database. Schleiter et al. (1999) used a "senso-net" method employing an additional weight for each input neuron representing the relevance of that variable. These sensitivities were then adapted by the training process, enabling an effective online feature reduction system. Recknagel and Wilson (2000) used a scenario analysis technique that grouped input variables into relevant subsets (such as physical or chemical conditions). The effects of changes to these subsets on the output variable were then observed.

Statistical validation of the relationships observed by sensitivity analysis was first performed by Baxt and White (1995). In this application, the bootstrap sampling was used to generate 1000 perturbed training sets from which 1000 ANN models were trained. Sensitivity analysis of each bootstrap model created distributions of effects. Statistical testing was then used to determine the significance of the observed effects of inputs on the output variable. Embrechts et al. (2001) employed a similar approach utilising bootstrapping to allow statistical significance testing which is described above in section 2.3.1. Furthermore, this author extended the sensitivity analysis procedure to account for non-monotonic relationships within overall input sensitivity. Sensitivities of each input variable to multiple perturbations across the range were observed, with blocked values held at their median values (as per Lek's algorithm). The calculated sensitivity for a single input record was calculated from the output responses thus;

$$S_{obs} = (|R_{pos}|_{max} + (|R_{pos}|_{max} - |R_{1.0}|)) + (|R_{neg}|_{max} + (|R_{neg}|_{max} - |R_{-1.0}|)) \tag{2.11}$$

Figure 2.10 shows an example of how the approach outlined above in equation 2.11 accounts for non-monotonic relationships between input and output variables.

Neural Interpretation Diagrams, initially used in an ecological application by Özesmi and Özesmi (1999), enable a qualitative assessment of the effect of each

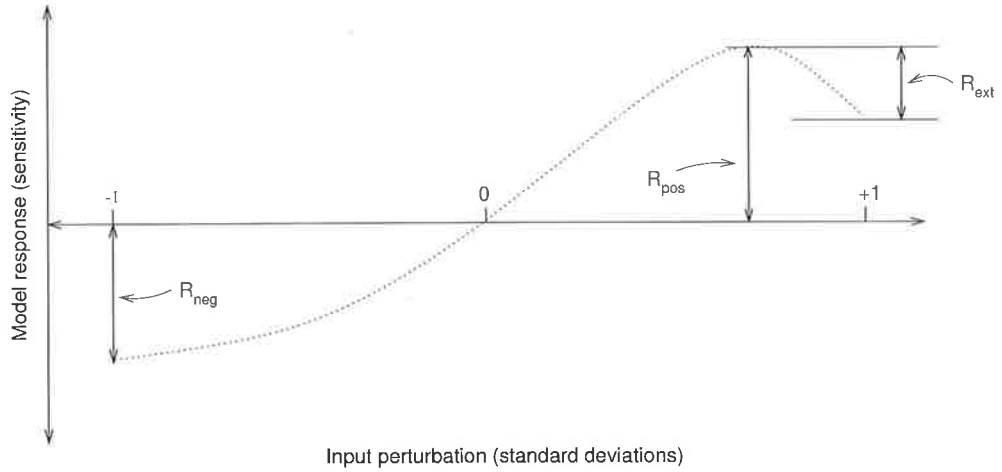

Figure 2.10: An example of total sensitivity calculation. $S = R_{neg} + R_{pos} + R_{ext}$.

input variable by representing the relative strengths and signs of connection weights visually. The weights of connections are represented by the pixel weights of lines and the sign of the connection by the colour of the line. An advantage of this approach is that it allows determination of interactions between different input variables through the investigation of the relative signs of input to hidden layer connection weights. However, Olden and Jackson (2002) points out that interpretation can be difficult in complex ANNs and that there is no way to differentiate between significant and insignificant connection weights. Garson's method (Garson, 1991) is an approach that numerically quantifies the relative contributions of input variables. While this approach is effective at determining overall contribution of input variables, it does not differentiate between excitory or inhibitory effects of variables.

Olden and Jackson (2002) proposed that a randomisation approach be used to improve aspects of both the NID interpretation and Garson's method. Specifically, it performs a significance test on the connection weights of a trained ANN, facilitating the interpretation of NID by eliminating non-significant weights, thus reducing network complexity. It determines whether the overall contribution of an input variable, as discovered by Garson's method, is statistically significant.

2.4 A Review of ANN Models of Eutrophication Variables

2.4.1 Introduction

As discussed in section 2.2, the properties of ANNs as universal approximators make them a very attractive method for modelling ecosystem properties. This is because ecosystems are known to be characterised by multivariate, non-linear processes that are difficult to handle using conventional deductive and inductive modelling approaches (Lek et al., 1996, 2000). According to Lek et al. (2000), Colasanti (1991) was the first to propose that ANNs might be a useful ecological modelling technique due to the similarities to ecosystems. French and Recknagel (1994) were early adopters of the technique, developing an ANN application for the prediction of algal blooms in Lake Saidenback, Germany. Since then there has been an increasing number of applications of ANNs and other machine learning methods for modelling all kinds of natural resource variables; including problem domains such as taxonomy, plant physiology, pollution assessment, forestry, weather forecasting, soil science, ornithology and others. Table 2.2 reviews the modelled variables and study sites of 22 applications of ANNs and other machine learning approaches to modelling eutrophication variables. These models share a number of attributes in that;

- They consider time-series data.
- They aim to make predictions or forecasts of phytoplankton biomass or productivity.
- They use supervised training methods – in other words, model approximation is guided by targets in the training data.

Most applications model variables in lakes and reservoirs; however 6 studies are applied to rivers (Jeong et al. (2001); Maier et al. (1998, 2000); Recknagel et al. (1997); Wilson and Recknagel (2001); Whitehead et al. (1997)) and 4 studies are applied to marine environments (Barciela et al. (1999); Sivonen and Jones (1999); Scardi (2001); Lee et al. (2003)). The following sections review how the studies presented in table 2.2 have undertaken the ANN modelling procedure outlined above in section 2.3

2.4.2 Model Design

Table 2.2 shows that nearly all applications consider algal biomass as an output variable; expressed either as an overall productivity indicator such as chlorophyll *a* or as cell counts of individual species, genera or functional groups. The most commonly modelled output variables are chlorophyll *a* concentration and cell

counts of species of cyanobacteria. By contrast, Sivonen and Jones (1999) and Scardi (2001) train models to predict primary productivity in terms of mg Cm^{-2} day^{-1} in marine environments.

Most applications listed in table 2.2 consider inputs known to have a deterministic link with the model outputs. Thus they consider various chemical, physical and biological conditions known to influence the processes of photosynthesis and trophic interactions. The following sections review the common variables used and how they may have a causative or correlative relationship with phytoplankton growth and species succession.

2.4.2.1 Inputs Describing Nutrient Availability and Chemical Properties

Nearly all models reviewed in table 2.2 consider concentrations of the macronutrient elements nitrogen and phosphorus. Nitrogen is expressed as concentrations of NO_2, NO_3, NH_4, total dissolved nitrogen or some combination of these species. Similarly, phosphorus is expressed as concentration of either PO_4, total dissolved phosphorus, total phosphorus or a combination thereof. These two elements are key plant macronutrients and have been well established to be frequent growth limiting factors in lakes and rivers. Measures of the availability of these two nutrients, particularly phosphorus, are widely used by simple empirical models for classifying the trophic state of waters (eg Vollenweider (1970), Vollenweider (1976), Sakamoto (1966) and Dillon and Rigler (1974)). Nitrogen is known to become a more important growth limiting factor in tropical lakes (Harris, 1986). Increasing overall nutrification is known to lift primary productivity and reduce the diversity of phytoplankton species observed in favour of dominance by cyanobacteria (Dokulil and Teubner, 2000; Zevenboom and Mur, 1980).

As well as the extent of nutrification, dynamics of the ratio of nitrogen to phosphorus are known to play a key role in determining the species dominance observed (Teubner et al., 1997; Takamura and Aizaki, 1991; Takamura et al., 1992). Low N:P ratios tend to favour cyanobacteria for a number of reasons including;

- Certain species of cyanobacteria fix atmospheric nitrogen, giving them a competitive advantage against non N-fixing phytoplankton. However, despite this fact, low N:P ratios tend to favour even non-fixing species of cyanobacteria over other algae (Dokulil and Teubner, 2000).
- Cyanobacteria have relatively high phosphorus requirements leading to a competitive disadvantage in high N:P conditions.
- High nitrogen levels tend to favour faster growing species such as green algae (Winder and Cheng, 1995).

The species of nitrogen may also influence the species dominance observed – for example, non N-fixing cyanobacteria have been shown to prefer NH_4, while N-fixing species prefer NO_3 (Dokulil and Teubner, 2000).

Table 2.2: ANN phytoplankton models – site and modelled variables

Reference	Site	Output(s)	Nutrients & Chemical properties	Physical properties	Biological inputs
Barciela et al. (1999)	Ria de Arousa (Spanish coast)	chl-a	N	temp, rad, mix depth, upwelling	chl-a
Bobbin and Recknagel (2001)	Lake Kasumigaura (Japan)	3 blue–green spp.	N, P, pH	temp, trans	
French and Recknagel (1994)	Lake Saidenbach (Germany)	Cyanophyceae, 3 green spp., Chlorophyceae, nanoplankton	N, P, pH, DO	temp, rad, trans	chl-a, zoo
Jeong et al. (2001)	Nakdong River (Korea)	chl-a	N, P, Si, pH	temp, trans, rad, flow, precip	zoo
Karul et al. (2000)	Keban, Mogan, Eymir Lakes (Turkey)	chl-a, tot. cell count, 3 blue–green spp.	N, P, pH, EC	temp, trans	zoo
Lee et al. (2003)	Hong Kong coast	Chlorophyll–*a*, *Skeletonema* spp.	N, P, DO	temp, rad, trans, wind, tide, precip	lag chl-a
Maier et al. (1998)	Murray River (Australia)	*Anabaena* spp	N, P, Fe	temp, turb, colour, flow	
Maier et al. (2000)	Murray River (Australia)	*Anabaena* spp.	N, P, Fe	temp, turb, colour, flow	
Olden (2000)	Grenadier Pond (Canada)	chl-a, phyto. community composition	N, P		zoo
Recknagel et al. (1997)	Lake Kasumigaura, Lake Biwa (Japan), Darling River (Australia), Lake Tuusulanjärvi (Finland)	blue–green spp., functional groups	N, P, Si, pH, DO, HCO_3, EC	temp, trans, colour, flow, strat, wind, cloud, depth	zoo
Recknagel and Wilson (2000)	Lake Kasumigaura (Japan)	3 blue–green spp.	N, P	temp, trans, rad, depth	chl-a, zoo
Recknagel et al. (1998)	Lake Kasumigaura (Japan)	5 blue–green spp.	N, P	temp, trans, rad, depth	chl-a, zoo
Scardi and Harding Jr (1999)	Chesapeake Bay (USA)	Primary productivity	salinity	temp, rad, euph. depth, station depth, lat, long, ext. coeff	chl-a
Scardi (2001)	Western Mediterranean, East Pacific	Primary productivity		temp, rad, depth, lat, long, date, day length	chl-a
Todorovski et al. (1998)	Lake Glumsoe (Denmark)	Primary productivity	N, P	temp	zoo
Whitehead et al. (1997)	River Thames (UK)	Chlorophyll–*a*		temp, rad, flow	upstream chl-a
Wilson and Recknagel (1997)	Lake Kasumigaura (Japan)	8 blue–green spp.	N, P	temp, rad, trans, depth	chl-a, zoo
Whigham and Recknagel (2001)	Lake Kasumigaura (Japan)	chl-a	N, P, pH, DO	temp, trans	zoo
Wilson and Recknagel (2001)	Lake Biwa, Lake Kasumigaura (Japan), Burrinjuck Dam, Darling River, Myponga Dam (Australia), Lake Soyang (Korea)	chl-a	N, P	temp, trans	chl-a
Walter et al. (2001)	Burrinjuck Dam (Australia)	chl-a	N, P	temp, rad, depth, volume, surface area	
Wei et al. (2001)	Lake Kasumigaura (Japan)	4 blue–green spp.	N, P, pH, DO, COD	temp, turb,	zoo
Yabunaka et al. (1997)	Lake Kasumigaura (Japan)	chl-a, 5 blue–green spp.	N, P, Si, pH, DO	temp, trans	zoo

Five applications listed in table 2.2 (Jeong et al. (2001); Maier et al. (1998, 2000); Recknagel et al. (1997); Yabunaka et al. (1997)) consider one or both of the micronutrients silica and iron. Available silica is an important nutrient for the growth of diatoms. Also cyanobacteria are known to have a higher demand for trace elements in general (Dokulil and Teubner, 2000). Many models also consider dissolved oxygen level, which may be be an indirect driving variable since anoxic conditions lead to phosphorus release from sediments (Trimbee and Prepas, 1988). Since pH determines availability of dissolved CO_2 in the water column, this too can play a role in species dominance. It has been shown that cyanobacteria compete well in environments with relatively high pH and resultant low carbon availability (Dokulil and Teubner, 2000).

Sivonen and Jones (1999), Scardi (2001) and Whitehead et al. (1997) do not consider any nutrient data in their models.

2.4.2.2 Inputs Describing Physical Conditions

Temperature, considered by all models reviewed, is a key driving variable as it determines rates of chemical and biological processes. Different species of phytoplankton have varying temperature optimums, with higher temperatures tending to favour cyanobacterial growth (Dokulil and Teubner, 2000; Robarts and Zohary, 1987). However, temperature alone does not determine dominance – there is generally a complex interaction with other conditions (Robarts and Zohary, 1987). For example Takamura and Aizaki (1991) and Takamura et al. (1992) report a succession in Lake Kasumigaura, Japan, from *Microcystis* spp to *Oscillatoria* spp. dominance arising from an interaction between temperature and nutrient availability. This succesion caused a drop in overall primary productivity.

Temperature can also affect other physical conditions such as the mixing regime. Thermal stratification can act as a physical barrier separating regions of high light/low nutrient conditions from regions of high nutrient/low light conditions. The presence of stratification tends to favour species of cyanobacteria adapted, through buoyancy control, to overcoming the physical separation of light and nutrients (Ganf and Oliver, 1982; Reynolds, 1987). Recknagel et al. (1997) represents the presence of thermal stratification to the model explicitly, while Jeong et al. (2001), Lee et al. (2003), Whitehead et al. (1997), Maier et al. (1998) and Maier et al. (2000) included variables that may affect the mixing regime such as wind or flow rate in the case of rivers.

All models also considered one or more variables indicating the level of light available for photosynthesis – incident solar radiation, cloud cover, secchi-disc depth, turbidity and colour. Light availability is a key driving variable for photosynthesis and light attenuation can have a significant impact on overall primary productivity (Ruley and Rusch, 2002). Light intensity may affect species dominance. For example, cyanobacteria are able to harvest low light intensities of

wavelengths unusable by other species making them highly competitive in low light conditions (Mur et al., 1999). Light availability has also been shown to interact with the level of nutrification in determining species succession. Zevenboom and Mur (1980) showed that non N-fixing *Oscillatoria* spp. dominates over N-fixing types in severely hypertrophic lakes even when growth is N-limited as a result of superior low light efficiency.

Some applications reviewed, such as Walter et al. (2001), use morphometric information such as depth, surface area and volume as inputs. This information has been shown to have a deterministic relationship with eutrophication and species dominance. Shallow lakes have been observed to favour dominance by filamentous cyanobacteria, while deeper lakes favour colony forming types (Schreurs, 1992) (cited by Ruley and Rusch (2002)). Recruitment of cyanobacteria is assumed to decrease as lake depth increases due to the reduction of the sediment area/volume ratio (Trimbee and Prepas, 1988). Fetch, combined with wind speed, has an effect on the mixing conditions and thus the degree of thermal stratification within a water body.

In addition to inputs with an established deterministic relationship with photosynthetic or ecological processes, Scardi (2001) used "co-predictor" variables as inputs – that is, variables known to be correlated to dynamics of the dependent variable, but not necessary causative. It is argued that since ANNs are relatively robust with regards to redundant inputs, extra correlative information provide a low risk means of improve model prediction accuracy. In this study and in Sivonen and Jones (1999), information regarding the latitude and longitude of measuring stations, date and day length were found to be helpful for model predictions.

2.4.2.3 Inputs Describing Biological Factors

11 studies consider zooplankton abundance in some form and 8 studies include overall algal abundance expressed as chlorophyll *a* in the input layer. Zooplankton can impose a top-down control of algal biomass through grazing and may have an impact on species dominance, since certain species of cyanobacteria limit their grazing mortality due to adaptations to make them inedible (Dokulil and Teubner, 2000) giving them a competitive advantage. Chlorophyll *a* has an impact on light availability, with higher concentrations reducing light availability through the shading properties of algal cells. Also it may influence the chemical properties in terms of nutrient availability (as a result of consumption) or pH.

2.4.2.4 Modelling time-series Interactions

Table 2.3 reviews the length and sampling interval of databases used, model handling of time and the division of data into training and testing sets. It can be seen that the lifespan of the time-series used for modelling varies considerably, from

2 months for Todorovski et al. (1998) to 18 years for one case study outlined in Whigham and Recknagel (2001). Most applications have between 5 and 10 years of data available. In most cases, the sampling intervals in the time-series used is in the order of weeks, although it varies from 4-5 days (Todorovski et al., 1998) to 1 month (Wei et al., 2001). It can be seen that the model time step is generally more frequent than the actual sampling frequency of the databases, with daily time steps being most commonly used. This increase in sampling frequency is achieved by interpolation of the modelled variables between the actual sample dates. The mode of interpolation was not described by any applications except for Yabunaka et al. (1997), who use linear interpolation and Todorovski et al. (1998), who use predictions by domain experts.

Out of the 22 applications reviewed, 9 trained models to make forecasts of future phytoplankton abundance given current environmental values, while the remaining models were trained to predict phytoplankton abundance on the same observation date as the input variables. Where a forecasting structure was implemented, a TDNN structure was applied whereby the input variables lagged the outputs by the required forecast period. This lag period was generally between 1 and 4 weeks which is consistent with the real sampling frequency of the time-series used. In some studies, such as Maier et al. (1998), Recknagel et al. (1998) and Recknagel and Wilson (2000), multiple lags of a single variable were used.

Jeong et al. (2001) and Walter et al. (2001) used a recurrent network structure (RNN) to further extend the time dynamic behaviour of the models. The RNN copies hidden to output layer activations at time $t-1$ and uses them as inputs for time t[2]. These authors state that the recurrent ANN paradigm offers a superior structure for time-series modelling, as the activations of the recurrent connections represent the model state at previous time steps in a manner comparable to many deterministic modelling approaches.

In contrast with the majority of studies reviewed, Sivonen and Jones (1999), Scardi (2001), Karul et al. (2000) and Whitehead et al. (1997) did not present the model predictions as a time-series since there was no intention in these cases to display the model's handling of dynamics in phytoplankton abundance over time. This differentiates these applications from the remaining studies where the models were explicitly trained and validated to exhibit time-dynamic behaviour.

2.4.3 Model Inference

Table 2.4 outlines technical aspects of the reviewed models including the model structure, approximation method (ie training algorithm), *a-priori* bias, approach to complexity tuning, elucidation method and controls used. The following sections provides further description of each of these aspects of the reviewed studies.

[2]See Pineda (1987); Elman (1990); Connors et al. (1994) for more complete descriptions of the methodology regarding recurrent ANNs.

Table 2.3: ANN eutrophication models – time series information, train/test set partitioning and forecast interval

Reference	Time series length	Sampling interval	Model time step	Forecast interval	Train/Test set
Barciela et al. (1999)	3 years	1 week	1 day, 1 week, season	0	3/3 years
Bobbin and Recknagel (2001)	11 years	2–4 weeks	2 – 4 weeks	0	9/2 years
French and Recknagel (1994)	5 years	7–10 days	1 day	1 day	3/2 years
Jeong et al. (2001)	5 years	1 week	1 day	0–40 days	5/1 years
Karul et al. (2000)	4 years	?	?	0	?/?/? years
Maier et al. (1998)	7 years	1 week	1 week	1–4 weeks	6/1 years
Maier et al. (2000)	7 years	weekly	weekly	1–4 weeks	6/1 years
Lee et al. (2003)	4, 18 years	1, 2 weeks	1 day & 1, 2 weeks	1–15 days	10/8 years
Olden (2000)	1 year	2 weeks	2 weeks	0, 2 weeks	?/?
Recknagel et al. (1997)	8–12 years	1–4 weeks	daily	0	6–10/2 years
Recknagel and Wilson (2000)	10 years	2–4 weeks	daily	0	8/2 years
Recknagel et al. (1998)	10 years	2–4 weeks	daily	0	8/2 years
Scardi and Harding Jr (1999)	12 years	?	?	0	100/226 records
Scardi (2001)	5, 7 years	1 day	1 day	0	1261/630/631 records
Todorovski et al. (1998)	2 months	4–5 days	0.1 day	1, 2, 5 days	2/2 months
Whitehead et al. (1997)	3 years	1 week	1 week	0	80%/20%
Wilson and Recknagel (1997)	10 years	2–4 weeks	daily	0	10/10 years
Whigham and Recknagel (2001)	10 years	2–4 weeks	daily	0	8/2 years
Wilson and Recknagel (2001)	8–18 years	2–4 weeks	1 month	0, 1 month	8–18/8–18 years
Walter et al. (2001)	18 years	1–4 weeks	1 day	1 week	15/3 years
Wei et al. (2001)	15 years	1 month	1 month	0	10/5 years
Yabunaka et al. (1997)	12 years	2–4 weeks	1 day	1 week	11/1 years

2.4.3.1 Approximation

Table 2.4 shows that most studies utilise MLP structures trained with the backpropagation algorithm. As mentioned above, Jeong et al. (2001); Walter et al. (2001) augment the traditional MLP approach with recurrent connections to better model time dynamics. Three studies utilised a second order training method instead of backpropagation – Wilson and Recknagel (1997) and Wilson and Recknagel (2001) employ the conjugate gradient training method and Karul et al. (2000) use a Levenberg-Marquardt algorithm.

Several more recent studies use alternative model representations and approximation methods. Bobbin and Recknagel (2001) apply principles of evolutionary computation to evolve a ruleset from data to predict algal biomass. Similarly, Whigham and Recknagel (2001) utilises the same principles to evolve equations to achieve the same task. Todorovski et al. (1998) uses a similar approach to equation discovery as Whigham and Recknagel (2001), the difference being that instead of using genetic algorithms for the training process, a non-linear optimisation technique is used. Maier et al. (2000) uses a B-spline associative network trained using a linear optimisation method. In each of these cases, the aim is to achieve a more transparent representation of knowledge learned from the training data than is possible with ANNs within a *model-free* approximation framework. These approaches have been developed in response to the common criticism of ANNs that they are an opaque means of representing knowledge compared to traditional deductive and inductive modelling approaches.

2.4.3.2 Generalisation

A few of the models reviewed impose an *a-priori* bias to somehow restrict the range of models that may be discovered during training to those that make sense from the point of view of domain experts. Scardi (2001) achieves this by means of a "constrained training" approach, whereby an error penalty is applied during training if the model approximates undesired solutions. In this case, the model was biased towards primary productivity response surfaces that had 1 maximum and 4 minima with respect to irradiance and biomass values. It is claimed that such a bias has the effect of restricting the complexity of the trained model thus reducing overfitting and ensuring that the model retained a degree of "biological soundness". Todorovski et al. (1998) and Whigham and Recknagel (2001) imposed a "declarative language bias" on their equation discovery models that restricted the model terms and grammar that could be used to those that made ecological sense. In addition, Todorovski et al. (1998) further biased the model by using synthetic data predicted by domain experts for training.

All models reviewed, apart from Bobbin and Recknagel (2001) and Lee et al. (2003), used some means of tuning the model complexity to prevent overfitting. The most common approach for models based on MLPs was through empirical

Table 2.4: ANN eutrophication models – methodology - model inference, tuning, elucidation and controls

Reference	Structure	Approx. method	*a–priori* model bias	Complexity tuning	Elucidation method	Controls
Barciela et al. (1999)	MLP	backprop	none	hidden layer size	none	deterministic model
Bobbin and Recknagel (2001)	Ruleset	GA	none	none	Examination of learned ruleset	none
French and Recknagel (1994)	MLP	backprop	none	hidden layer size	none	none
Jeong et al. (2001)	recurrent ANN	backprop	none	hidden layer size	sensitivity curves	none
Karul et al. (2000)	MLP	Levenberg–Marquardt	none	early stopping of training	none	
Lee et al. (2003)	MLP	Backprop	none	none	sensitivity analysis, weight interpretation	
Maier et al. (1998)	MLP	Backprop	none	hidden layer size	sensitivity analysis	none
Maier et al. (2000)	B–spline AMN	LMS	none	no. basis functions	fuzzy interpretation of B–spline basis functions	backprop ANN
Olden (2000)	MLP	Backprop	none	hidden layer size, connection pruning	neural interpretation diagrams	none
Recknagel et al. (1997)	MLP	Backprop	none	hidden layer size	sensitivity analysis	empirical, deterministic, time–series, heuristic & fuzzy models
Recknagel and Wilson (2000)	MLP	Backprop	none	hidden layer size	sensitivity, scenario analysis	none
Recknagel et al. (1998)	MLP	Backprop	none	hidden layer size	none	none
Scardi and Harding Jr (1999)	MLP	Backprop	constrained training	jitter, early stopping, hidden layer size	sensitivity analysis	none
Scardi (2001)	MLP	Backprop	constrained training	jitter, early stopping, hidden layer size	sensitivity surfaces	none
Todorovski et al. (1998)	Equations	Levenberg–Marquardt	language bias, data synthesis	equation length	analysis of discovered equations	naive models, linear model
Whitehead et al. (1997)	MLP	Backprop	none	training time	extraction of equation	time series analysis, dynamic mass balance model
Wilson and Recknagel (1997)	MLP	Conjugate Grad.	none	hidden layer size	none	none
Whigham and Recknagel (2001)	Equations	GA	language bias	depth of program tree	analysis of discovered equations	none
Wilson and Recknagel (2001)	MLP	SCG	none	hidden layer size, early stopping	none	perceptron
Walter et al. (2001)	recurrent MLP	Backprop	none	early stopping	sensitivity curves	deterministic model
Wei et al. (2001)	MLP	Backprop	none	hidden layer size	sensitivity analysis	none
Yabunaka et al. (1997)	MLP	Backprop	none	hidden layer size	sensitivity analysis	none

determination of the optimum number of hidden layer units by cross-validation. In addition to this measure, Sivonen and Jones (1999), Scardi (2001) and Wilson and Recknagel (2001) stopped training early as a further control against overfitting. Karul et al. (2000) and Walter et al. (2001), by contrast, only used early stopping and did not perform any optimisation of hidden layer size. Olden (2000) used an online connection pruning approach in addition to hidden layer optimisation, where connection weights below a threshold value were pruned from the network. Sivonen and Jones (1999) and Scardi (2001) added a small Gaussian noise component to the input data at each training epoch with $\mu = 0$ and $\sigma = 0.01$ in addition to hidden layer tuning and early stopping. This perturbation, known as "jitter"[3], is claimed to smooth or regularise function approximations by ANNs leading to superior generalisation characteristics.

With respect to the alternative model structures reviewed, Maier et al. (2000) determined the optimum number of basis functions to include in the AMN. Todorovski et al. (1998) and Whigham and Recknagel (2001) defined a maximum equation length to prevent the model getting too complex and thus overfitting training data.

2.4.4 Validation

The final column in table 2.3 shows the division of data between training and validation sets in terms of time. Most applications used 1-3 years of data for validation and the remainder for training, with the training and validation data being retained as discrete blocks in terms of time (such as years). In general, this meant that the majority of the data (80% or more) was used for training. Karul et al. (2000) and Scardi (2001) also used a third "tuning" dataset which was used to tune the training time of the MLP (this is denoted in table 2.3 as train/tune/test rather than train/test). Wilson and Recknagel (1997, 2001) used all the data for training and validation by means of 10-fold-crossvalidation and the leave-one-out bootstrap estimator respectively.

The final column in table 2.4 shows whether the machine learning model in each case was compared to a conventional modelling approach. It can be seen that in 6 of the 22 studies reviewed, some reference was made to other model types, whether by direct comparison or through discussion of the calibre of performance observed.

Barciela et al. (1999) found that MLPs were capable of more accurate predictions of marine primary productivity than a deterministic model, particularly at short time scales. It was reasoned by these authors that the performance of the deterministic model is hampered by difficulties associated with estimating some parameters. Maier et al. (2000) compared the prediction accuracy of B-spline

[3]See Györgyi (1990) for more background to this technique.

AMN networks with conventional MLP networks for forecasting *Anabaena* spp. biomass in the Murray river. It was concluded that the AMN networks performed slightly better than MLPs and had the additional feature of providing a more transparent knowledge representation. Recknagel et al. (1997) argued that MLPs are capable of superior prediction accuracy than existing empirical, deterministic, time-series, heuristic and fuzzy ruleset models due to their ability to resolve to species level compared to chlorophyll *a* or functional groups and their ability to resolve timing of growth to day or weeks compared to months, seasons or years.

Todorovski et al. (1998) compared the models developed by means of equation discovery system LAGRAMGE with a linear model form and 2 "naive" models – "no-change" and "same-change". The no-change model predicts that the value of the output for the next time step will be the same value as for the current time step ($\hat{\text{phyt}}(t+h)$=phyt(t)). The same-change model predicts that the change in the output will be the same as the change from the previous time step ($\hat{\text{phyt}}(t+h)$-phyt(t)=phyt(t)-phyt($t-h$)). It was found that the LAGRAMGE model had superior prediction accuracy to the linear model and the no-change model. The same-change model performed better than LAGRAMGE at small prediction intervals, but was not as robust in that it did not perform as well when the prediction interval was increased. Similarly, Wilson and Recknagel (2001) compared the performance of MLP models with perceptron models to determine the importance of the ability of MLPs to map non-linear decision boundaries. It was found in a study involving forecasting chlorophyll *a* abundance in 5 lakes and 1 river that the MLPs generally performed marginally better than the perceptron models, although in 2 instances the perceptron model performed best.

Whitehead et al. (1997) compared the MLP model with both conventional time-series analysis and a deterministic model for predicting chlorophyll *a* biomass in the River Thames. These authors found that the MLP model performed very similarly to the conventional modelling approach, although they commented that the MLPs had the advantage that no "subjective information is required to determine the model structure or estimate parameters". However, it was noted that, since ANNs do not explicitly represent processes, interpretation of processes could only be made at the most general level compared to conventional modelling approaches. Thus, as is the general consensus regarding ANNs, these authors found that ANNs are most useful in situations where analysis of large datasets is required without *a-priori* knowledge.

2.4.5 Knowledge Discovery

Many studies reviewed employed some technique to elucidating knowledge from trained models. The most common approach for the MLP models was sensitivity analysis (see section 2.3.5). Jeong et al. (2001), Scardi (2001) and Walter et al. (2001) further enhance the information retrieved by reporting the sensitivity

"surfaces" or "curves" retrieved with respect to the output when the input or inputs of interest are varied over their entire range (ie "Lek's algorithm") enabling elucidation of non-linear relationships. Recknagel and Wilson (2000) applied a scenario analysis by means of observing the effect on the output by modifying input variables in related groups. In this case, the effect on the output of 4 separate scenarios was observed where data was swapped between the 2 validation years for the following groups of inputs – i) nutrients, ii) zooplankton, iii) physical data, iv) chlorophyll *a*. This enabled the elucidation of the causes of a species succession that occurred between the two validation years.

Several authors used a more direct approach to knowledge discovery than sensitivity analysis by interpreting ANN connection weights. Olden (2000) used neural interpretation diagrams (NID) to visualise the strongest connection weights and thus the most important driving variables. Similarly, Lee et al. (2003) used a method to interpret the connection weights to elucidate the driving variables. Whitehead et al. (1997) used a method to extract an equation from the connection weights, although the exact method of interpretation was not outlined in this case.

Where novel forms of model approximation such as GA or equation discovery were used, knowledge discovery was facilitated by the transparent nature of knowledge representation learned. In the case of the models developed by Bobbin and Recknagel (2001), Whigham and Recknagel (2001) and Todorovski et al. (1998), elucidation was simply a matter of interpreting either the rulesets or equation sets discovered by training. Maier et al. (2000) used an interpretation method to gain a fuzzy ruleset from the trained associative network.

2.4.6 Discussion and Conclusions

2.4.6.1 Choice of Input Variables

It is widely recognised that none of the driving variables considered by applications of ANNs and machine learning to modelling eutrophication variables can determine primary productivity or species dominance alone. A number of studies have illustrated complex interactions between nutrient levels, light availability, lake morphometry and temperature with respect to the spectrum and abundance of algal species favoured (Ruley and Rusch, 2002). It is not surprising, therefore, that the models reviewed in table 2.2 generally utilise input layers that are highly multivariate considering input variables from each of the 3 classes specified. Indeed, this is the type of modelling that ANNs are well suited to, since they can handle multivariate modelling tasks where the relationships are complex and unknown.

However, there are several conflicts in approaches to model design in evidence amongst the reviewed papers. Two studies, (Maier et al., 1998; Lee et al., 2003), stress the need to reduce the risk of redundant inputs being included in the model. This is reasonable since Aussem and Hill (1999), Aoki et al. (1999) and others

point out that noise introduced by redundant inputs can reduce performance and increase the likelihood of overfitting. Scardi (2001), on the other hand, states that the robust nature of ANNs with respect to input redundancy enables a wider range of inputs to be considered than was previously possible. As discussed in section 2.3.1, there is a need to achieve the right balance between the "curse of dimensionality" on the one hand and the flexibility to include novel input variables on the other. The contradictions apparent in the literature with regards to this issue suggests there is a need to determine the practical importance of this tradeoff. Furthermore, with the issue of database compatibility in mind, a relevant question is whether a generalised, or "generic" set of highly available inputs can be identified that can be used to create models that have similar performance to *ad-hoc*, database specific architectures.

In terms of the choice of input variables, the review showed that no models explicitly considered the possibility of interspecific competition through inclusion of species cell counts in input layers. This is in spite of the fact that many studies reviewed make predictions of species abundances. Therefore, the question arises whether species succession in freshwater ecosystems is driven primarily by environmental conditions, interspecific competition, or both? Also, it is known that there may be considerable spatial variability in algal density in water bodies depending on the extent of stratification and the effects of wind. This is particularly the case with respect to cyanobacterial blooms which tend to be concentrated on the surface due to overbuoyancy (Reynolds, 1987) and therefore highly vulnerable to the effects of wind. Yet no studies reviewed explicitly included information regarding spatial variability of algal biomass in the structure of the model, or made mention of this factor as a possible influence on the observed data.

Given this review, the following can be concluded;

- The problem of input selection needs further investigation to determine how robust ANNs are with regards to redundant input variables.
- There is a need to identify "generic" models compatible with a wide range of databases, as well as application specific models.
- There is a need to investigate the importance of input variables describing spatial variability and/or competition between species or groups of phytoplankton.

2.4.6.2 Modelling Time Series

Table 2.3 shows that about 50% of the models reviewed do not explicitly represent links between past and present states in the model design, either by means of lag inputs or by recurrent network connections. This is in spite of the fact that nearly all models are trained using time-series data and are evaluated in terms of their ability to handle dynamics in algal biomass over time. Lee et al. (2003) points

out that models not designed to make forecasts are not useful in a management sense, since they predict algal biomass for the same time as the input variables are measured. It can also be argued that, unless a time component is explicitly considered, a model's use as an elucidatory tool is compromised by a lack of clarity with regards to the direction of causality being modelled. For example, it is possible that a "same day prediction" ANN model has learned to predict algal biomass from the consequences of high primary productivity, such as nutrient consumption and low secchi disk depth, rather than the causes.

Many studies reviewed enforce a higher frequency model time step than the actual sampling frequency by means of interpolation. In some cases it is stated that interpolation creates a regular time step necessary for compatibility between the dataset and TDNN and/or RNN structures. However, Lee et al. (2003) argues that using interpolated data to train a TDNN model may cause a blurring of past and future conditions giving the model access to information that, when applied to new data, would not be available. Figure 2.11 illustrates the potential problem – it can be seen that on the right hand side of the diagram where the lag interval falls below the actual sample intervals, the input data is calculated from observations that are *in the future* relative to the model output. This "temporal contamination" is a particular problem where the model considers an autoregressive component, that is, where one of the input variables is the output variable with a time lag imposed, since, in such a case, it means that the input data has been derived from the same real values in the time-series as the output (Lee et al., 2003). These authors provide an elegant demonstration of the pitfalls of this approach by showing that an ANN with lagged inputs can model a series of interpolated random numbers very accurately, despite there being no model underlying the real, uninterpolated values.

Logically, the use of interpolation to enforce dataset compatibility with time-series modelling structures such as time delay or recurrent connections raises a number of other problem issues including;

- The method of interpolation used (linear, splines etc) will affect the model learned by the ANN. As yet, there are no documented results known to the author that empirically differentiate between interpolation methods for an ANN application modelling natural resource variables.
- It adds significantly to the overall data processing task, since it necessitates extensive, error prone preprocessing. Also, the inflation of dataset sizes by up to 10-30 times their original record count increases training times.
- The assumptions made about the dynamics of variables between sample dates may be incorrect – particularly for highly dynamic variables such as phytoplankton abundance.

Given the issues raised by this discussion, it can be concluded that more research is needed to develop a time-delay model representation that is compatible with the

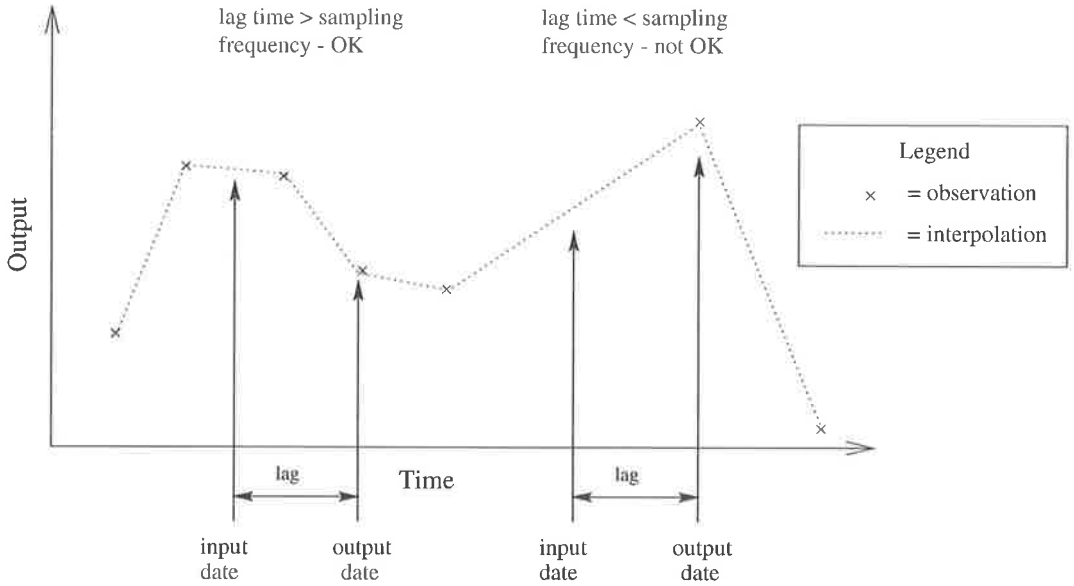

Figure 2.11: Use of lag inputs with interpolated data.

sampling structure of typical limnological datasets. Such a representation needs to be characterised by the following qualities;

- It must reduce the probability of "temporal contamination" when used with time delay or recurrent connections.
- It must, as far as possible, avoid assumptions or simplifications regarding dynamics between sample dates.

2.4.6.3 Approximation and Generalisation

This review shows that a number of innovations in ANN and machine learning techniques are being applied to the eutrophication modelling problem in recent years. These include the use of time delay and recurrent network connections to represent prior model states, alternative model representations and approximation methods to either speed up learning or enable the use of more transparent knowledge representations and the use of *a-priori* bias to guide model approximations. These innovations are motivated by recognition of the following requirements;

- The desire to develop models that are sound from an ecological perspective.
- The desire to "peak into the black box" to gain a better explanation of the system being modelled.
- The desire to raise the efficiency and robustness of the model approximation process.

The review of applications clearly shows that practitioners have few problems obtaining reasonable model approximation in terms of mapping training data despite the theoretical difficulties with ill-conditioning and local optima outlined in section 2.3.2. Conversely, achieving reasonable generalisation evidently requires greater attention since nearly every application reviewed biased modelling outcomes by limiting complexity to prevent overfitting. With respect to ANNs, combinations of up to 4 meta-parameters determining complexity are being tuned, including hidden layer size, training time, weight decay and jitter. Where alternative model approximation methods were used, some means of limiting overall size or complexity of outcomes was often applied.

In a review of 43 papers describing ANN application to modelling ware resource variables, Maier and Dandy (2000) concluded that the process of choosing ANN complexity for a task (ie stopping criteria, hidden layer geometry etc) is generally described poorly and/or carried out inadequately leading to sub-optimal performance and difficulty in attaining meaningful comparisons between models. It can be argued, given the review of parameters controlling ANN generalisation in section 2.3.3, that the reason that sub-optimal generalisation of models may be occurring in practice is that the problem of minimising total bias+variance is fraught with difficulties. Analytical approaches to the problem are unwieldy and yield loose complexity bounds (Moody, 1991) and empirical approaches based on cross-validation are inherently error prone in the context of an unstable classifier such as ANNs (Breiman, 1996b).

Also, it is possible that interactions between the effects of different complexity determining meta-parameters and/or model approximation methods further reduces the efficiency of this optimisation task. Studies such as Alpsan et al. (1995) and Lawrence and Giles (2000) go some way towards determining the effects of these issues on generalisation performance. However, with the proliferation of methods available today, the task of searching the space of possible combinations of methods and parameters is a task beyond even the most well equipped water resource manager.

It can be concluded that the issue of achieving optimal ANN model generalisation requires further research.

2.4.6.4 Validation

As stated in section 2.3.4, Flexer (1995) proposed a list of five minimum requirements for evaluation of ANN models (listed in table 2.1). No applications reviewed satisfied all five requirements, with most only satisfying the first requirement for separate training and validation sets. Furthermore, for most applications, only one to two years of data expressing extreme conditions with respect to the output variable were chosen for validation – for example, "bloom" and "non-bloom" years. While such an approach seems reasonable for estimating

model performance in extreme conditions, a number of obvious shortcomings are evident;

- No information is yielded regarding model performance under more moderate output conditions.
- There is no consideration of the values of the independent variables. For example, it may be desired to validate model performance for unusually warm or cool years.
- For datasets where the sampling interval is high, such an approach leads to questions about the significance of the performance estimates.
- Validation data is grouped into contiguous blocks of one or two years. This means that most validation set records will separated in time from training set records by more than one sample interval. This does not provide a realistic validation for a real-time forecasting application where the ANN is constantly retrained using the most up to date data.

Few applications reviewed compare modelling outcomes of machine learning approaches with more conventional inductive or deductive approaches. It can be argued that comparisons need to be made over a wider range of case studies in order to put the utility of ANNs in this application in a broader context. In particular, only 2 studies, Wilson and Recknagel (2001) and Todorovski et al. (1998), make a direct comparison between model inference constrained to multiple linear relationships and the unconstrained, non-linear machine learning approach. This is in spite of the fact that a key hypothesis to be investigated in any *model-free* model inference application is that unconstrained model inference provides outcomes that are superior as a result of increased flexibility.

It can be concluded that more robust approaches to validation need to be employed that provide the following features;

- Improved validation set representation.
- Does not use validation data for model selection purposes.
- Enables the use of statistical tests for performance comparisons.
- Provides comparisons with conventional or naive modelling techniques to provide a meaningful context for performance estimates.

2.4.6.5 Knowledge Discovery

This review shows that sensitivity analysis is the most used approach to elucidating knowledge and that the methodology has in recent times advanced to the point where it can illustrate non-linear relationships between input and output variables and interactions between input variables. However, it can be observed that in most cases, input variables besides the input under investigation are blocked at median

values. If there are interactions between inputs with respect to their effects on the output variable, which is a reasonable assumption given the non-linear nature of ANNs, observed sensitivity will depend on the values of the blocked input variables. This means that the sensitivity determined will only be relevant given a tiny region in the input space rather than representative of the generalised effect. Furthermore, it is possible that blocking all inputs at median values may push the input space into a region outside that which was used to create the model, because it is likely that in the reservoir or lake under investigation, all the environmental variables are never observed to be at median values simultaneously. This means that the sensitivities often reported in the literature may actually be indicative of the model's behaviour when it is effectively extrapolating, which, as Geman et al. (1992) explains, is when *model-free* inference methods such as ANNs are inherently unreliable.

It can be concluded that there is a need to determine how sensitivity analysis can be implemented in a way that assumes the following facts about learned models;

- Inputs are likely to have non-linear relationships with output variables.
- Inputs may have complex interactions with other input variables with respect to relationships with output variables and
- ANNs and other *model-free* inference methods are inherently unreliable when asked to make extrapolations.

2.5 Proposals for ANN Model Representation

The previous discussion arrived at a number of conclusions regarding requirements for further research. This section proposes a number of developments to the ANN model representation and methods that are intended to overcome the identified shortcomings of existing approaches. It is hypothesised that the suggested changes enhance the performance, stability and compatibility of the ANN modelling paradigm in the context of a decision support framework for operational control of algal blooms[4].

2.5.1 An "Input Window" Model Representation

It is proposed that an input window model representation can ensure compatibility between raw time-series data and TDNN models without the need to interpolate a regular sampling frequency. This approach represents input (and/or output) variables as summary statistics over a defined window of time relative to the

[4]See appendix A for a list of operational control measures for dealing with algal blooms that may benefit from short term forecasts.

output date as illustrated by figure 2.12. The summary statistic may be the mean, median, variance, range, trend, or any other appropriate statistic or model output. It is hypothesised that such an "input-window" representation has a number of benefits in the context of typical limnological time-series, since;

- Given a sufficiently long summary interval, the technique guarantees a high proportion of matching input records for a given output variable and lag interval. A short summary period will be adequate where the sampling frequency of the input variable is similar to that of the output variable. Longer summary periods can be used to access input variables that are sampled less frequently than the output variable without altering the overall data representation.
- By defining strict bounds on the time period input data is summarised, the technique eliminates the problem identified by Lee et al. (2003) that blurred boundaries between past, present and future states will bias performance expectations. It can be guaranteed that the model will not access information effectively "in the future" relative to the output date when making a prediction unless it is explicitly intended.
- Since interpolation is no longer necessary, any bias caused by assumptions of dynamics is eliminated. Also, the overall information processing task is significantly reduced.
- It provides scope for experimentation with the summary method, since inputs can be represented quantities describing many dynamic and/or statistical properties.

2.5.2 Improving Generalisation Qualities by Bagging

It is proposed that *bagging* (Breiman, 1994) can be used to stabilise ANN models to improve generalisation qualities. It is hypothesised that, as shown by Cannon and Whitfield (2002) and Wilson and Recknagel (2001), when bagging is applied, model error first decreases with increasing fitting power and then stabilises at a minimum, since aggregation effectively "cancels out" the variance component of model error typical in the overfitting phase[5]. Thus, as long as sufficient fitting power in terms of hidden layer configuration and training epochs is applied to prevent *underfitting*, bagging guarantees optimum generalisation eliminating the need for error prone analytical or empirical determination of model complexity.

[5]The relative importance of bias and variance in overall prediction error with increasing fitting power is illustrated in figure 2.9.

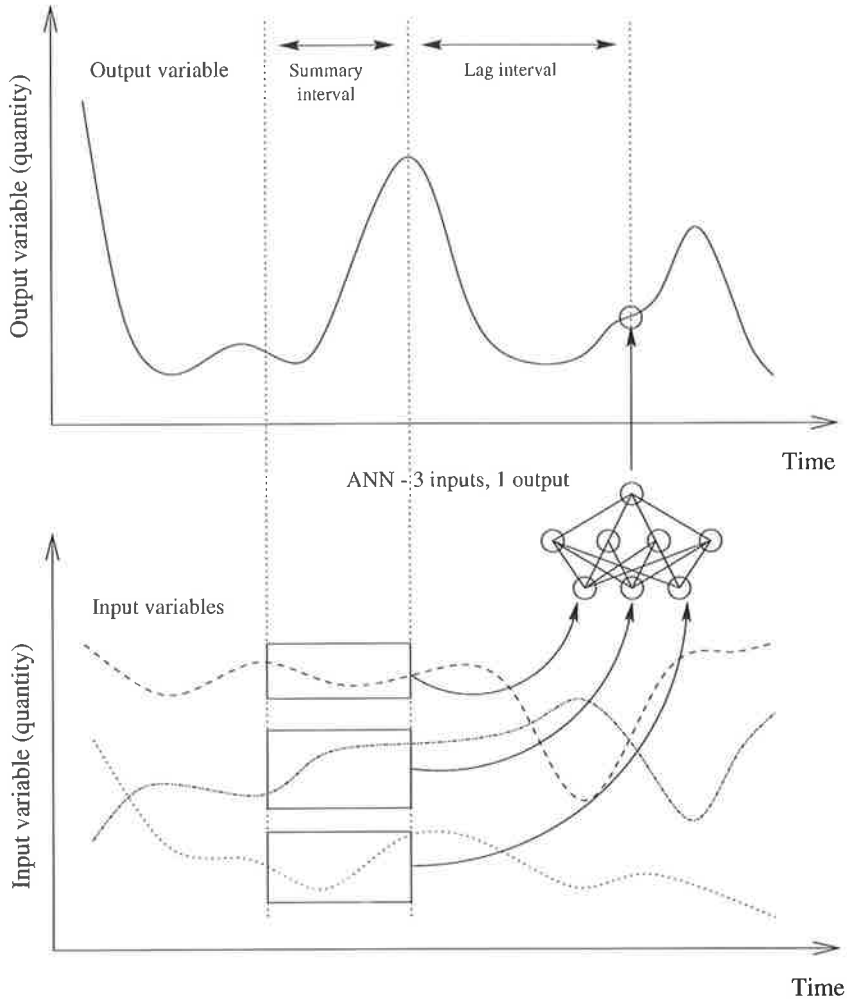

Figure 2.12: Time delay model using summary period.

2.5.3 Model Validation by Rotation Performance Estimators

It is proposed that representation of validation sets be increased by use of rotation estimators, where multiple subsamples of training and validation sets are used to gauge model performance. The following two approaches are suggested;

- **Leave-one-out bootstrap** is efficiently carried out in combination with bagging. N records are randomly subsampled with replacement from a training set of size N giving a training set of 67% of the available data on average (Weiss and Kulikowski, 1991). The "out-of-bag" records (on average, 33% of data) left behind by the subsampling process are used as validation set records. This process of bootstrap sampling and model training and validation is repeated a number of times until a reasonable distribution of model predictions on the out-of-bag data is obtained.
- **Leave-k-out cross-validation** is conducted by dividing sample into k equally

Table 2.5: Procedure for blocked 20-fold-crossvalidation with bagging

Loop for $i = 30$ **replicates**
  Block time-series into k "sub time-series" – starting position t
  Loop for $k = 20$ **blocks**
    Sample training pool – all data except block k
    Take a bootstrap sample of training data from pool
    Train ANN
    Generate predictions on current block and save.
  End block loop
  $t = t + 1$
End replicate loop

sized subsamples each used in turn as a validation set while the remainder are pooled as a training set.

Logically, random sampling of validation records from throughout the time-series may lead to different expectations of performance than if the validation data is blocked into a contiguous period. This is because non-stationarities in the time-series may lead to differences in the ANNs ability to generalise on short term "local" time scales as opposed to longer term "global" time scales. To investigate this possibility, it was elected to define a blocked leave-k-out crossvalidation, where the data is divided into k smaller time-series as illustrated in figure 2.13. Furthermore, a hold-out period from both training and testing of 90 days was designated after the final record in each block. This hold-out period was enforced when a given block was used for validation (but not training) (see figure 2.14). The purpose of this hold-out period is to decrease the potential for *temporal contamination* of validation data by reducing the likelihood of short-term serial correlation existing between it and data that is *in the future* in the time-series.

Blocked leave-k-out was combined with bagging by taking bootstrap samples of 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. With each replicate, the times at which the divisions are defined is altered to maximise training and set variability. The entire procedure is summarised in table 2.5. It is proposed that double cross-validation (ie use of a third "tuning" set) is not necessary, since the bagging methodology eliminates the need to tune complexity related meta-parameters to maximise generalisation (thus satisfying the third requirement outlined in table 2.1). The use of bagging in combination the rotation performance estimator effectively permits the remaining requirements in table 2.1 to be satisfied.

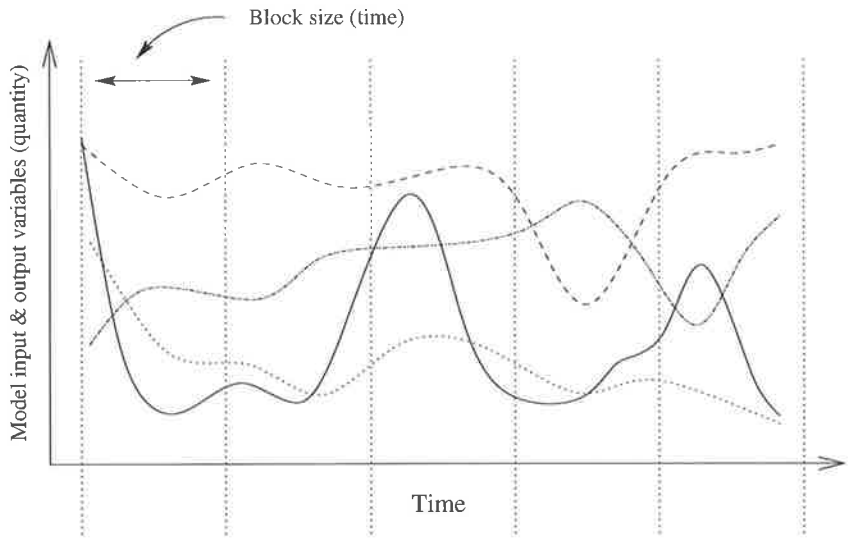

Figure 2.13: Dividing data into $k = 5$ discrete blocks in the time series.

2.5.4 Sensitivity Analysis Through Time

It is proposed that the sensitivity analysis method generally applied be broadened in several ways to provide a more accurate measure of the relative importance of inputs and to pave the way to discovering more detailed information about interactions between inputs. To take account of the assumption that ANNs learn non-linear relationships between input and output variables, it is proposed that the input variable in question be swept over a range of discrete values as per Lek's algorithm (Lek et al., 1996). Furthermore, to gain data regarding interactions between the effects of inputs, it is proposed to conduct the sensitivity analysis for each input where the values of the remaining inputs are blocked at each dataset value in turn (ie a *sensitivity analysis through time*). Implementation of these two strategies is described by the procedure in table 2.6. This procedure results in a database of model responses indexed by the input variable, perturbation value and output date. The following information can be retrieved from this database;

- Overall model sensitivity to a defined input.
- Model sensitivity to a defined input where other inputs and outputs fall within a given range of values.
- Model sensitivity to a defined input within a give time period where the time period is defined either by dates, or by a relationship to other unmodelled variables.

It is proposed that overall model sensitivity to a given input be calculated to take account of non-monotonic relationships that may have been learned between input and output variables. This can be achieved using the approach developed by Embrechts et al. (2001) described by equation 2.11 in section 2.3.5.

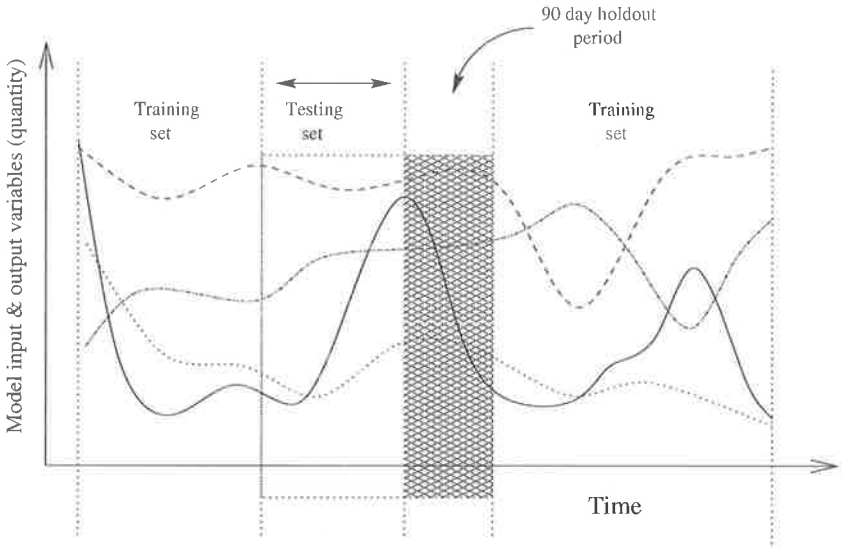

Figure 2.14: Division into train and test data with a hold-out period.

Table 2.6: Procedure for bagged sensitivity analysis through time

Loop for i **bootstrap samples**
  Loop for every j **test set dates**
    Loop for k **inputs**
      Loop for m **perturbations**
        Calculate output from input for date j using model i substituting input k with input k + perturbation m
      End perturbation loop
    End input loop
  End date loop
End bootstrap loop

If bagging is employed, the sensitivities will be represented as a bootstrap distribution allowing statistical significance tests.

2.5.5 "LakeNet" – a Platform for ANN Model Implementation

Implementing the model representations and methodology proposed above requires repeated preparation of training and validation sets followed by ANN training and testing. This process is computationally intensive because the modelling task considers a number of dimensions including;

- 6 datasets (see chapter 3).

- Many possible model designs, encompassing a range of input/output variables, lag times and input windows lengths and types.
- Many ANN "meta-parameter" settings such as training algorithm, number of hidden layer nodes and training time where experiments regarding these features is required.
- Training of many member models to make up a bagging ensemble.
- The use of rotation performance estimators requiring multiple training and validation samples.
- The use of the *sensitivity analysis through time* procedure with a number perturbations for each date-input.

Such a procedure requires training of thousands of individual ANNs with unique training and test sets. Even with access to fast computers, this is a huge computational and data management task. Such an undertaking would be time consuming and error prone using an interactive approach, where data preprocessing is done using spreadsheet applications and ANNs are trained using desktop software applications.

It is proposed that the middleware application "LakeNet" be developed to facilitate the information processing task. Middleware is software designed to be an intermediary between a client program and a database server. LakeNet performs the task of retrieving and preprocessing data and messaging the ANN client with appropriate control information and data. This messaging is achieved by means of the application programming interface (API) of the ANN client software. At the completion of training, validation and sensitivity analysis, LakeNet then retrieves the corresponding predictions from the ANN client and, after performing any necessary post-processing, inserts the information back into the database. It is proposed that LakeNet be implemented in the Java language (Sun Microsystems Inc., 2001) to maximise platform independence.

For the purposes of LakeNet, it is proposed that all data, including monitoring data from each of the study sites, configuration data for experiments and ANN clients and model predictions, be stored in tables in a relational database management system (RDBMS)[6]. Storing data in a RDBMS permits data to be defined in terms of its relationships to certain keys. This means that, unlike a "flat file" system such as a spreadsheet or text file, the positional information of a record or piece of information is not relevant. Compared to a flat file system, a RDBMS has the following advantages;

[6]In the present study, MySQL version 3.23.47 MySQL AB (2002) was used as the database platform and the command line client of SNNS version 4.1 was used as the ANN simulator. Any other RDBMS software that supports Structured Query Language (SQL), such as Microsoft Access, Oracle, or Postgres, would also be suitable.

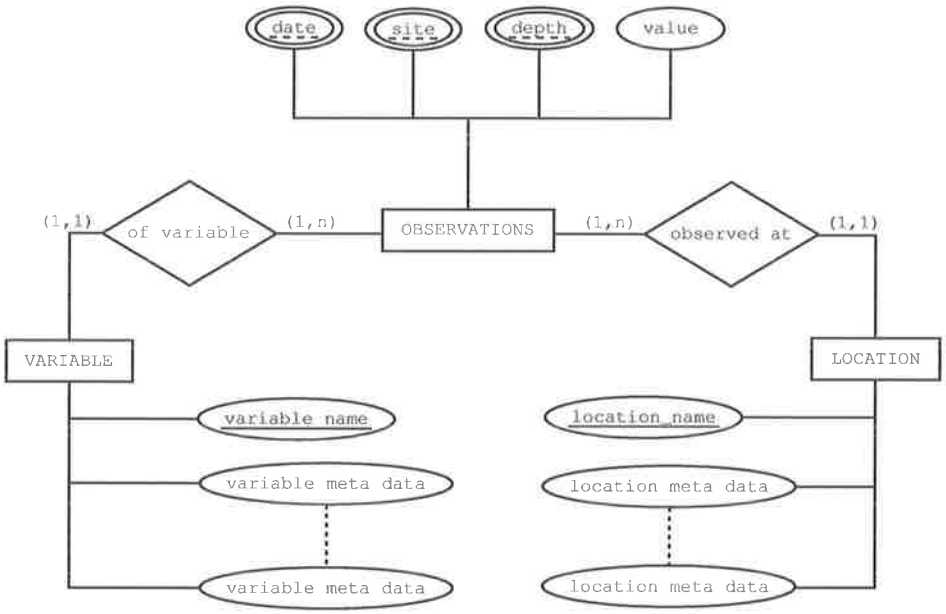

Figure 2.15: Database entity relationship diagram (ERD)

- The structure is much more flexible with respect to the insertion of new records or variables, since the interface with retrieving applications is kept constant.
- There is less risk of errors being made with respect to data manipulation or retrieval.

Figure 2.15 shows an entity relationship diagram describing the tables used to store water quality data.

2.6 Conclusion

ANNs are very flexible model representations that;

- May approximate a decision surfaces characterised by non-linear decision boundaries.
- In the context of a suitable training algorithm and proper "data conditioning", may learn decision surfaces from data within specified distortion criteria.

These properties are particularly useful when tackling modelling tasks for which there is a lack of sufficient domain knowledge to apply conventional modelling approaches. In addition, unlike conventional empirical models, they are not constrained by simplifying assumptions with respect to the data such as normality or linearity. The process model for developing models using supervised ANNs is

well established and it has been proven that ANNs can be used to develop models with prediction accuracy comparable or superior to conventional models.

However, it was argued that a number of practical issues concerning *compatibility of time-series representations with typical historical datasets, stability of model inference, performance estimation* and *knowledge discovery* affect use of ANNs in a decision support role. The following changes to model representation and approximation were proposed with respect to each of these issues;

- Representation of inputs as *sliding time windows* rather than discrete lags to increase the compatibility of TDNN structures with "typical" environmental datasets without extensive data preprocessing.
- Representation of models as bootstrap ensembles by *bagging* to increase the stability of model inference and increase resistence to the effects of *overfitting*.
- Use of rotation performance estimators such as k-fold cross-validation or the leave-one-out bootstrap to improve the accuracy of performance estimation leading to better judgements about the model's usefulness and more accurate tuning of parameters that effect model performance.
- Use of the *sensitivity analysis through time* to yield more accurate information regarding the relative importance of model inputs.

This thesis will investigate the utility of each of these proposals.

Chapter 3

Study Sites and Data

3.1 Introduction

One of the aspects that sets the present study apart from other related studies is that, through the kindness of a number of many scientists and ecologists, there is a wide range of datasets available from which ANN models can be trained and validated. This is fortunate, since, as the aim of the present study is to develop more generalised "compatible" approaches to ANN modelling, it is important to that the methods be tested on the widest possible range of data. This allows determination of interactions between the effects of model design and site specific properties such as eutrophication, the morphometry and residence time of the water body, various chemical attributes of the water, climate, the regularity and span of monitoring and other relevant attributes. Historical data collected by water quality management authorities was donated from 6 sites, including;

- Lake Biwa, Japan
- Burrinjuck Dam, New South Wales, Australia
- Darling River, New South Wales, Australia
- Lake Kasumigaura, Japan
- Myponga reservoir, South Australia
- Lake Soyang, South Korea

Sections 3.2.1 to 3.2.6 briefly review the conditions of each of these water bodies and the characteristics of the available data. Section 3.3 compares the trophic states of each site discovered by investigation of the available data. Section 3.4 proposes *site generic* and *site specific* model designs for ANNs based on the observed data availability.

Table 3.1: Six freshwater bodies: water quality, monitoring, and database information

	Lake Biwa (Japan)	Lake Burrinjuck (NSW, Australia)	Darling River (NSW, Australia)	Lake Kasumigaura (Japan)	Myponga Reservoir (SA, Australia)	Lake Soyang (South Korea)
Water Quality						
mean Chl a (μg/l)	9.32	15.8	20100*	60.5	7.45	4.30
max Chl a (μg/l)	38.5	579	281000*	280	41.6	98
std dev Chl a (μg/l)	6.5	28.7	26100*	42.5	6.77	6.81
mean annual min temp (°C)	4.9	9.1	9.7	4.5	9.7	5.1
mean annual max temp (°C)	29.5	25.6	27.2	28.8	22.3	27.0
mean transparency	1.76 m**	1.55 m**	101 NTU***	0.84 m**	5.09 NTU***	3.9 m**
Morphometry						
max depth (m)	103	63.5	n/a	7	36	118
mean depth (m)	41	56.6	n/a	4	not avail.	35.3
area (km^2)	670	4.2	n/a	220	not avail.	46.5
volume (10^6 m^3)	27800	756	n/a	900	26.8	1650
retention time (years)	5.5	>2	0.002	0.55	<1	0.77
Database						
lifespan of database	1984–91	1978–97	1978–93	1978–93	1970–97	1984–95
no. chl-a sample dates	151	330	628	125	646	120
approx. sampling interval (days)	17	22	9	47	16	37
no. variables sampled	31	56	65	105	236	21
no. sample sites	1	13	1	3	2	1
depth of sampling?	no	yes	no	no	no	yes

* Chlorophyl a data not available. Total cells/ml used instead.
** Depth of secchi disk.
*** Turbidity

3.2 Study Sites

3.2.1 Lake Biwa

Lake Biwa, located on the island of Honshu approximately 10 km from the city of Kyoto, is the largest freshwater lake in Japan. As one of the oldest freshwater lakes in the world, it has attracted considerable scientific interest on account of its unique biota and fossil rich sediments. It is of great economic importance to the surrounding Shiga prefecture as the host of Japan's largest freshwater fishery and an important source of freshwater for domestic, industrial and agricultural purposes.

Figure 3.1 shows that the morphometry of Lake Biwa is comprised of two basins joined by a 1.3 km wide narrows. The northern or "main" basin is the largest and has an average depth of 43 metres making it the second deepest lake in this study (see comparison in table 3.1). The southern or "secondary" basin has an average depth of only 4 metres. The two basins have considerably different water quality and biological conditions as a result of the different morphometry. The large surface area of 670 km^2 and high average depth mean that lake Biwa has a volume of approximately 27800 * 10^6 m^3 which is by far the largest of all the lakes studied. Also it has the longest water retention time of 5.5 years.

Figure 3.1: Lake Biwa (Japan).

This lake was once considered oligotrophic. Many wetlands and smaller lakes surrounding Lake Biwa have had a beneficial effect on water quality. However, since World War 2, increasing land reclamation of the wetlands and industrialisation in surrounding areas have caused a decline in water quality to the point that it is now considered meso-eutrophic. In recent years, "red-tides" resulting from

dinoflagellate blooms and potentially toxic cyanobacterial blooms have become a regular occurrence.

Figure 3.2 shows that the summer climate of Kyoto near Lake Biwa is sub-tropical with high rainfall and relatively warm maximum and minimum temperatures. The winter months however are cool to cold with average minimum temperatures from December to February near freezing point.

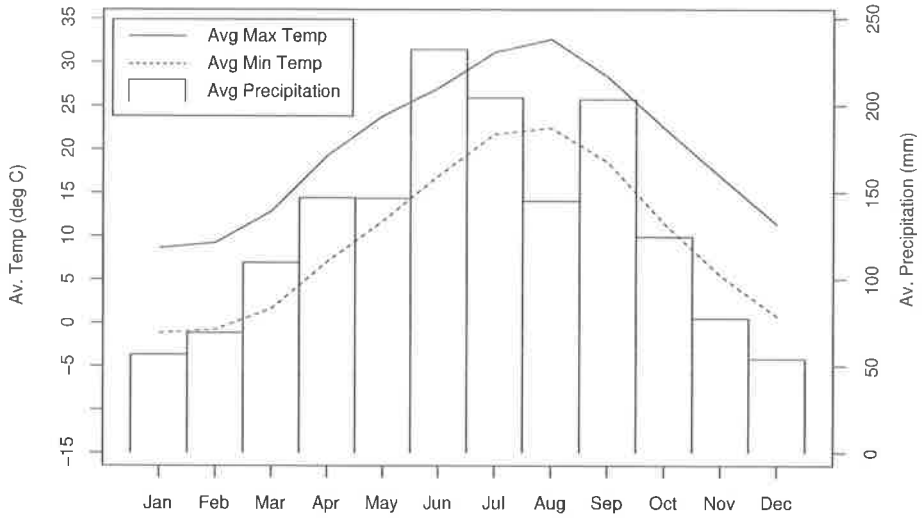

Figure 3.2: Average temperature and precipitation – Lake Biwa

A total of 31 measured variables were available for this lake measured over an eight year period from 1984 to 1991. Table 3.2 shows that the database for lake Biwa contains commonly measured chemical water quality parameters including plant macro-nutrients (nitrate and phosphate), dissolved oxygen, pH and a single micro-nutrient (Si). Physical information is represented by temperature and underwater light penetration (Secchi disk depth). In addition to these common physical variables, there is basic information on weather conditions such as wind speed and a trinary variable indicating fine, cloudy and rainy conditions.

The biological variables include chlorophyll *a* as a measure of total algal biomass. Phytoplankton data (see table 3.3) are resolved to species level. The average, maximum and standard deviation information regarding cell densities indicate that the mixotrophic flagellate *Euglena americana* is the most abundant and dynamic phytoplankton species in Lake Biwa. This species is known to form an algal bloom every spring when phosphorus becomes limiting (Urabe et al., 1999). Other dominant algae include species of diatoms (eg *Melosira granulata*) and green algae (eg *Pediastrum biwae*).

Table 3.2 shows that there were between 104 and 190 unique sampling dates. Chemical properties such as NO_3, PO_4 and Si were the least well represented with 104 dates. Chlorophyll *a* measurements are present for 151 dates and the remaining variables are available for 190 dates. For months when observations were

Table 3.2: Lake Biwa: Sampling Frequency

	Lifespan	Obs. dates	Obs months	Obs. per mo.
Water quality & physical variables				
Chlorophyll *a*	1984–91	151	96	1.6
Dissolved oxygen	1984–91	190	96	2.0
NO_3	1984–91	104	96	1.1
PO_4	1984–91	104	96	1.1
pH	1984–91	190	96	2.0
Secchi depth	1984–91	190	96	2.0
Si	1984–91	104	96	1.1
Water temperature	1984–91	190	96	2.0
Weather (0 fine 05 cloudy 1 rain)	1984–91	190	96	2.0
Wind speed	1984–91	190	96	2.0
Phytoplankton				
Ankistrodesmus fal v mirabile	1984–91	190	96	2.0
Asterionella formosa	1984–91	190	96	2.0
Coelastrum cambricum	1984–91	190	96	2.0
Cyclotella glomerata	1984–91	190	96	2.0
Euglena americana	1984–91	190	96	2.0
Melosira granulata	1984–91	190	96	2.0
Micractinium pusillum	1984–91	190	96	2.0
Pediastrum biwae	1984–91	190	96	2.0
Planktosphaeria spp.	1984–91	190	96	2.0
Rhodomonas spp.	1984–91	190	96	2.0

Table 3.3: Lake Biwa: 10 most abundant phytoplankton species (cells/ml)

Var. name	av. var.	stdev. var.	max. var.
Euglena americana	549	1789	18780
Melosira granulata	304	618	3345
Pediastrum biwae	141	555	5056
Asterionella formosa	128	525	4400
Ankistrodesmus fal v mirabile	120	279	2526
Rhodomonas spp.	89	205	1838
Coelastrum cambricum	88	572	6000
Dictyosphaerium spp.	81	230	2175
Cyclotella glomerata	77	163	963
Micractinium pusillum	48	135	1300

conducted, there was an average sampling rate of between 1.1 and 2 observations per month.

3.2.2 Burrinjuck Reservoir

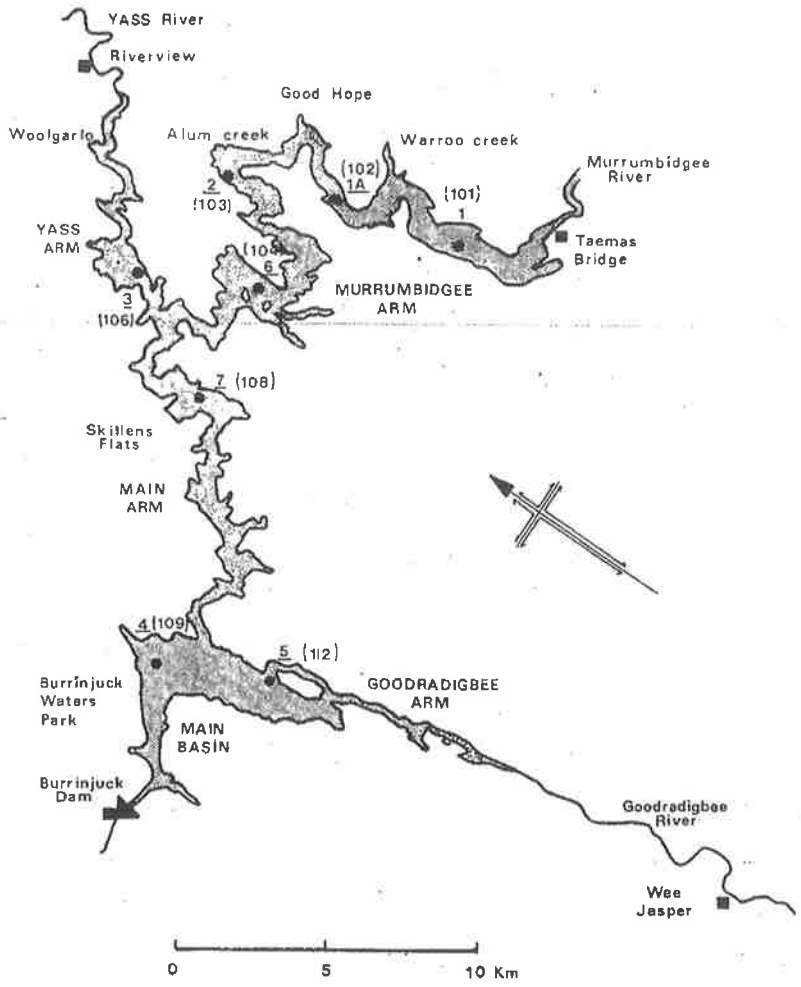

Figure 3.3: Lake Burrinjuck (NSW, Australia).

Lake Burrinjuck is a major storage for the Murrumbidgee Irrigation Area in New South Wales, Australia. It was created in 1908 by damming the Murrimbidgee river downstream of the junctions with the Yass and Goodradigbee rivers. It is located approximately 340 km southwest of Sydney near the township of Yass. As well as being an important water supply for irrigators, it is popular for recreational activities such as fishing, boating and water-skiing. Figure 3.3 shows that Lake Burrinjuck is dendritic in shape with the main basin being joined by several long, narrow arms corresponding to the Goodradigbee, Murrimbidgee and Yass rivers.

There are many small bays and inlets along its edges. This lake is considered to be meso- to eutrophic and has experienced recurrent algal blooms since the 1960's.

Figure 3.4 shows that Lake Burrinjuck has a warm temperate climate with more precipitation occurring in winter months than summer months. Whilst the average summer maximum temperatures are similar to those of the 3 Asian lakes featured in this work, the summer minimum temperatures are considerably warmer. Winters in South Eastern Australia are relatively mild with average maximum temperatures in the mid teens and minimum temperatures above freezing.

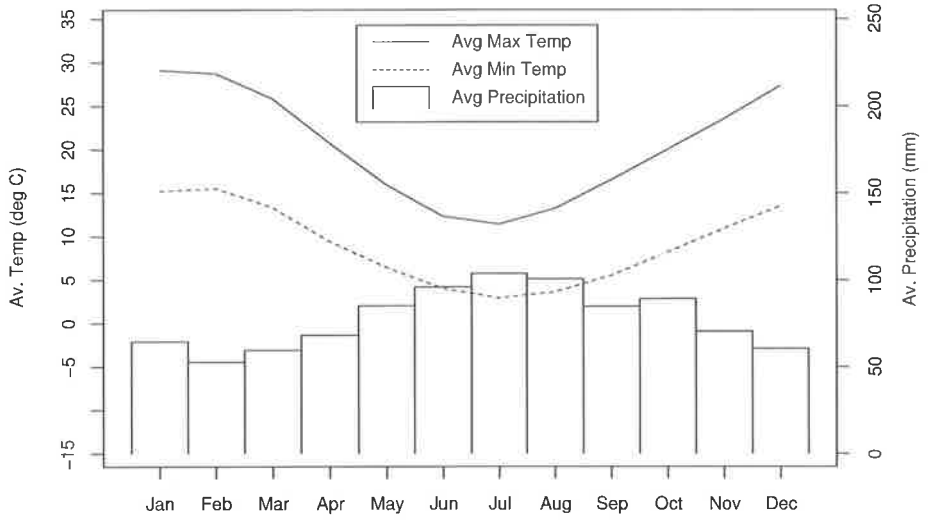

Figure 3.4: Average temperature and precipitation – Lake Burrinjuck

A total of 56 variables were available for this lake measured over a 22 year period from 1976 to 1997. Table 3.4 shows that the chemical information in the database includes several representations of both phosphorus and nitrogen including dissolved and total phosphorus, ammonia, oxidised nitrogen and total nitrogen. Also included are data for dissolved oxygen. However, there are no data for important micro-nutrients such as silica. The physical data includes variables describing water temperature, underwater light (Secchi depth) weather information (sunshine hours, precipitation, wind and evaporation) and water inflow rates from a number of tributary streams and rivers ("Ginnind & Charnwood", "Goodradigbee", "Molonglo Coppins", "Mountain Creek", "Murrum Mt McD", "Yass" and all variables commencing with "S410"). Also there is data for the lake volume and surface area – information particularly relevant for this site given the large fluctuations in water level caused by irrigation drawdown and evaporation in summer.

The biological data (see table 3.5) include variables describing total algal biomass (chlorophyll *a*) and abundance of a number of families of zooplankton and a single macro-invertebrate group (nymphs). Phytoplankton abundance is resolved at functional group level rather than species level. It is clear from this table cyanobacteria are by far the most productive and dynamic phytoplankton group

Table 3.4: Burrinjuck Dam: Sampling Frequency

	Lifespan	Obs. dates	Obs months	Obs. per mo.
Water quality & physical variables				
Air temp – maximum	1976–96	7055	250	28.2
Air temp – minimum	1976–96	7236	250	28.9
Area	1976–96	7671	252	30.4
Chlorophyll *a*	1977–97	283	190	1.5
Dirn 900	1976–91	2515	89	28.3
Dissolved oxygen	1978–97	207	162	1.3
Dissolved phosphorous	1977–97	359	205	1.8
Evaporation	1976–96	7455	247	30.2
Ginnind & Charnwood	1976–97	7673	253	30.3
Goodradigbee	1976–96	7671	252	30.4
Molonglo Coppins	1976–97	7673	253	30.3
Mountain Creek	1976–96	7671	252	30.4
Murrum MtMcD	1976–97	7673	253	30.3
NH_4	1977–97	353	207	1.7
NO_x	1977–97	363	207	1.8
Precipitation	1976–96	7572	250	30.3
Relative humidity 1500	1976–96	7671	252	30.4
Relative humidity 900	1976–96	7671	252	30.4
S410008	1976–96	7671	252	30.4
S410700	1976–97	7673	253	30.3
S410731	1976–97	7673	253	30.3
S410745	1976–97	7673	253	30.3
S410761	1976–97	7673	253	30.3
Secchi depth	1979–97	173	143	1.2
Stratification	1977–97	295	193	1.5
Sunshine hours	1976–96	7640	251	30.4
TKN	1982–96	165	129	1.3
Total nitrogen	1977–97	50	42	1.2
Total phosphorous	1977–97	358	204	1.8
Volume	1976–96	7671	252	30.4
Water level	1976–97	7672	253	30.3
Water temperature	1977–97	350	200	1.8
Wind speed 1500	1976–96	6916	250	27.7
Wind speed 900	1976–96	7090	250	28.4
Yass	1976–96	7671	252	30.4
Phytoplankton				
Chlorophyta	1977–97	295	193	1.5
Chrysophyta	1985–97	126	105	1.2
Cyanophyta	1977–97	295	192	1.5
Diatoms	1977–97	289	195	1.5
Euglenophyta	1985–97	113	94	1.2
Total Algae	1977–97	302	196	1.5
Xanthophyta	1985–97	114	93	1.2
Zooplankton				
Ciliophora	1984–97	45	41	1.1
Cladocera	1982–97	116	106	1.1
Copepoda Calanoida	1985–97	113	101	1.1
Copepoda Cyclopoida	1985–97	106	97	1.1
Copepoda Harpacticoida	1993–97	38	34	1.1
Nymphs	1982–97	134	122	1.1
Rotifera	1983–97	80	74	1.1
Total zooplankton	1982–97	135	123	1.1

Table 3.5: Burrinjuck Dam: Most abundant phytoplankton groups (cells/ml)

Var. name	av. var.	stdev. var.	max. var.
Cyanophyta	27499	624228	34122240
Chlorophyta	2612	8983	386640
Diatoms	1527	3496	77100
Chrysophyta	1162	3039	77100
Euglenophyta	30	228	8253
Xanthophyta	10	84	1500

in this reservoir. Also, the maximum values column indicates the occurrence of significant blooms of green algae and diatoms during the monitoring period.

Table 3.4 shows that data availability varies from 38 dates in 34 months for Copepoda Harpacticoida to over 6900 dates in approximately 250 months for variables describing meterological conditions and inflow. The observation density ranges from 1.1 to 1.8 dates per month for most water quality variables including phytoplankton, to an average of 1 observation per day for the data describing meterological conditions and inflow.

3.2.3 Darling River

The Darling river is a significant part of the Murray-Darling system, which at 3780 km in total combined length is Australia's largest and the world's forth largest, river system. This river system drains the Murray Darling Basin which is a region of approximately 1 million km^2 comprising a variety of alpine, temperate and arid landscapes to north and west of the Great Dividing Range. This area has great economic importance to Australia as it supports a large pastoral, cropping and irrigation based agricultural industry. Also, it provides domestic and industrial water supply to many towns and cities – the most significant being Adelaide which derives approximately 50% of its domestic water supply from the Murray River (McKay and Moeller, 2001).

The Darling River drains the north part of the Murray-Darling basin including much of northern NSW and southern Queensland. It is a major source of water for urban, industrial and agricultural purposes throughout this region. The Darling river joins the Murray river close to the township of Mildura which is on the western part of the border between Victoria and NSW. At this juncture it has an average annual flow rate of 1890 GL (Weston, 1987).

The Darling River has a number of water quality problems including taste, salinity, turbidity and occasional cyanobacterial blooms. In 1991, the Darling River experienced the world's largest recorded algal bloom, with floating scums of cyanobacteria being observed over a 1000km length of the river (McKay and

Moeller, 2001). In general, these problems are attributed to poor quality water flowing from tributaries arising from excessive soil erosion and agricultural runoff and natural salinity compounded by high evaporation rates (Weston, 1987).

Figure 3.5 shows the climate information for Wentworth, NSW which is approximately 60 km south of Burtundy where the data used in this study was collected. This region of Australia is warm and arid with low average rainfall and high summer maximum temperatures. Winters are mild with average maximum temperatures in the teens.

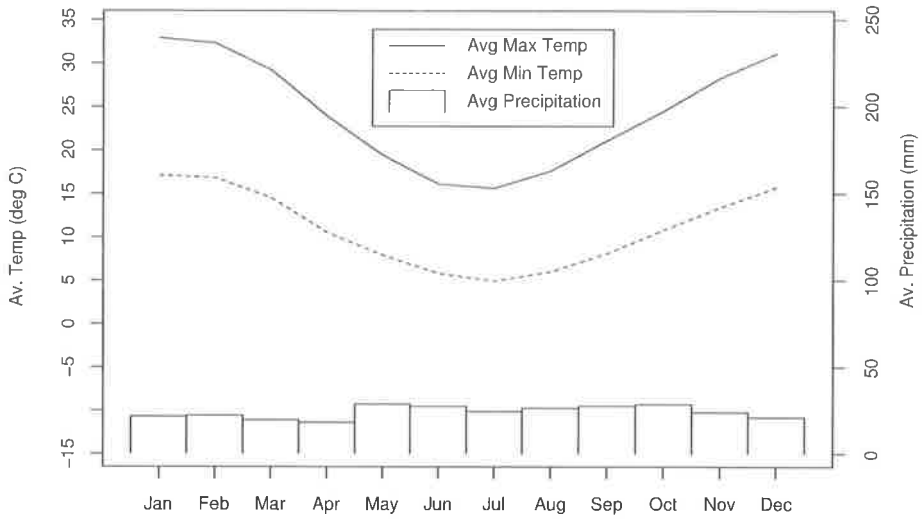

Figure 3.5: Average temperature and precipitation – Darling River (location Wentworth NSW)

A total of 65 variables were available in the Darling River database measured over 16 years from 1978 to 1993. Table 3.6 shows that macro-nutrient data is represented by dissolved and total phosphorus, oxidised nitrogen and TKN. Also it is clear that salinity is a particular concern in the Darling river as there is data for electrical conductivity and a number of relevant ions. Silica concentration and pH are also represented, but no dissolved oxygen data is available. Physical variables include temperature, but not Secchi disk depth. Instead, underwater light conditions are represented by turbidity and colour. The presence of flow data distinguishes this water body as a river rather than a lake. There are no weather data available.

Chlorophyll *a* concentration is not represented amongst the biological variables in this database. However, unlike the other databases, heterocyst counts are available indicating the innoculum levels for certain species of cyanobacteria. Phytoplankton data, represented by functional groups rather than species (see table 3.7), shows that the Darling River is equally dominated by cyanobacteria and chlorophyta. These two groups were responsible for approximately double the overall phytoplankton biomass of flagellates which were the next most dominant

Table 3.6: Darling River: Sampling Frequency

	Lifespan	Obs. dates	Obs months	Obs. per mo.
Water quality & physical variables				
Bicarbonate	1978–93	390	175	2.2
Calcium	1978–93	513	164	3.1
Chloride	1978–93	494	176	2.8
Colour	1979–93	175	140	1.3
E.C. - field	1978–93	721	172	4.2
E.C. - lab	1978–93	662	170	3.9
Flow	1978–93	5479	180	30.4
Magnesium	1978–93	513	164	3.1
NO_x	1978–93	600	177	3.4
pH - field	1978–93	710	175	4.1
pH - lab	1978–93	574	166	3.5
Potassium	1978–93	510	164	3.1
Silica	1978–93	611	177	3.5
Sodium	1978–93	512	164	3.1
SRP	1978–93	529	164	3.2
Sulphate	1978–93	398	175	2.3
Water temperature	1978–92	656	159	4.1
TKN	1979–91	454	139	3.3
Total phosphorous	1978–93	613	178	3.4
Turbidity	1978–93	744	178	4.2
Phytoplankton				
Chlorococcales	1980–92	628	148	4.2
Chlorophyta	1980–92	628	148	4.2
Cyanophyta	1980–92	628	148	4.2
Centric diatoms	1980–92	628	148	4.2
Unicellular diatoms	1980–92	628	148	4.2
Ditomophyta	1980–92	628	148	4.2
Flagellates	1980–92	628	148	4.2
Planctonema spp.	1980–92	628	148	4.2
Scenedesmus spp.	1980–92	628	148	4.2
Ulothricales spp.	1980–92	628	148	4.2

Table 3.7: Darling River: 10 most abundant phytoplankton groups (cells/ml)

Var. name	av. var.	stdev. var.	max. var.
Cyanophyta	3541	5896	67533
Chlorophyta	3170	5887	63233
Flagellates	1658	2398	25000
Ulothricales spp.	1414	5063	62729
Chlorococcales	1195	1703	19725
Planctonema spp.	972	4543	62729
Ditomophyceae	923	1465	11388
Centric diatoms	754	1351	11388
Unicellular diatoms	668	1141	9689
Scenedesmus spp.	557	883	13280

group. The maximum value column in table 3.7 indicates that significant blooms of the flagellates, *ulothricales* and *planctonema* have also occurred at some point in the time-series.

Table 3.6 shows that the data availability varies from 175 observations over 140 months for colour, to 5479 observations over 180 months for flow. Most variables, including phytoplankton data, have a reasonably high sampling density of between 3 and 4.2 observations per month. Flow rate was observed every day.

3.2.4 Lake Kasumigaura

Lake Kasumigaura, located on the Kanto plain 50km north-east of Tokyo, is the second largest lake in Japan after Lake Biwa. It has had high economic importance as a fishery throughout the twentieth century (Otsuki et al., 1987) and is also popular for recreational uses such as boating. The lake is large and shallow with an average depth of 4m and a surface area of 171km^2. It is very low lying with an elevation of only 1m. Figure 3.6 shows that it comprises a main basin of approximately 20 km length and 10 km wide at the widest points. There are two bays approximately 10 km length and widths ranging from 1 to 5 km. The shallow depth means that significant thermal stratification is a rare event in this lake.

This lake has become a highly eutrophic water body over the course of the twentieth century, especially following industrialisation and urbanisation in the vicinity after World War 2, the introduction of net pen carp culture in 1965 and flow regulation with the construction of a dam in 1974. Cyanobacteria such as *Microcystis*, *Anabaena* and *Aphanizomenon* were observed as early as 1910 and became dominant after succeeding *Melosira* in 1957 (Takamura et al., 1987). A succession in dominance from *Microcystis* to *Oscillatoria* occurred in 1987 (Takamura and Aizaki, 1991). In general, cyanobacterial blooms start forming in both of the bays

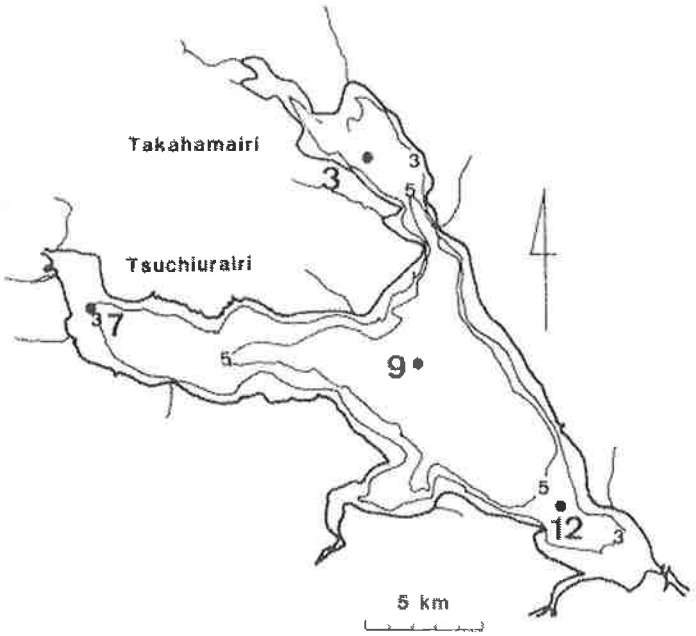

Figure 3.6: Lake Kasumigaura (Japan).

in late spring and early summer and eventually spread out over the entire lake by late summer (Otsuki et al., 1987).

Table 3.7 shows that Tokyo (approx 60 km south west of Lake Kasumigaura) has a subtropical climate with a warm summer and high summer and autumn rainfall. However, like Lake Biwa, this area has mild to cool winters with minimum temperatures close to freezing.

A total of 105 variables were available in the Lake Kasumigaura database measured over a 16 year period from 1978 to 1993 in the case of phytoplankton and from the early 1980's to 1993 for the remaining variables. Table 3.8 shows macronutrient data being represented by NO_2, NO_3, NH_4, dissolved inorganic nitrogen, total nitrogen, orthophosphate, dissolved total phosphorus and total phosphorus. Also among the available chemical variables are silica, pH and dissolved oxygen. The physical variables include transparency (Secchi depth), water temperature and water depth data. Weather information is also present with rainfall, radiation time and intensity variables available from a number of stations in the vicinity of the lake. Wind information is not present.

The biological database for this lake is relatively rich with data available for chlorophyll *a*, a number of groups of zooplankton and phytoplankton cell counts resolved to either species or genus level (see table 3.9). The phytoplankton data indicates that Lake Kasumigaura is hypertrophic with very high average and maximum cell counts for numerous species of cyanobacteria (in particular *Microcystis spp.* and *Oscillatoria spp.*). Diatoms and flagellates are also important

Table 3.8: Lake Kasumigaura: Sampling Frequency

	Lifespan	Obs. dates	Obs months	Obs. per mo.
Water quality & physical variables				
Chl-a	1983–93	125	122	1.0
Water depth	1983–93	93	88	1.1
Dissolved inorganic nitrogen	1981–92	155	140	1.1
Dissolved oxygen	1983–93	127	119	1.1
Dissolved total phosphorous	1983–93	135	125	1.1
Light	1981–93	222	155	1.4
NH_4	1983–93	135	125	1.1
NO_2	1983–93	128	120	1.1
NO_3	1983–93	128	120	1.1
pH 0m	1983–93	124	116	1.1
PO_4	1983–93	135	125	1.1
Radiation time (Kashima)	1981–92	4299	144	29.9
Radiation time (Tsuchiura)	1981–92	4299	144	29.9
Rain (Kashima)	1981–92	4299	144	29.9
Rain (Tsuchiura)	1981–92	4299	144	29.9
Total silica	1981–92	155	140	1.1
Total nitrogen	1983–93	134	125	1.1
Total phosphorous	1983–93	135	125	1.1
Water temperature	1983–93	134	124	1.1
Phytoplankton				
Anabaena flos-aquae	1978–93	277	178	1.6
Cyclotella sp. 1	1978–93	277	178	1.6
Gomphosphaeria spp.	1978–93	277	178	1.6
Merispodeia spp.	1978–93	277	178	1.6
Microcystis aeruginosa	1978–93	277	178	1.6
Microcystis wesen	1978–93	277	178	1.6
Ochromonas spp.	1978–93	277	178	1.6
Oscillatoria spp.	1978–93	277	178	1.6
Phormidium spp.	1978–93	277	178	1.6
Synedra rumpens	1978–93	277	178	1.6
Zooplankton				
Bosmina fatalis	1981–92	211	136	1.6
Cladocera	1981–92	211	136	1.6
Copepoda	1981–92	211	136	1.6
Diaphanosoma brachyurum	1981–92	211	136	1.6
Rotifera	1981–92	211	136	1.6
Total zooplankton	1981–92	211	136	1.6

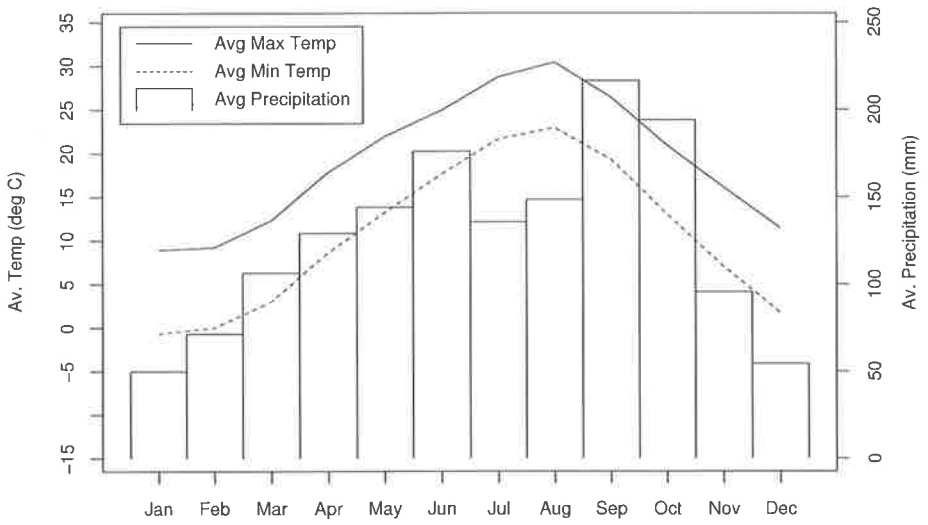

Figure 3.7: Average temperature and precipitation – Lake Kasumigaura

Table 3.9: Lake Kasumigaura: 10 most abundant phytoplankton species (cells/ml)

Var. name	av. var.	stdev. var.	max. var.
Microcystis aeruginosa	36899	87793	670000
Oscillatoria spp.	18302	53049	502320
Gomphosphaeria spp.	9773	31204	331240
Phormidium spp.	9097	31920	385476
Anabaena flos-aquae spp.	4247	18073	230000
Ochromonas spp.	3267	8294	127399
Synedra rumpens	3149	10913	143634
Merismopedia spp.	2738	26926	604250
Cyclotella spp. 1	2403	7556	75420
Microcystis wesen	1767	10414	141232

groups with high maximum cell counts indicating blooms of *Synedra spp.* and *Ochromonas spp.* taking place during the time-series.

Table 3.8 shows data availability varying from 93 dates in 88 months for water depth to 4299 dates in 144 months for rain and radiation time. Most water quality data availability is in the region of 120-150 dates and measurements of phytoplankton were made on 277 dates. The sampling density varies between 1 and 1.6 observations per month, although the density for rain and radiation time is approximately once per day.

3.2.5 Myponga Reservoir

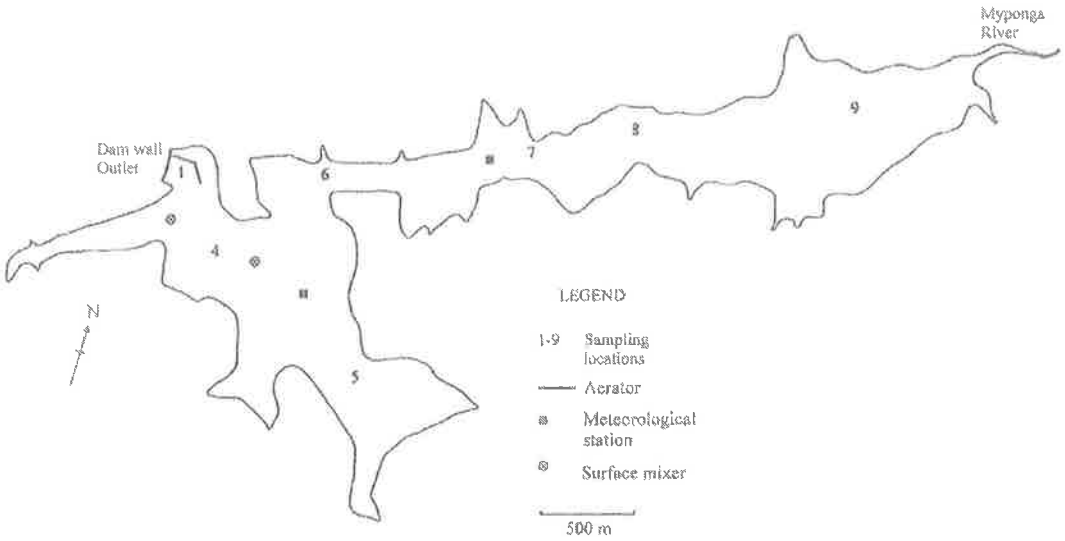

Figure 3.8: Myponga Reservoir (SA, Australia) (from Lewis et al. (2002)).

Myponga Reservoir, located approximately 60 km south of Adelaide, South Australia, was constructed in 1962 by damming the Myponga river in order to provide reticulated water supply to the towns and industries in the region from the township of Myponga to the southern metropolitan area of Adelaide. The catchment area for this reservoir is 125 square kilometres of predominantly agricultural land which takes in the township of Myponga (Government of South Australia, 1962). With an area of $3.2 * 10^2$ m^2 and a volume of $26.8 * 10^6$ m^3, Myponga reservoir is the smallest of the lakes investigated in this study. Figure 3.8 shows that the reservoir is a dendritic shape with a number of bays extending from the main basin. The long arm corresponds to the inundated path of the Myponga River.

The water quality has been characterised by high colour and recurrent cyanobacterial blooms since the dam's construction (Harvey, 1992; Velzeboer et al., 1991). Copper sulphate has been applied up to 3 times a year since 1963 to combat the cyanobacterial blooms and the associated water quality problems (McAuliffe and

Rosich, 1989). Artificial destratification by means of aeration was commenced in 1980 primarily to prevent problems with phantom midge larvae entering the reticulated water supply. Unfortunately this has had no significant effect on cyanobacterial growth (Harvey, 1992).

Figure 3.9 shows that the Myponga reservoir experiences a Mediterranean climate with warm dry summers and mild wet winters. The Myponga region has the coolest summers of all the lakes studied with average maximum temperatures in the mid to high twenties. Like the other Australian lakes, winter minimum temperatures are relatively mild being not less than 5 degrees Celsius.

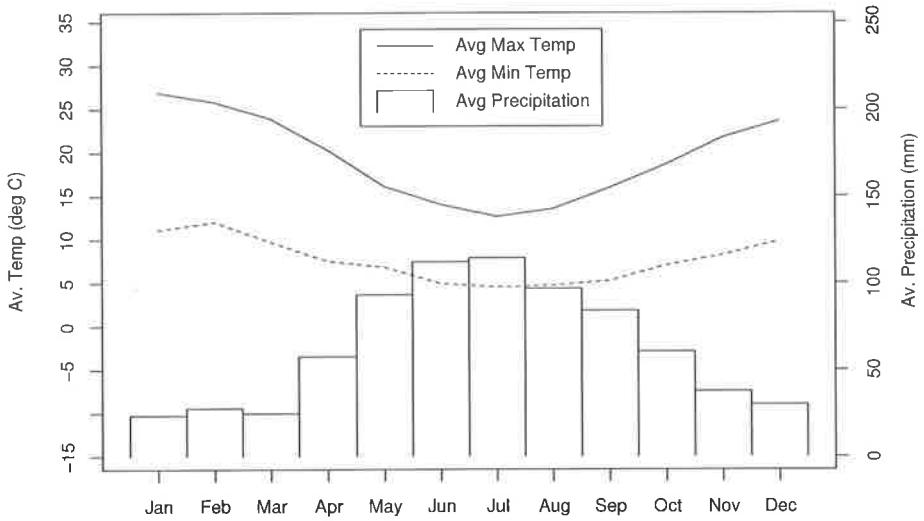

Figure 3.9: Average temperature and precipitation – Myponga Reservoir

In terms of the number of variables measured, Myponga presents the richest database in this study with 236 fields. However a number of fields contained less than 20 observations making them unsuitable for use for ANN modelling purposes. Table 3.10 shows the better represented variables with 20 or more records. It can be seen that sampling occurred from 1970 to 1997 for a few variables such as turbidity, although the database is richest in the period from 1984 to 1997.

Macro-nutrient data include variables representing NH_3, NO_2, NO_3, TKN, orthophosphate and total phosphorus. A range of trace elements are recorded including aluminium, iron and manganese but not silica. Copper data is also available which is to be expected in light of the regular copper sulphate dosing to control phytoplankton growth. Other chemical variables include pH, dissolved oxygen and dissolved organic carbon data. Also there is a number of variables alluding to taste or odour not listed in table 3.10 ("geranium", "sweetish", "fishy", "earthy", "peaty", "musty", "mouldy", "odour" and "vegetable"). Amongst the physical data, information is available for temperature and underwater light properties

Table 3.10: Myponga reservoir: Sampling Frequency

	Lifespan	Obs. dates	Obs months	Obs. per mo.
Water quality & physical variables				
E. coli	1970–97	560	136	4.1
Aluminium – soluble	1984–97	148	147	1.0
Aluminium – total	1984–97	148	147	1.0
Chlorophyll *a*	1985–97	645	143	4.5
Chlorophyll *b*	1985–97	643	143	4.5
Coliforms	1970–97	365	97	3.8
Colour – true (395nm)	1971–91	1066	240	4.4
Colour – true (456nm)	1991–97	273	73	3.7
Conducitvity	1984–92	92	76	1.2
Copper – soluble	1984–97	148	147	1.0
Copper – total	1984–97	148	147	1.0
Cyclopoida	1971–78	30	25	1.2
Dissolved organic carbon	1984–97	145	136	1.1
Dissolved oxygen	1992–97	175	61	2.9
Iron – soluble	1984–97	205	147	1.4
Iron – total	1984–97	204	148	1.4
Iron – total	1984–97	185	148	1.3
Manganese – soluble	1984–97	185	148	1.3
Manganese – total	1984–97	204	148	1.4
NH_4	1984–97	194	140	1.4
NO_2	1984–97	205	148	1.4
NO_3	1984–97	206	148	1.4
Odour – cold	1970–97	1326	309	4.3
Odour – hot	1970–97	1259	302	4.2
pH	1984–97	76	75	1.0
Secchi depth	1981–85	176	46	3.8
TKN	1984–97	186	147	1.3
Total dissolved solids	1984–92	92	76	1.2
Total organic carbon	1984–90	67	62	1.1
Total phosphorous	1984–97	205	147	1.4
Turbidity	1971–97	1329	308	4.3
Water temperature	1983–97	630	160	3.9
Phytoplankton				
Anabaena circinalis	1991–97	91	26	3.5
Anabaena spp.	1972–90	32	24	1.3
Ankistrodesmus sp. 1	1988–96	163	39	4.2
Ankistrodesmus sp. 2	1989–97	96	33	2.9
Chlorella spp.	1984–97	50	20	2.5
Dictyosphaerium spp.	1972–97	444	114	3.9
Dictyosphaerium very small sp.	1988–95	24	12	2.0
Microcystis spp.	1975–93	109	48	2.3
Pseudanabaena spp.	1992–97	20	10	2.0
Scenedesmus spp.	1971–97	379	138	2.7

Table 3.11: Myponga reservoir: 10 most abundant phytoplankton species (cells/ml)

Var. name	av. var.	stdev. var.	max. var.
Chlorella spp.	10476	27123	170000
Scenedesmus spp.	7530	18444	139392
Anabaena spp.	5216	24295	150000
Microcystis spp.	3633	12288	114200
Ankistrodesmus sp 1	1674	2862	20160
Ankistrodesmus sp 2	1360	2871	27714
Pseudanabaena spp.	762	2437	10500
Dictyosphaerium very small spp.	749	3121	17136
Dictyosphaerium spp.	642	1568	22102
Anabaena circinalis	588	1488	11400

(Secchi depth, turbidity and colour) and weather data including radiation and wind direction.

The biological database includes total algal biomass (chlorophyll *a* and *b*) and individual numbers of several zooplankton groups. However, compared to other variables, zooplankton data is relatively poorly represented with less than 35 records available from a short time span in the 1970's. *E. Coli* presence is well represented over the entire time-series. The phytoplankton data (see table 3.11) is resolved to genus or species level. It is evident that the Myponga reservoir experiences significant blooms of green algae (*Chlorella spp.* and *Scenedesmus spp.*) and cyanobacteria (*Anabaena spp.* and *Microcystis spp.*). Unfortunately, *Chlorella spp.* and *Anabaena spp.* are not well represented in this database with only 52 and 33 records respectively.

Table 3.10 shows that data availability is highly variable between fields, with some variables measured over a long period of time, such as turbidity and colour, having over 1000 sampling dates. The sampling density ranges from 1 observation per month to 4.5 observations per month in the case of chlorophyll *a*.

3.2.6 Lake Soyang

Lake Soyang is the largest and deepest reservoir in Korea. It is situated approximately 100 km to the north east of Seoul in the far north of South Korea. It was constructed by damming the North Han river in 1973. 90% of the lake inflow comes from the Soyang river. It is of great importance as a supply of drinking water and it has, at times, been host to a net cage fish farming industry (Kim et al., 2000).

The morphometry of Lake Soyang is a complex dendritic structure with one main arm and many branching inlets of up to 5 km in length (see figure 3.10). It is a deep lake with a mean depth of 42m and a maximum depth of 110m. It has a retention time of 0.7 years which is the highest of South Korea's freshwater lakes. Much of the inflow into Lake Soyang occurs in the monsoonal months of July and August, when over 50% of the region's annual rainfall occurs.

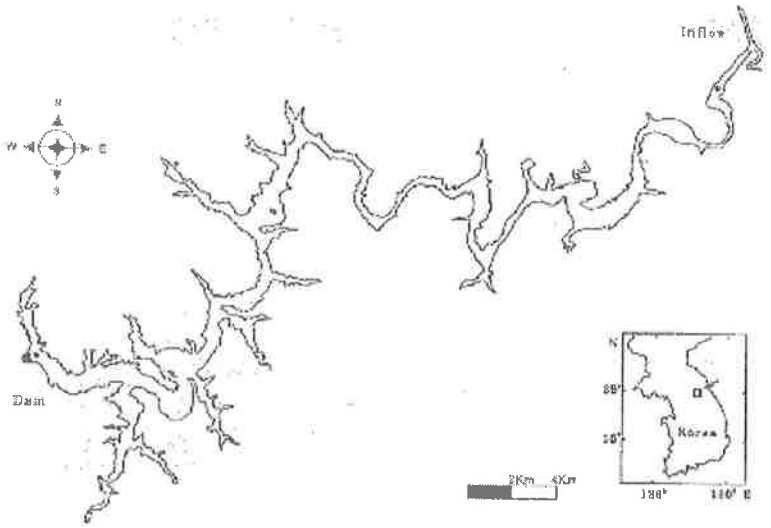

Figure 3.10: Lake Soyang (South Korea).

Nutrient loadings to Lake Soyang originate from non-point agricultural sources and the fish farming industry. By contrast there is an insignificant level of industrial influence on the reservoir. The lake is considered meso-eutrophic, with significant "monsoonal eutrophication" regularly occurring in the summer months as a result of several precipitation events of >100 mmday^{-1} (Kim et al., 2000). Like most Korean lakes, cyanobacteria are the dominant phytoplankton species in Lake Soyang during the summer months (Kim et al., 1997). Kim et al. (1999) observed that by late summer in the years 1996 – 98, large surface crops of *Anabaena* had formed.

Figure 3.11 shows that, like Biwa and Kasumigaura, Lake Soyang is under the influence of a monsoonal climate with high summer temperatures and rainfall. The winters are the coldest of all the lakes investigated with average maximum temperatures in January close to freezing and minimum temperatures well below freezing.

A total of 34 variables were available in the Lake Soyang database measured from 1984 to 2000. Table 3.12 shows that macro-nutrient availability is represented by NO_3 and orthophosphate concentration. No micro-nutrient, pH, or dissolved oxygen data is available, although there is data for electrical conductivity. Amongst the physical variables there is data for underwater light (Secchi depth), temperature and climatic data including rainfall and radiation. As with Lake Burrinjuck,

Table 3.12: Lake Soyang: Sampling Frequency

	Lifespan	Obs. dates	Obs months	Obs. per mo.
Water quality & physical variables				
Alkalinity	1993–00	101	85	1.2
Biological oxygen demand	1987–00	136	122	1.1
Chlorophyll *a*	1984–00	339	183	1.9
COD	1990–97	84	49	1.7
Cosmarium bioculatum (0m)	1984–89	39	39	1.0
Dissolved inorganic phosphorous	1987–00	287	148	1.9
Dissolved organic carbon	1995–99	136	51	2.7
Dissolved oxygen	1987–00	294	158	1.9
Dissolved total phosphorous	1990–00	238	120	2.0
Electrical conductivity	1984–00	351	176	2.0
Elevation	1990–99	3485	116	30.0
Inflow	1984–99	3641	168	21.7
NO_x	1984–89	39	39	1.0
NH_3	1987–97	111	82	1.4
NO_3	1984–00	343	177	1.9
Outflow	1990–99	3550	119	29.8
pH	1987–00	307	155	2.0
POC	1995–00	135	51	2.6
POC/TOC	1995–00	96	39	2.5
Productivity	1987–00	158	143	1.1
Radiation	1984–95	2162	129	16.8
Rainfall	1984–99	3633	178	20.4
Rainfall (chun)	1990–99	3569	118	30.2
Rainfall (inje)	1990–99	3536	117	30.2
Secchi depth	1984–00	360	185	1.9
Soluble reactive phosophorous	1984–95	128	111	1.2
Stratification	1984–95	155	116	1.3
Suspended solids	1993–00	94	73	1.3
TCO2	1993–00	101	85	1.2
Total nitrogen	1988–00	278	133	2.1
Total organic carbon	1995–00	104	41	2.5
Total phosphorous	1987–00	282	134	2.1
Turbidity	1987–00	310	158	2.0
Water temperature	1984–00	355	179	2.0

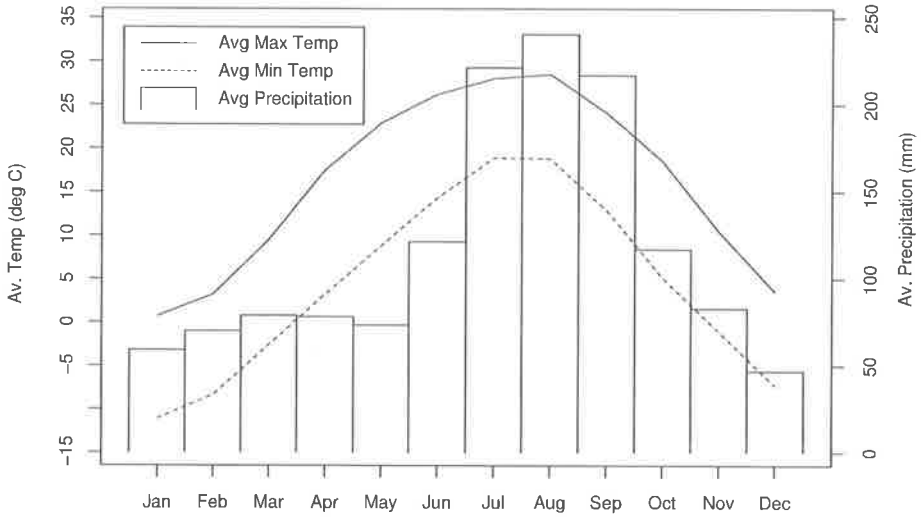

Figure 3.11: Average temperature and precipitation – Lake Soyang

Table 3.13: Lake Soyang: 10 most abundant phytoplankton species: (cells/ml)

Var. name	Var. lifespan	No. obs.	av. var.	stdev. var.	max. var.
Anabaena macrospora	1984–89	39	82	428	3342
Fragilaria crotonensis	1984–89	39	52	440	4747
Asterionella gracillima	1984–89	39	51	139	1064
Microcystis aeruginosa	1984–89	39	42	270	2702
Eudorina elegans	1984–89	39	21	174	1857
Sphaerocystis schroeteri	1984–89	39	19	111	1000
Cosmarium bioculatum	1984–89	39	9	67	561

there is a field describing inflow. This is potentially important for the biological conditions in this reservoir as it is known that "monsoonal eutrophication" in the summer months is directly related to increases in in-flowing water bringing nutrients (Bomchul Kim, pers. comm.).

Total algal biomass of this reservoir is represented by chlorophyll *a* concentration, total number of species observed and a variable of unknown units describing productivity. Phytoplankton data (see table 3.13) is resolved to species level. However its usefulness for modelling purposes is limited by poor representation with only 39 observations present for the time-series. The data indicate that compared to the other reservoirs in this study, Soyang is relatively oligotrophic with low average and maximum cell counts for all species. Cyanobacteria, diatoms and green algae are amongst those present in this reservoir.

Table 3.12 shows that data availability is highly variable, with over 3500 records for rainfall, elevation, inflow and outflow and less than 100 records for 5 fields. Sampling density is between 1 observation per month and 1 observation per day. Chlorophyll *a* was measured 1.9 times per month on average.

3.3 A Comparison of Trophic State

Forsberg and Ryding (1980) recommends a multi-dimensional approach to classifying trophic state where multiple water quality variables are considered. Tables 3.14, 3.15 and 3.16 illustrate classification standards according to Ryding and Rast (1989), Vollenweider and Kerekes (1982) and Forsberg and Ryding (1980). It can be observed that each of these standards considers levels of plant macro-nutrients and algal biomass as indicated by chlorophyll *a* concentration and secchi disc depth. The German classification standard (Ryding and Rast, 1989) also considers pH in the epilimnion and dissolved oxygen levels in the hypolimnion.

Table 3.17 shows the data from each of the 6 water bodies investigated as it corresponds to the variables outlined in the standards in tables 3.14, 3.15 and 3.16. Note that in many cases, missing data prevents assessment using all of the variables available in the standards. However, there is sufficient data available to achieve a reasonable comparison of trophic states in each case. Table 3.18 shows the relevant classifications for the 5 lakes investigated given all the variables in each of the 3 classification systems. It can be seen from this table that a distinct hierarchy of trophic state emerges amongst the case studies with Lake Soyang being the least eutrophic and Kasumigaura the most eutrophic.

Soyang is considered as mesotrophic according to most variables, although there is a very high concentration of inorganic nitrogen. This is possibly a result of the net cage fish farming activities that have been conducted in the lake. The high maximum chlorophyll *a* values and low minimum secchi depths indicates that the lake has, at times, been subjected to substantial algal blooms. Myponga is the next

Table 3.14: German lake classification standard (after Ryding and Rast (1989))

Criterion	Quality class					
	1	2	3a	3b	4	5
Orthophosphate (mg/L)	0–0.002	0–0.005	0–0.1		> 0.1	> 0.5
DIN (mg/L)	≤ 0.01	≤ 0.03	≤ 0.1		> 0.1	> 0.5
Chlorophyll *a*∗ (μg/L)	≤ 3	< 10	10–20	20–40	40–60	> 60
Secchi depth (m)	≥ 6	≥ 4	≥ 1		≥ 0.5	< 0.5
pH – epilimnion	6.5–8	7–8.5	7–9	7–9.5	6.5–10	6–11
O_2 (mg/L) – hypolimnion	≥ 6	≥ 1	anaerobic		n/a	
Trophic state	oligo–	meso–	eu–strat.	eu–unstrat.	poly–	hyper–

∗ Average in warmer 6 months of year

Table 3.15: OECD lake classification standard (after Vollenweider and Kerekes (1982))

Trophic category	TP	mean chl	max chl	mean secchi	min secchi
Ultra–oligotrophic	< 4.0	< 1.0	< 2.5	> 12.0	> 6.0
Oligotrophic	< 10.0	< 2.5	< 8.0	> 6.0	> 3.0
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

note: secchi refers to secchi depth (m)
chl refers to chlorophyll *a* (μg/L)
TP refers to in–lake total phosphorous (μg/L)

Table 3.16: Trophic levels according to Forsberg and Ryding (1980)

Trophic State	Chlorophyll mg/m^3	Transparency m
Oligotrophic	< 3	> 4.0
Mesotrophic	3–7	2.5–4.0
Eutrophic	7–40	1.0–2.5
Hypertrophic	> 40	< 1.0

Table 3.17: Observed water quality

Variable		Biwa	Burr.	Darl.	Kasu.	Mypo.	Soya.
Chl *a* (μg/L)	mean	9.32	15.8	n/a	60.5	7.45	4.30
	max	38.5	579.0	n/a	280.0	41.6	98.0
	summer	9.47	15.3	n/a	75.6	7.34	6.21
PO_4 (mg/L)	mean	0.00319	0.0210	0.157	0.0109	0.0180	0.00348
TP (mg/L)	mean	n/a	0.622	0.317	0.0995	0.0637	n/a
NO_3 (mg/L)	mean	0.102	0.541	0.157	0.463	0.125	0.925
TN (mg/L)	mean	n/a	1.71	n/a	1.42	n/a	n/a
NO_3/PO_4	ratio	31.2	25.8	1.0	42.3	6.94	266.0
Secchi (m)	mean	1.82	1.36	n/a	0.81	1.85	3.71
	min	0.70	0.12	n/a	0.20	0.90	0.40
pH	mean	7.3	n/a	7.9	8.6	n/a	n/a

most eutrophic lake, with meso-eutrophic scores according to most classifications. Biwa can be considered a little more eutrophic than Myponga with a greater number of eutrophic scores. Burrinjuck is eutrophic to hypertrophic with an equal number of each of these scores, while Lake Kasumigaura is clearly suffers the worst water quality of all the lakes investigated being considered hypertrophic by most classifications. The average NO_3/PO_4 ratios show that the Darling River and Myponga reservoir are generally nitrogen limited and that Lake Soyang and to a certain extent, Lake Kasumigaura, are likely to exhibit phosphorus limitation.

Figure 3.12 compares the distributions of chlorophyll *a* data for the 5 lakes using box and whisker plots (Tukey, 1977; McGill et al., 1978) to represent the distributions of observations (see appendix B for a more complete description of box-and-whisker plots.). The data expressed in this plot takes into account all measuring sites and dates for each lake. It can be seen that Lake Kasumigaura has by far the highest chlorophyll *a* concentration of all the lakes as indicated by the fact that the lower quartile is aligned with the maximum of the ranges observed for the other studies. Also, this plot shows that while Lake Burrinjuck has a similar median chlorophyll *a* levels to the other lakes in this study, the large number of circles above the box-plot indicate a number of outliers in this data. It is likely these are the result of occasional severe algal blooms leading to chlorophyll *a* levels much higher than normal.

Figure 3.13 shows a plot of chlorophyll *a* levels where Kasumigaura data was excluded. Also, all outliers above 35 μg/ml were excluded to reduce the range of the *y*-axis in order to gain a clearer comparison. It can be seen that lake Soyang has the lowest chlorophyll *a* levels as indicated by the median values, interquartile

Table 3.18: Trophic state classifications

	Biwa	Burrinjuck	Kasumigaura	Myponga	Soyang
German Ryding and Rast (1989)					
PO_4	mesotrophic	eutrophic	mesotrophic	mesotrophic	mesotrophic
DIN∗	eutrophic	hypertrophic	polytrophic	polytrophic	hypertrophic
Summer Chl *a*	mesotrophic	eutrophic	hypertrophic	mesotrophic	mesotrophic
Secchi	eutrophic	eutrophic	polytrophic	eutrophic	mesotrophic
pH	oligotrophic	n/a	$\geq$ eutrophic	n/a	n/a
OECD Vollenweider and Kerekes (1982)					
TP	n/a	hypertrophic	eu–hypertrophic	eutrophic	n/a
Chl *a*	eutrophic	eutrophic	hypertrophic	mesotrophic	mesotrophic
Max chl *a*	eutrophic	hypertrophic	hypertrophic	eutrophic	hypertrophic
Secchi	eutrophic	hypertrophic	hypertrophic	eutrophic	mesotrophic
Min secchi	eutrophic	hypertrophic	hypertrophic	eutrophic	hypertrophic
Forsberg and Ryding (1980)					
Chl *a*	eutrophic	eutrophic	hypertrophic	eutrophic	mesotrophic
Secchi	eutrophic	eutrophic	hypertrophic	meso–eutrophic	mesotrophic

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

ranges and total ranges. The remaining lakes have more similar productivity ranges, but it is clear that Myponga is less eutrophic in general than Lake Biwa and Burrinjuck Dam as it has significantly lower median chlorophyll *a* as indicated by the non-overlapping notches of the box-plots and a comparatively reduced lower quartile value.

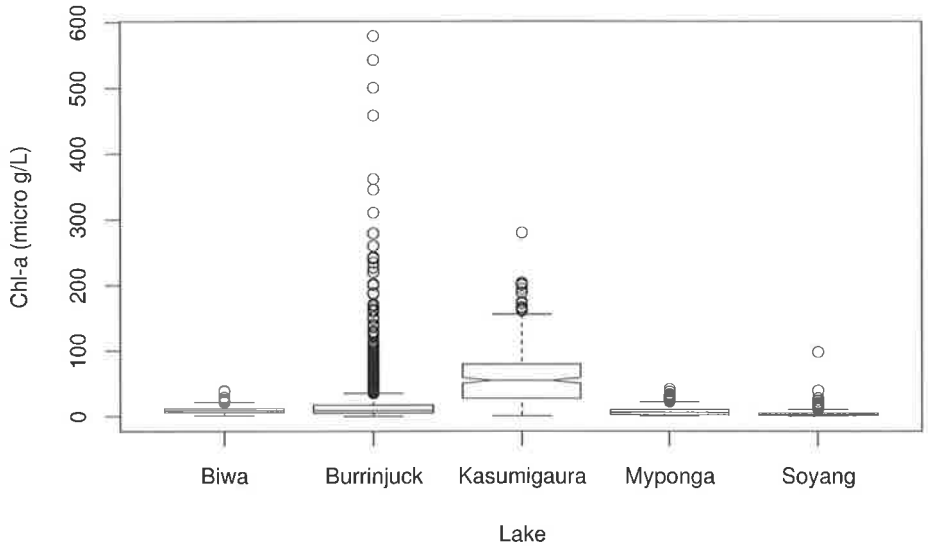

Figure 3.12: Total algal biomass – a comparison of 5 lakes

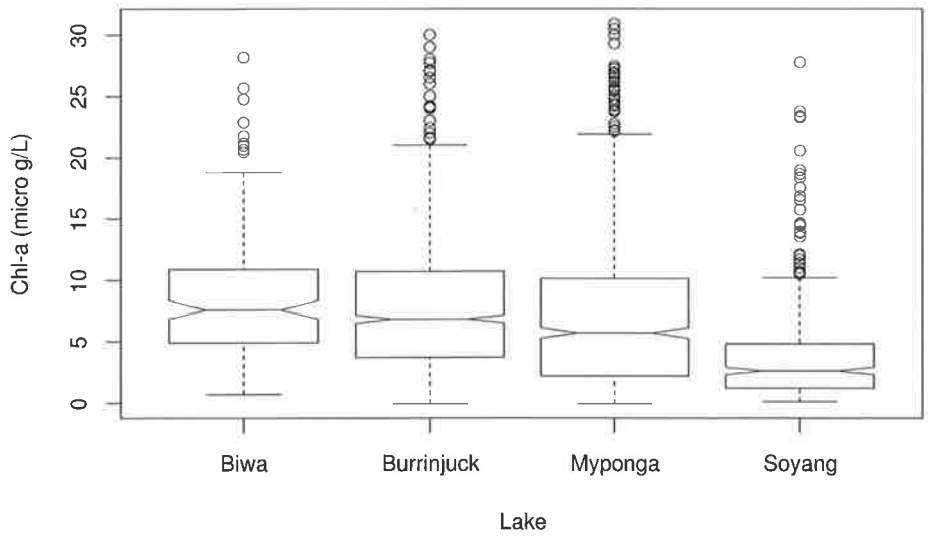

Figure 3.13: Total algal biomass – a comparison of 4 lakes

The comparison of the distributions of secchi disk depth data illustrated in figure 3.14 shows that Soyang enjoys by far clearest water with higher median and interquartile ranges. However, it can be seen that the secchi depth data for this lake covers a large range indicating high variability of trophic state over time.

Biwa and Myponga have the next highest readings with median values falling between 1.5 and 2 m and reasonably compact interquartile ranges. Burrinjuck has median values of approximate 1 m, while Kasumigaura has the least transparent water with a median value between 0.5 and 1 m.

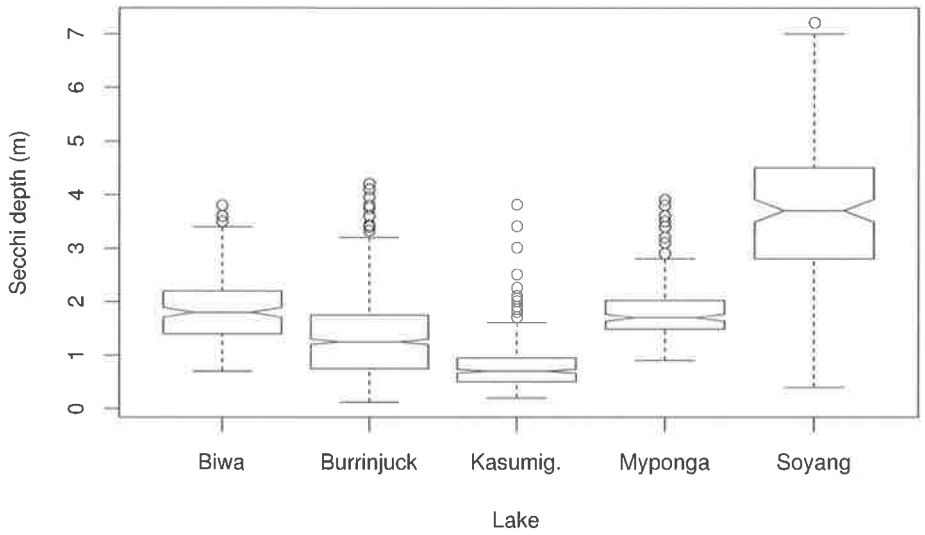

Figure 3.14: Secchi disk depth – a comparison of 5 lakes

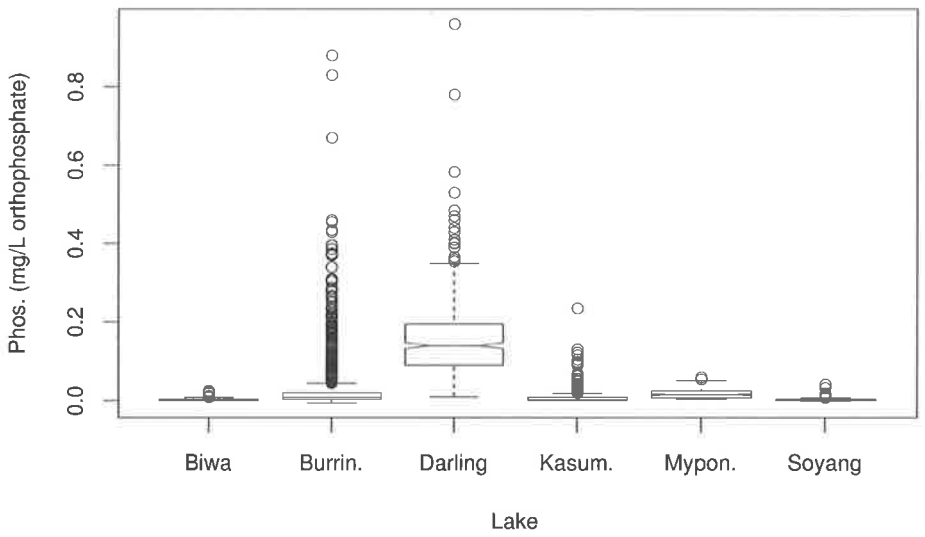

Figure 3.15: Phosphorous concentration – a comparison of 6 lakes and 1 river

Figure 3.15 compares the 6 case studies with regards to orthophosphate concentration. This comparison shows that the Darling River has levels of bioavailable phosphorus that are an order of magnitude higher than any of the lakes. For purposes of clarity, figure 3.16 shows the orthophosphate data with the Darling River excluded and the y-axis limited to a maximum range of 0.055 mg/L. This

plot shows a distinct hierarchy of orthophosphate levels with Myponga being the highest, followed by Burrinjuck, Kasumigaura, Biwa and finally Soyang.

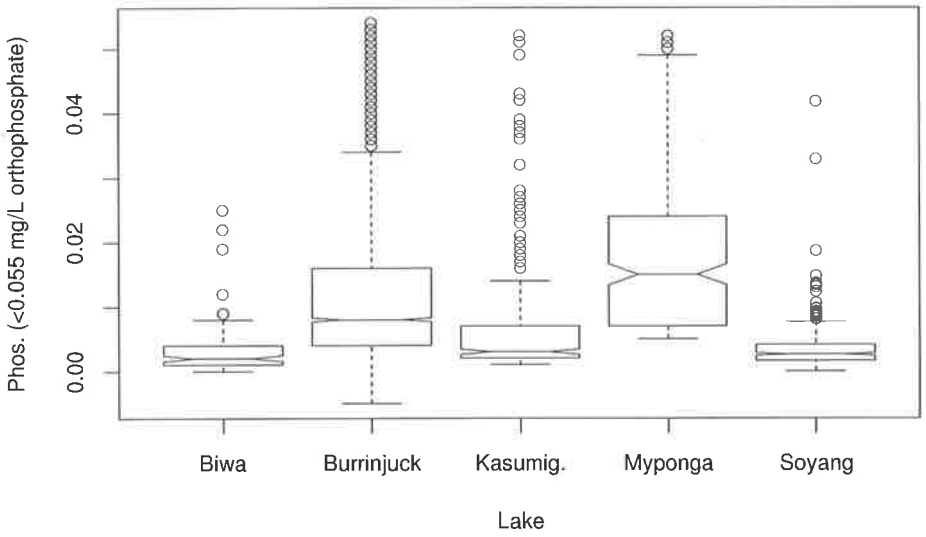

Figure 3.16: Phosphorous concentration – a comparison of 5 lakes

Figure 3.17 compares nitrogen levels in the 6 case studies. Nitrogen was expressed as mg/L NO_3 for all lakes except Burrinjuck and the Darling River where it was expressed as mg/L NO_x. Since it is probable that total oxidised nitrogen is dominated by the nitrate faction, it is reasonable to compare these units. There are a large number of high outliers in the data for lake Burrinjuck, so in the interests of clarity, the y-axis was limited to a range of 2 mg/L. It can be seen that Lake Soyang has significantly higher nitrogen levels than the other case studies. Burrinjuck and Kasumigaura are the next highest. Biwa, Darling and Myponga have significantly lower concentrations of nitrogen than the other case studies.

3.3.1 Discussion and Conclusions

According to the classification systems, it is clear that Lake Kasumigaura and Burrinjuck can be considered, on average, hypertrophic water bodies. However examination of the box-plot of chlorophyll *a* values for Lake Burrinjuck shows that the average value for this lake is probably somewhat skewed by a number of extremely high values. Therefore, it is likely that the eutrophication in Burrinjuck is of a more intermittent, "acute" nature than the more "chronic" conditions observed in Kasumigaura. The data shows that these two lakes share a number of features contributing to eutrophication. They both tend towards P limitation, although they have relatively high orthophosphate concentrations. They are also have reasonably high nitrate concentrations as well. With an average depth of 4m, Lake Kasumigaura is very shallow meaning that it may suffer from significant internal nutrient loadings as a result of the high sediment area to volume ratio and

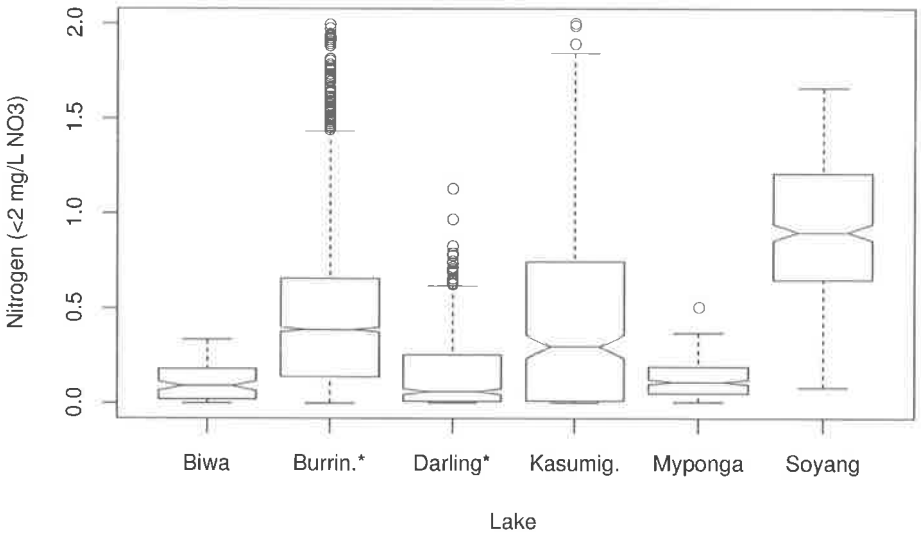

Figure 3.17: Nitrogen concentration – a comparison of 5 lakes and 1 river

(* mg/L NO_x)

a low dilution factor. While Burrinjuck Dam has a deeper mean depth of 57 m, it is likely to be significantly shallower at the end of summer as a result of irrigation draw down and dry conditions. Thus it can be hypothesised that the intermittent extreme eutrophication in this lake is a result of occasional shallow conditions and high temperatures intensifying the effect of internal and external nutrient loadings.

Lake Biwa and Myponga Reservoir are somewhat less productive with a meso-eutrophic classification according to most standards. Both of these reservoirs have lower NO_3 concentrations than the two hypertrophic lakes and Biwa has lower PO_4 levels as well. Myponga is most likely N limited for much of the time, while Biwa is more likely to be P limited. Interestingly, while Myponga has higher concentrations of macro-nutrients than Biwa and has a lower mean depth and volume, it is generally classified as being a little less eutrophic than Lake Biwa with lower chlorophyll *a* concentrations. A possible reason for this observation is that Myponga reservoir is more intensively managed with the aim of deterring algal growth with regular $CuSO_4$ doses and aeration. Also, a comparison of climate conditions in these two lakes (see figures 3.2 and 3.9) shows that Myponga has lower summer temperatures and precipitation than the lake Biwa region meaning i) cyanobacteria may not be as well favoured and ii) less nutrient inflow in the summer months to fertilise algal blooms.

Lake Soyang is clearly the least eutrophic lake to be included in this study with a mesotrophic classification according to most indices. The nutrient data clearly shows that algal growth is likely to be severely P limited. However, the classifications and the plots of secchi depth and chlorophyll *a* concentrations show that this lake suffers occasional periods of intense eutrophication. This may be a result of nutrient inflow as a result of strong monsoonal rains in the summer months.

Another feature of this reservoir is the very high NO_3 concentrations. This may arise from fish farming activities that were carried out in this water body over certain periods and also a lack of flux into the organic N pool as a result of the strong P limitation.

The Darling River, while not being classified according to the standards in tables 3.14, 3.15 and 3.16, can probably be considered highly eutrophic with average and peak total algal cell counts of 20100 and 281000 cells/ml respectively. It is worth noting that table 3.17 and plot 3.15 shows that the PO_4 levels in this river are in general an order of magnitude higher than those experienced in the lakes. This is reasonable given that a river has a very short comparative residence time meaning that P will not be lost as a result of sedimentation processes. Also, the Darling River has a vast catchment compared to the lakes in this study that includes numerous point and non-point sources of P such as towns, industries, intensive agriculture and areas of significant soil erosion. Furthermore the lower reaches of the Darling river system are in highly arid regions with high evaporation rates compounding problems of P loading and salinity.

3.4 Model Design

3.4.1 A Generic ANN Model Design

Choice of inputs to ANN models can be seen as a compromise between hypotheses of causal relationships between available data and the output variable/s and the abundance of records for likely input and output variables that are matching in time. The analysis of the water quality databases in this chapter revealed there is abundant data in all case studies for variables describing bioavailable phosphorus and nitrogen, water transparency and temperature. These parameters are known to be important driving variables for algal growth and are widely used as independent variables in other eutrophication models (eg Benndorf and Recknagel (1982); French and Recknagel (1994); Maier et al. (1998)). Wilson and Recknagel (2001) suggested these variables form a generic ANN structure that is widely compatible with available data.

The "generic ANN model template" is illustrated in figure 3.18 showing these 4 inputs and a "feedback" input comprising the output variable in each case. Table 3.19 shows the actual input layers used for each dataset. Note that there are differences in the expression of nitrogen availability and transparency between different case studies, with NO_x being used instead of NO_3 for the Burrinjuck and Darling models and turbidity being used instead of secchi disk depth for the Darling and Myponga models. These differences are result of variations in data availability. However, it must be stressed that the "deterministic intent" of the inputs sets remains the same for all datasets. In addition to the generic model inputs, flow was added as an input to the Darling model since it has been shown

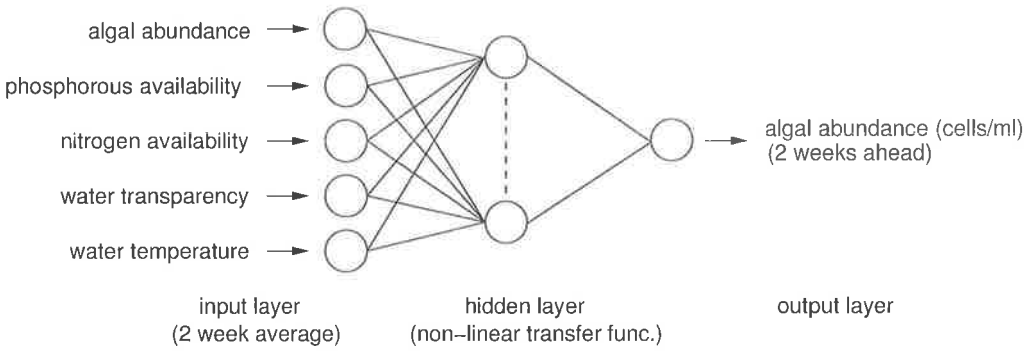

Figure 3.18: Generic ANN design for 2 week forecasts of algal abundance

Table 3.19: Input layers for the generic model

Site	Inputs
Biwa	PO_4, NO_3, secchi depth, water temp.
Burrinjuck	PO_4, NO_x, secchi depth, water temp.
Darling	PO_4, NO_x, turbidity, water temp., flow
Kasumigaura	PO_4, NO_3, secchi depth, water temp.
Myponga	PO_4, NO_3, turbidity, water temp.
Soyang	PO_4, NO_3, secchi depth, water temp.

in the ANN modelling work reported by Recknagel et al. (1997), Maier et al. (1998) and Jeong et al. (2001) to be an important factor influencing phytoplankton productivity in rivers.

As proposed in section 2.3.1, time-series dependencies are modelled using the TDNN structure, whereby input variables lag the output by the forecast interval. The proposed "input window" approach to defining model inputs described in section 2.5.1 is utilised meaning that inputs to the ANN were average values of each of the driving variables falling between 2 and 4 weeks prior to the output date. The summary period was chosen to maximise compatibility with the 6 datasets available for this study.

It is proposed that models be developed predicting a measure of overall algal abundance and the 3 most abundant algal species/functional groups in terms of average cell counts. Chlorophyll *a* concentration in μg/l is used to indicate total abundance for all datasets except the Darling River, where the total algal cell count is used instead. The remaining 3 outputs for each dataset are listed in table 3.20. Note that species abundance models are developed for Lake Biwa, genera abundance models are developed for Lake Kasumigaura and Myponga Reservoir and functional group abundance models are trained for Burrinjuck Dam and the

Table 3.20: Model outputs

Site	Outputs
Total abundance	
All except Darling	chlorophyll a
Darling	total cell count
Site specific outputs	
Biwa	*Euglena americana, Melosira granulata, Pediastrum Biwae*
Burrinjuck	Cyanophyta, Chlorophyta, Diatoms
Darling	Cyanophyta, Chlorophyta, Flagellates
Kasumigaura	*Microcystis spp., Oscillatoria spp., Gomphosphaeria spp.*
Myponga	*Ankistrodesmus spp., Scenedesmus spp., Dictyosphaerium spp.*
Soyang	n/a

Darling River. In the case of Lake Soyang, there was only data available for a total abundance model. Table 3.21 compares the number of records available for modelling using the generic model definition outlined.

A generic model design may have the following benefits;

- Meaningful comparisons between models developed for different lakes using established elucidation techniques for ANNs such as sensitivity analysis.
- It provides scope for increasingly general ANN eutrophication models trained with data aggregated from many lakes or classes of lakes.
- It potentially reduces modelling effort through rationalisation of monitoring, database and modelling approaches.

3.4.2 Case Specific ANN Model Design

It is reasonable to hypothesise that, for best performance, input layers should be specific to both output and dataset and that all variables known to have a causative relationship with the output variable should be considered. Furthermore, co-predictors with a correlative relationship with modelled variables may also be useful, as may be multiple lags of certain variables.

The choice of variables for inclusion as inputs in dataset specific models is constrained by the availability of data with matching measurement dates to the output variable. Tables 3.2, 3.4, 3.6, 3.8, 3.10 and 3.12 compare the variables measured and data availability for each variable for the 6 datasets used in this study. These tables show that different sets of variables were measured in each case meaning that, unlike the generic model, specific models will be dataset specific. Additionally, these tables compare the time spans of measurement, the number of discrete

Table 3.21: Sampling frequency.

dataset	output	count	median	mean	min	max
			(days between samples)			
Biwa	chlorophyll *a*	88	22	31.6	7	119
	Euglena americana	112	15	25.4	6	77
	Melosira granulata	112	15	25.4	6	77
	Pediastrum biwae	112	15	25.4	6	77
Burrinjuck	chlorophyll *a*	85	28	76.8	6	1084
	chlorophyta	87	28	75.0	6	1084
	cyanophyta	84	28	77.7	6	1084
	diatoms	82	28	79.6	6	1084
Darling	total phytoplankton	508	7	8.8	3	224
	chlorophyta	508	7	8.8	3	224
	cyanophyta	508	7	8.8	3	224
	flagellates	508	7	8.8	3	224
Kasumigaura	chlorophyll *a*	70	55	51.7	14	182
	Gomphosphaeria spp.	87	28	43.4	1	182
	Microcystis aeruginosa	87	28	43.4	1	182
	Oscillatoria spp.	87	28	43.4	1	182
Myponga	chlorophyll *a*	428	7	10.1	1	139
	Ankistrodesmus spp.	108	6	11.3	2	129
	Dictyosphaerium spp.	213	5	22.2	1	550
	Scenedesmus spp.	136	7	34.7	1	431
Soyang	chlorophyll *a*	222	8	22.1	2	634

Table 3.22: Input summary periods for different sampling densities.

Av. obs per mo.	Summary length	Summary lag period
≤ 3.0	30 days	7 – 37 days
≤ 8.0	14 days	7 – 21 days
> 8.0	7 days	7 – 14 days
all	60 days	7 – 67 days

measurement dates and months and the average data density in months where measurements took place. Comparison of these properties permits determination of the "time compatibility" of each variable with the output variable, where time compatibility is defined as agreement to the following points;

- Start and end date of variable lifespan overlaps that of output variable.
- No observation months is $\geq$ no. observation months of output.
- Few months with missing data.

In terms of of the input summary definition, it was elected to define the end of all input windows 7 days prior to the output date making the model capable of 7 day forecasts. However, a question arises regarding how long the input window should be to achieve reasonable representation of data. The answer to this question depends on the sampling density with respect to each input. The column showing the average number of samples per month in tables 3.2, 3.4, 3.6, 3.8, 3.10 and 3.12 show that, in months when data was collected, sampling frequency varied from once per day to once per month. Thus, in order maximise data representation, the length of the input window needs to be specific to each variable to account for differences in data density. Table 3.22 shows the policy used to choose window lengths.

It was elected to utilise the maximum possible number of variables as model inputs for each of the respective datasets. This means that many variables may be included that would not normally be considered by a strictly deterministic modelling approach. However, it is reasonable to hypothesise that such variables may be considered *co-predictors* (Scardi, 2001), since they were collected with the intention of gaining insight into water quality in each case. Tables J.1 to J.6 show the starting models for each dataset given the model selection policies outlined. These models are characterised by a number of noteworthy features;

- The starting models have large input layers compared to previously published ANN models (see table 3.23).
- Variables describing abundance of algal species or functional groups are included in models developed for Biwa, Burrinjuck, Darling and Kasumigaura datasets to investigate hypotheses of competition and/or mutualism.

Table 3.23: Comparison of starting model input layer sizes

Dataset	No. inputs
Lake Biwa	40
Burrinjuck Dam	68
Darling River	60
Lake Kasumigaura	70
Myponga Reservoir	38
Lake Soyang	28

- Variables describing inflow volume are present for models developed for the Burrinjuck and Soyang datasets. This may have a bearing on the water residence time and thus the effect of flushing on the lake ecosystem.
- Variables describing zooplankton abundance are included in models developed for Lake Kasumigaura. Zooplankton exert top down control of phytoplankton abundance through grazing.
- Variables describing concentration of silica are present in the Biwa, Darling and Kasumigaura datasets. Availability of silica is thought to have an effect on the biology of diatoms.
- Variables describing various aspects of weather information are included in models developed for Biwa, Burrinjuck, Kasumigaura and Soyang. The weather affects light availability for photosynthesis, the mixing conditions in the water body and the likelihood of significant inflow events.
- Variables describing concentrations of a number of heavy metals and trace elements are included in models developed for the Myponga dataset. Of particular interest in this case is data for copper concentration, since regular $CuSO_4$ dosing for control of phytoplankton abundance is a feature of this reservoir.
- pH is included as an input for models developed for the Biwa, Darling, Kasumigaura and Soyang datasets. pH is known to be correlated to the structure of phytoplankton communities, with cyanobacteria favouring high pH conditions (Reynolds, 1984; Harris, 1986; Shapiro, 1990).

Table 3.24 compares the data availability for each output given the input layers defined in tables J.1 to J.6. As with the generic model (see table 3.21), the data availability in terms of the total number of records available and the time interval between samples varies between datasets and outputs. The Darling River models enjoy the most abundant data with 388 records per model, while the Burrinjuck models only have $\approx$ 100 records per model available for training and validation. The median sampling interval shows that, when data is available, sampling was carried out approximately monthly in Burrinjuck Dam and Lake Kasumigaura,

Table 3.24: Sampling frequency.

dataset	output	count	median	mean	min	max
			(days between samples)			
Biwa	chlorophyll *a*	146	15	19.8	7	37
	Euglena americana	183	14	15.8	6	28
	Melosira granulata	183	14	15.8	6	28
	Pediastrum biwae	183	14	15.8	6	28
Burrinjuck	chlorophyll *a*	99	28	55.4	7	448
	chlorophyta	102	27	53.8	6	448
	cyanophyta	99	27	55.4	6	448
	diatoms	100	27	54.9	6	448
Darling	total phytoplankton	388	7	11.3	3	273
	chlorophyta	388	7	11.3	3	273
	cyanophyta	388	7	11.3	3	273
	flagellates	388	7	11.3	3	273
Kasumigaura	chlorophyll *a*	89	29	39.6	14	154
	Gomphosphaeria spp.	111	28	31.8	2	147
	Microcystis aeruginosa	111	28	31.8	2	147
	Oscillatoria spp.	111	28	31.8	2	147
Myponga	chlorophyll *a*	519	7	8.1	1	105
	Ankistrodesmus spp.	112	6	10.9	2	105
	Dictyosphaerium spp.	243	4	17.2	1	510
	Scenedesmus spp.	155	7	27.0	2	352
Soyang	chlorophyll *a*	187	11	18.6	2	251

fortnightly in Lake Biwa and weekly in the Darling River, Myponga Reservoir and Lake Soyang. Comparison of the median, mean, minimum and maximum sampling intervals shows that there are large gaps, or holes in the time-series for some models. In the case of Myponga, the data have low median sampling intervals of ≤ 7 days, but relatively high average and maximum sampling intervals indicating frequent sampling indispersed with long breaks.

3.5 Conclusion

Six datasets are available for the present study. Three originated from Australia, two from Japan and one from South Korea. This chapter showed that these datasets vary significantly in terms of climate, situation, catchment and morphometry. Also, an analysis of the data showed wide variation in trophic state. Lake Soyang was demonstrated to be mesotrophic, Lake Biwa and Myponga

Reservoir are meso-eutrophic, while Burrinjuck Dam, the Darling River and Lake Kasumigaura are eutrophic to hypertrophic.

Model design for ANNs depends on the hypothesised relationships (either correlative or causative) between variables and data availability. It was found that the datasets varied considerably in terms of the variables measured, the frequency of measurement and the time over which monitoring took place. In terms of outputs, it was elected to specify a measure of total phytoplankton abundance, such as chlorophyll *a* and the three most abundant species/genera/functional groups for each dataset.

With respect to inputs, it was elected to define two input layers as starting points for model design. The *generic* model, comprising variables describing bioavailable nitrogen and phosphorus, water temperature, secchi disk depth and a lag input for algal abundance, is compatible with all six datasets. The *specific* model represents the maximum possible input parameterisation for each dataset assuming an input-window TDNN structure that will achieve reasonable (ie $>$ 80 records) representation.

Chapter 4

Model Complexity and Bootstrap Aggregation

4.1 Introduction

In chapter 2, it was found that generalisation of ANN models predicting phytoplankton abundance is usually "tuned" by selection of parameters affecting model complexity – for example, hidden layer configuration, training time, jitter, weight decay or other parameters that influence the number and weight of network connections. However, it was concluded in section 2.4.6.3 that optimisation of these parameters is a difficult task meaning that, in practice, many applications may be characterised by sub-optimal performance. According to Maier and Dandy (2000), there is a lack of clear guidelines with respect to configuration of these and other aspects of ANN training.

In section 2.5.2, it was proposed that adoption of *bagging* may be a solution to this problem, since it reduces the variance component of model generalisation error associated with overfitting. Thus it can be proposed that application of bagging minimises the total (bias + variance) model error, providing the following conditions have been satisfied;

- There is sufficient ANN model "parameterisation" in terms of hidden layer units for the problem at hand,
- The learning algorithm is able to approximate a reasonable mapping of the training data and
- A sufficiently large number of "bootstrap models" are trained in order to simulate the probability distribution underlying population data.

If bagging is proven to be effective at minimising generalisation error for an ANN modelling problem, there is no longer a need for analytic or empirical model "complexity tuning". This will lead to greater confidence in modelling

outcomes and improved data efficiency, since there is no need to perform "double cross-validation" where a second validation set is held out from training to tune complexity parameters.

Also, it was concluded in section 2.4.6.3 that the training algorithm used may influence the generalisation properties of ANN models. Lawrence and Giles (2000) suggested that less efficient algorithms, such as backpropagation, may actually result in better performing ANN models on independent data than so-called "second order" algorithms because inefficiencies in approximation may result in a smoother, more *regularised* model.

This chapter presents the results of experiments designed to determine the effects of the following on validation performance of ANN models;

- Number of hidden layer units.
- Stopping error of training.
- First order versus second order training algorithms.
- Validation methodology.

Thus the aim of this work is to obtain data leading to specific recommendations regarding each of these parameters. In particular, it is hypothesised that application of bagging reduces model sensitivity to these parameters thus eliminating the requirement for cross-validation to determine an optimum ANN configuration for a modelling task.

4.2 Methods

4.2.1 Model Inputs and Outputs

7 models were specified for forecasting phytoplankton abundance. Table 4.1 summarises the design of each of the 7 models with respect to inputs, outputs, the site from which data was retrieved, the number of inputs and the number of records available for training and validation. The "generic" input layers refer to the generic model template discussed in chapter 3 and illustrated in figure 3.18. The models that only consider the generic structure are supported by all six datasets available for this study. The 10 input models (ie generic + input 1 ... input n) represent designs that are supported only by data from their respective sites.

Four models were developed predicting chlorophyll *a* abundance, while the remaining three models predict occurrence of either phytoplankton genera or species. Chlorophyll *a* was expressed as μg/l while the genera and species were expressed as cells/ml. The range of models were chosen to highlight interactions between

Table 4.1: Model designs

Model	Site, Output and Inputs	No. In	No. dat.
1	site = Lake Kasumigaura (site 9) output = chlorophyll *a* inputs = generic	5	82
2	site = Myponga Reservoir output = chlorophyll *a* inputs = generic	5	606
3	site = Lake Soyang output = chlorophyll *a* inputs = generic	5	87
4	site = Lake Kasumigaura (site 3) output = *Microcystis aeruginosa* inputs = generic + pH, DO, Si, Rotifers, Cladocera, Copepoda	10	125
5	site = Lake Kasumigaura (site 9) output = *Oscillatoria spp.* inputs = generic + pH, DO, Si, Rotifers, Cladocera, Copepoda	10	72
6	site = Lake Kasumigaura (site 3) output = chlorophyll *a* inputs = generic + pH, DO, Si, Rotifers, Cladocera, Copepoda	10	102
7	site = Myponga Reservoir output = *Scenedesmus spp.* inputs = generic	5	228

model design, study site and the experimental treatments. However, due to restrictions in data processing availability, it was not intended to conduct experiments for all datasets/models outlined in chapter 3.

Note that data from two sampling sites (stations 3 and 9) were retrieved for the Lake Kasumigaura models.

The *input window* input representation in combination with the TDNN model structure is used for all models (see section 2.5.1). Thus, each input is defined as the average values falling within a time window between a start and end lag with respect to the output variable. The "window end" date lags the output date by 14 days meaning that the models are trained to forecast algal biomass 2 weeks in advance. The inputs were defined as the average conditions over a 30 day summary period prior to the window end date meaning that the window start date lags the output date by 44 days.

4.2.2 Model Inference

It can be concluded from the review in section 2.3 that the prerequisites for reasonable model approximation given training data using ANNs are;

1. A training algorithm capable of adapting the ANN structure such that prediction error can be minimised within a reasonable time period.
2. Conditioning of data such that it is compatible with the ANN structure and learning algorithm.
3. Sufficient "degrees of freedom" in the ANN structure relative to the state space of the problem permitting the mapping of appropriate decision boundaries.

The following provides a brief review of the choices arrived at for each of these prerequisites.

4.2.2.1 Training Algorithms

Ideally, training algorithms should have the following qualities;

- Highly efficient convergence properties allowing rapid training with minimal computational overhead.
- Stability, meaning the ability to converge to a reasonable mapping regardless of initial connection weights (Alpsan et al., 1995).
- Robustness, meaning adequate convergence properties given a range of technique related meta-parameters and database dimensions (Alpsan et al., 1995).
- Capable of producing models with good generalisation characteristics.

There are too many training algorithms in existence claiming to meet these requirements to permit an exhaustive search. Preliminary experimentation using the algorithms available in the Stuttgart Neural Network Simulator (SNNS) software package (University of Stuttgart, 1999) reached a number of conclusions. Firstly, it was found that batch mode backpropagation (see section 2.3.2.2) was effective at learning training sets, but was much slower than incremental mode backpropagation. This was also the conclusion of Alpsan et al. (1995). Also, it was shown that novel approaches to training such as QuickProp (Fahlman, 1988), RProp (Riedmiller and Braun, 1992; Braun and Riedmiller, 1992) and scaled conjugate gradients (SCG) (Møller, 1993) are efficient training algorithms that can reach a target error rate on training data significantly more quickly than backpropagation. Furthermore, it was shown that the SCG approach can perform effectively without the need to tune any algorithm related "meta-parameters". On the basis of this preliminary work, it was elected to compare incremental mode backpropagation and SCG for all model training conducted in the present study.

Where backpropagation was applied, it was used in incremental weight update mode with a learning rate of 0.15 and a momentum of 0.9, which was found to achieve reasonable mappings of training data in preliminary experiments. By contrast the SCG algorithm requires no such parameters.

4.2.2.2 Numerical Conditioning

The review presented in section 2.3.2.1 shows that "numerical conditioning" can have a significant influence on the convergence properties of ANNs. Preliminary experimentation found that the convergence properties of ANNs was greatly enhanced by conditioning the input data such that each input had a mean of 0 and a standard deviation of 1. Training times decreased dramatically and the accuracy of the converged models was much improved. Therefore, input data was standardised in this way prior to training all models. Similarly, all output data was scaled to lie within the range of the activation functions of the hidden units.

4.2.2.3 Hidden Layer Configuration

As discussed in section 2.3, specification of ANN hidden layer neurons are the generally accepted approach to achieving a structure capable of mapping training data, since Hornik (1993) showed that an ANN with a single hidden layer with a sufficient number of units is capable of mapping any continuous input-output function within given distortion criteria. Preliminary experiments using a variety of data and hidden layer sizes showed that, in the context of the data and model designs implemented in the present study, 20 hidden layer units in a single layer was sufficient for achieving model approximations with negligible error on training sets. Therefore, 20 hidden layer units was set as the upper limit for network

size for all models trained in this study. 0 hidden unit ANNs (ie perceptrons) are the functional equivalent of a multiple linear regression, since they may only map linear decision boundaries (Cheng and Titterington, 1994). Therefore, it was elected to train 0 hidden unit models as controls for all experiments.

Why Not use Recurrent Connections? RNNs utilise activations from the hidden to output layer as an extra set of inputs. Such recurrent connections are argued to enhance ANN performance at modelling time-series because they learn a hidden "temporal context" or time delay. It has been shown by Jeong et al. (2001), Walter et al. (2001) and Jeong et al. (2003) that RNNs are effective for time-series modelling of phytoplankton dynamics. However, maintaining the integrity of the temporal context learned by the recurrent connections requires that training pairs be processed in time order. In the context of the present study, the sequence of training data may be disturbed by;

- implementation of bagging, where ANNs are trained on bootstrap samples of training pairs from the time-series and
- uneven sampling intervals and missing values in the datasets.

Thus it was elected *not* to use recurrent connections and rely only on explicit links between past and present states defined by the lags between input and output data.

4.2.3 Model Validation

The blocked bootstrapped 20-fold cross-validation performance estimator was utilised for all experimental treatments, except where the leave-one-out bootstrap estimator was used to determine the effect of validation method on model performance. The methodology for each of these approaches is outlined in section 2.5.3. To determine the bagged model performance, predictions of member ANNs (ie individual bootstrap ANNs) were aggregated by averaging and the aggregate prediction compared with observed values. Observations of model performance constituted the following;

1. Variance of bootstrap model predictions on training and validation data.
2. RMSE of model predictions on training and validation data for member models.
3. RMSE of model predictions on training and validation data for the ensemble, or bagged model.
4. Observations of the timing and magnitude of model predictions of algal blooms compared to observed events judged on the basis of visual examination of time-series plots.

4.2.4 Computational Platform

As suggested in section 2.5.5, three layers of software were used for experimentation;

1. Database storage
2. Middleware
3. ANN simulator client

As discussed in section 2.5.5, all data is stored in a relational database. MySQL v 3.23 was used as the relational database engine in the present study. The middleware facilitates communication between the database layer and the application layer. It permits automation of the data pre and post-processing stages. The middleware application used for the present study, "LakeNet", was purpose built using the Java programming language. An "off-the-shelf" ANN simulator software package was used in preference to a custom designed program. The Stuttgart Neural Network Simulator (SNNS) (University of Stuttgart, 1999) was chosen as a platform for all experimentation. This software provides many ANN architectures and training algorithms. Also, messaging between the client and middleware is facilitated using the "Batchman" command line client. SNNS is distributed under the conditions of an "open-source" license meaning that the source code can be examined to check its method of operation.

For the purposes of this work, all experimentation was conducted on PCs running the Linux operating system. The hardware included an Intel Pentium III 733 MHz workstation with 256 MBytes of RAM and an AMD Athlon 1800XP (1535 MHz) workstation with 512 MBytes of RAM. Running all the experiments documented in this chapter took several weeks.

4.2.5 Experimental Treatments

Table 4.2 shows the range of treatments applied to each of the models defined for this experiment. The experiment is factorial in design meaning that all combinations of learning algorithms, stopping error and hidden layer configurations were trained.

Early stopping of training was implemented by checking model prediction error on training set data after each training iteration. If it is less than or equal to the designated stopping error, training was halted. Note that SCG is a *batch mode* training algorithm and BP was implemented in this experiment in *incremental mode*. This means that the training set error will be checked after presentation of each record in the BP experiments and after presentation of the entire training set for SCG.

Table 4.2: Summary of experimental design.

Experimental Treatments	
Learning algorithms	SCG (Møller, 1993) Backprop with momentum (Rumelhart et al., 1986)
Stopping error	0 – 4
No. hidden nodes	0, 5, 10, 20

The stopping errors applied were unique to each model design/dataset. This is because preliminary experiments showed that the stopping error required for reasonable model approximation on training sets within a reasonable time limit varied according to both the model design and the database used.

4.2.6 Summary

The methodology with respect to model structure, inference, validation and computation is summarised in table 4.3.

4.3 Results and Discussion

4.3.1 Effect of Model Complexity

4.3.1.1 Model Error Rates

The co-plot feature of the R package for statistical computing The R Development Core Team (2001) was used to illustrate interactions between different treatment effects. Figures C.1 to C.7 illustrate the effects of stopping error on the distribution of model prediction error rates given the number of hidden layer units and model aggregation (bagging) for all 7 models. Part A of these plots shows the training error and part B shows validation set error.

With respect to training error, it can be seen from all these plots that the ANNs with 0 hidden units perform very differently from those with 5, 10 or 20 hidden units. In general, without a hidden layer, the stopping error of training had little influence on the final training set error[1]. However, where hidden layer units were present, there is a close relationship between the stopping error and the training error, with lower stopping errors leading to reduced prediction errors on

[1]Training and stopping errors have different units because the stopping error is calculated from scaled data, whereas the training and validation error rates have been calculated once the predictions have been rescaled.

Table 4.3: Summary of general methodology and computational platform.

Model	
No. inputs	model defined
No. outputs per model	1
Time component	TDNN with input windows
Input window stat.	average
Output representation	identity
Model approximation	
No. hidden layers	1
No. hidden layer units	≤ 20
Trans. func. – hidden	logistic
Trans. func. – output	identity
Architecture	feedforward MLP
Learning algorithms	incremental BP with momentum (Rumelhart et al., 1986)
	Scaled conjugate gradients (Møller, 1993)
Data preprocessing	inputs – standardised mean = 0, stdev. = 1
	outputs – scaled, min. = 0.1, max. = 0.9
No. training epochs	$\leq$10000
Validation	
Method	Bootstrapped blocked 20–fold–crossvalidation
No. replicates	30 per block
Computational Platform	
Database	MySQL v 3.23
Middleware	LakeNet
ANN Simulator	Stuttgart Neural Network Simulator (SNNS) v 4.1
Operating System	Linux kernel 2.4.x
Hardware	1 Intel Petium III 733 MHz, 256 MBytes RAM
	1 AMD Athlon 1800XP, 512 MBytes RAM

the training set. With the exception of model 2 (see figure C.2), a training error close to 0 was achieved where sufficient hidden layer units were present and the stopping error was 0. 5 hidden units was sufficient to achieve close to 0 training error for models 4, 5, 6, (figures C.4, C.5 and C.6) while 10 units were required for models 1 and 3 (figures C.1 and C.3) and 20 hidden units was required for model 7 (figure C.7). The bagged model predictions were generally characterised by lower training error than the unaggregated models[2].

These results indicate that the number of hidden layer units serves to define the maximum model complexity possible for a given input-output structure. Thus, where the number of hidden layer units is low relative to the data availability, the minimum training error will be significantly higher than 0. However, where sufficient hidden units exist, the minimum training error is close to 0. Thus, insofar as the present models and methods are concerned, obtaining a suitably bias free model *approximation* (ie 0 training error) is simply a matter of ensuring sufficient hidden layer units are present. This indicates that the training algorithms used (BP and SCG) have been configured correctly as they are able to fully exploit ANN structures. Also, it is clear that bad local optima do not appear to be interfering with model approximation.

Part B of figures C.1 to C.7 show that when models are not bagged, there is a strong interaction between effects of the number of hidden units and stopping error on validation performance. ANNs without a hidden layer appear almost completely insensitive to the effects of stopping error. However, ANNs with a hidden layer have sharply rising validation error rates at low stopping errors. In the case of models 2, 6 and 7 (figures C.2, C.6 and C.7), the effect of stopping error is dependent on the size of the hidden layer, with more hidden units leading to greater validation set error rates at low training error. Models 1, 2 and 7 (figures C.1, C.2 and C.7) show the classic "underfitting – optimum – overfitting" response to training error as it was decreased where hidden units were present. The remaining ANNs with hidden units simply suffered ever increasing validation error with reduced training errors. These results show that the hidden layer configuration has very little impact on model *generalisation* as opposed to *approximation*. Instead, where hidden units are present, stopping error of training is a much more important determinant of validation error rate. Specifically, it was observed that;

- Overfitting occurs at low stopping errors regardless of the size of the hidden layer (where hidden units were present).
- Optimum validation set error rates with respect to training error were very similar for all hidden layer configurations.

[2]There are two points for the aggregated plots since they represent the error rates of the models trained by BP and SCG.

- The training error at which the validation error optimum occurs is the same for each hidden layer configuration above 0 units for each model respectively.

These observations tend to support the notion that the size of the connection weights between artificial neurons is more important for determining ANN generalisation rather than the number of weights where a hidden layer is present. In other words, the magnitude of the dimensions in the model parameter space play a greater role than the number of dimensions. It can be concluded, as stated by Finnoff et al. (1993), that overfitting cannot be prevented by adjustments to hidden layer size alone – measures to control the parameter sizes must also be employed.

4.3.1.2 The 0 Hidden Unit Models

Clearly, the behaviour of the 0 hidden unit model (ie the perceptron) compared to the multi-hidden unit model (ie multi-layer perceptron or MLP) indicates that it is a different class of inference method. Cheng and Titterington (1994) stated that perceptrons are functionally equivalent to multiple linear regression estimation meaning that linear decision boundaries are assumed with respect to the model inputs. From the results, it can be concluded that such a simplifying assumption means that, while model approximation by perceptrons is not as accurate as by MLPs, they do not suffer from overfitting. However, as Paruelo and Tomasel (1997) demonstrated, linear regression starts to suffer overfitting with increasing input dimensionality. Thus it is likely that the input layers of the models investigated in the present study are too small to cause overfitting by perceptrons.

A surprising result of these experiments is that, in general, the MLPs do not generalise better than perceptrons despite lower training errors. Thus, it follows that any non-linear relationships being learned from the data in the context of the present study are no more valid than the linear relationships inferred by the perceptron. This finding contrasts with the majority of applications of ANNs to ecosystem modelling showing the superiority of the MLP approach due to their ability to generalise non-linear relationships. However, Hwarng and Ang (2001) demonstrated using simulated data that perceptrons may be superior to MLPs for both linear and non-linear time-series modelling. This issue clearly warrants further investigation using a greater variety of models and data than utilised in this chapter.

4.3.1.3 Model Variance

Figures 4.1 to 4.3 plot a number of model performance statistics with respect to the stopping error for the 20 hidden layer unit models. The upper 3 traces on each plot show the unbagged validation error, the bagged validation error and the unbagged training error. The lower 2 traces plot the average standard deviation of

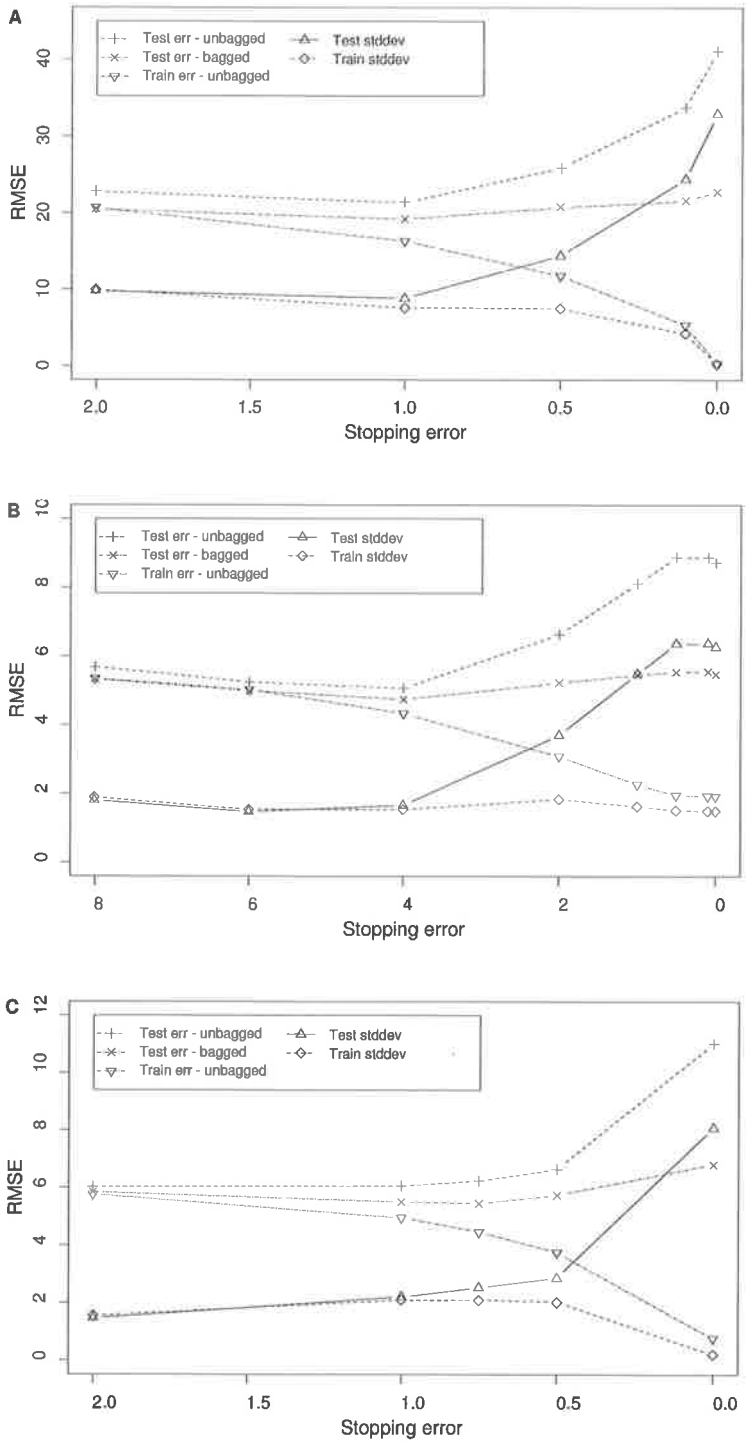

Figure 4.1: **A** RMSE v Stop error – model 1. **B** RMSE v Stop error – model 2. **C** RMSE v Stop error – model 3.

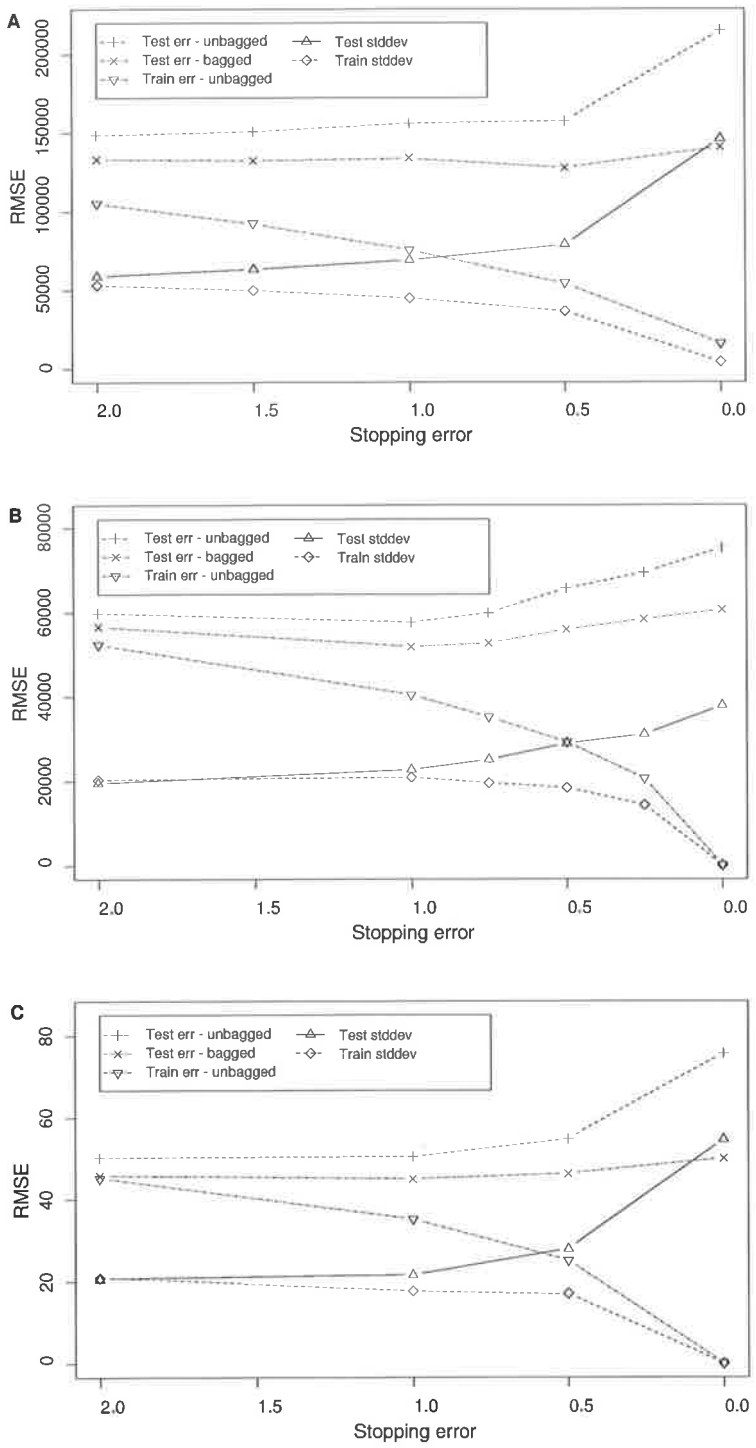

Figure 4.2: **A** RMSE v Stop error – model 4. **B** RMSE v Stop error – model 5. **C** RMSE v Stop error – model 6.

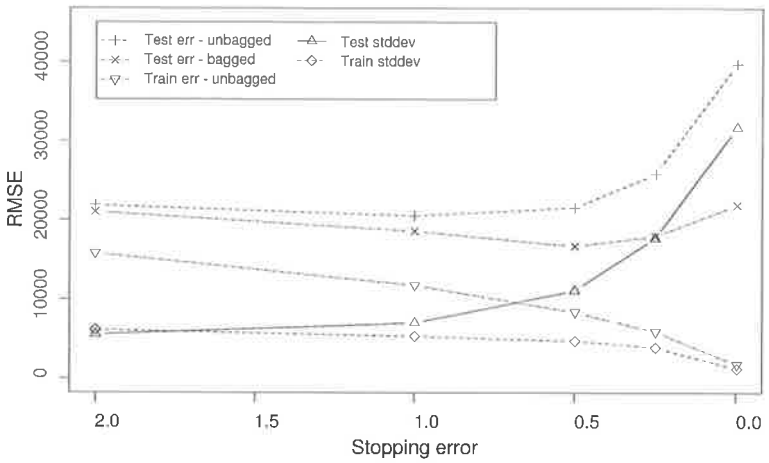

Figure 4.3: RMSE v stopping error – model 7.

bootstrap model predictions over the entire time-series of training and validation set data respectively. These plots show that, for every model, as the complexity of the ANN increases (ie stopping error decreases), the validation set error rates of the unbagged models first decreases, or is unchanged and then rises sharply at some threshold stopping error. The validation error of the bagged model appears correlated with that of the unbagged model, although it is apparent that the effects of overfitting are less pronounced as the effect of stopping error is far lower. Meanwhile, training set error appears well correlated with stopping error.

In every case, the average standard deviation of the bootstrap model predictions on validation data is almost perfectly correlated with the validation error rates. This suggests that the sharp increase in validation error at lower training error is caused by an increase in the variance component of overall (ie bias + variance) prediction error. This observation corresponds with conclusions of Moody (1991) and Geman et al. (1992). The average standard deviation of model predictions on training data declines with decreasing stopping error for all models except model 2 where it remains constant.

The observation of the correlation between validation error and model variance is important in that it signals that prediction variance may be used in some way as a goal function for optimising ANN complexity. In other words, instead of optimising the stopping error (or any other regularisation parameter) by observation of its effect on validation set error rate, it is possible instead to identify the point at which model variance starts to increase rapidly to signify the commencement of the overfitting phase. An advantage of using variance instead of RMSE is that the target values for predictions do not have to be known. Thus the perceived independence of the validation data is maintained eliminating the need for data inefficient "double crossvalidation" where a second validation set is used to estimate real world performance of the complexity tuned ANN.

4.3.1.4 The Overfitting Index

An "overfitting index" can be calculated by comparison of the average variance of model predictions with the variance of the observed output variable;

$$OF = \frac{\text{mean stdev of predictions}}{\text{stdev of observed data}} \tag{4.1}$$

where the numerator refers to the mean of the standard deviation of the bootstrap model predictions for each example in the validation set. The idea of this index is to indicate whether the variance of the models trained is small or large relative to the variance in the output variable. If OF is large, model variance is relatively high and thus is likely to be in an overfitting phase. It is hypothesised that, since this effectively produces a statistic that can be described as the "standardised model variance", it may be possible to identify a universal value that indicates whether or not the model is overfitting.

Figures 4.4 to 4.6 plot OF and bagged validation set RMSE with respect to the stopping error of training for 20 hidden unit models trained by the SCG training algorithm. The patterns of both validation RMSE and OF are correlated with RMSE and test set standard deviation results in figures 4.1 to 4.3. These results are characteristic of the classic underfit – optimum – overfit pattern of total prediction error with increasing ANN complexity. With respect to OF it is apparent that the threshold for the transition to the overfitting stage generally occurs in the region of 0.35 to 0.5. Thus it can be concluded that OF values > 0.5 indicate that the model is overfitting.

It must be conceded that these results are of a preliminary nature and that experimentation with more models and data is warranted to further explore the properties of OF. If OF is proven to be a reliable guide to the occurrence of overfitting, it has the same aforementioned advantages as model variance for use as a goal function in the optimisation of ANN complexity parameters. However, unlike variance, it would not be necessary to plot a curve of variance versus training error to identify the threshold at which overfitting commences, since the number itself conveys the relevant information.

4.3.1.5 Reservations

It may be noted that, with only three hidden layer configurations in this experiment (5, 10 and 20 units), the search of the space of possible configurations was hardly exhaustive. However, given the fact that the results were repeated for 7 models with different input-output layer structures and training set sizes, it seems likely that the conclusions are valid. Furthermore, as Prechelt (1998) points out, practitioners have generally found it more convenient to search the space of stopping

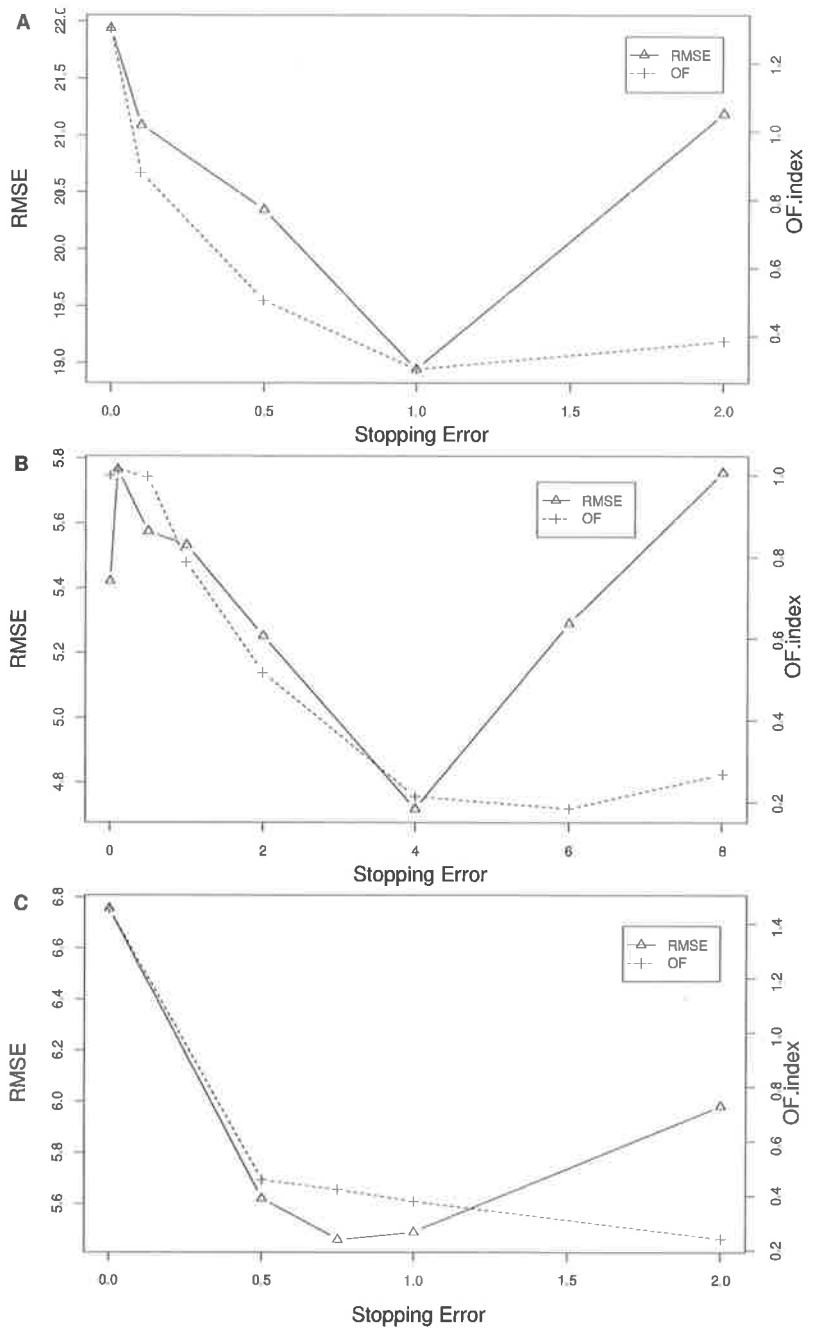

Figure 4.4: **A** RMSE and OF v Stop Error – Model 1. **B** RMSE and OF v Stop Error – Model 2. **C** RMSE and OF v Stop Error – Model 3.

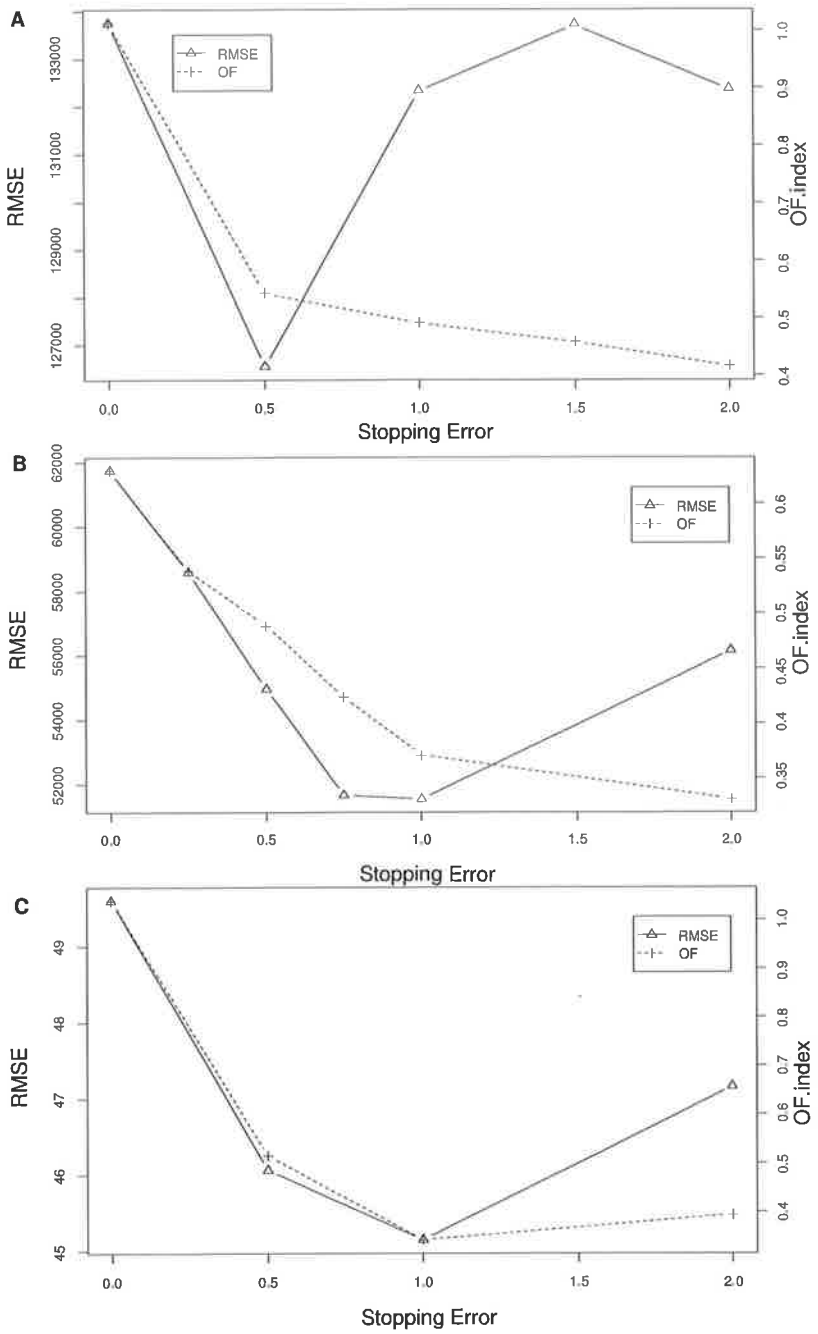

Figure 4.5: **A** RMSE and OF v Stop Error – Model 4. **B** RMSE and OF v Stop Error – Model 5. **C** RMSE and OF v Stop Error – Model 6.

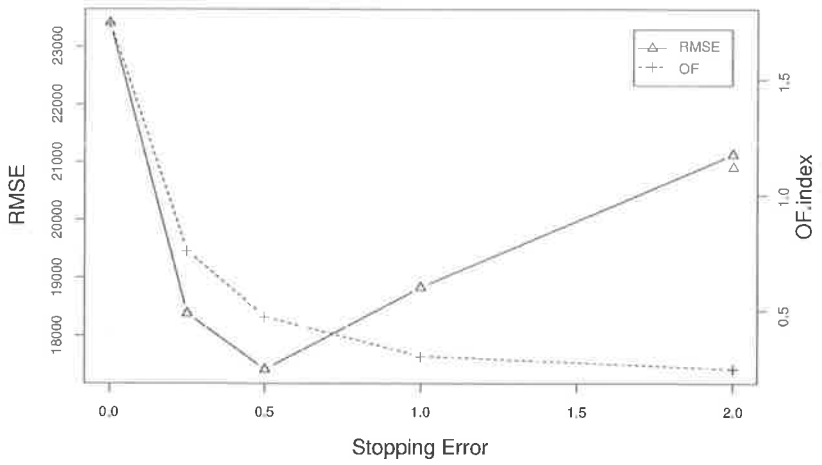

Figure 4.6: RMSE and OF v Stop Error – Model 7.

errors than hidden layer configurations for the simple reason that early stopping does not necessitate time consuming convergence of training.

4.3.2 Model Aggregation

4.3.2.1 Effect of Bagging on Model Performance

Part B of figures C.1 to C.7 show that error rates of the bagged models on validation data are less sensitive to the effects of stopping error or hidden layer configuration than the unaggregated models. This is particularly so for ANNs with hidden layer units trained to low training errors. Figure 4.7 shows the effect of decreasing training error on the predictions of the unaggregated and aggregated models by plotting observed values of chlorophyll *a* abundance versus time and superimposing boxplots of the distributions of validation set predictions made by model 1. Part A of this plot shows the predictions by the model with a stopping error of 1 for a 20 hidden unit ANN and part B shows predictions made by the same ANN trained with a stopping error of 0. It can be observed from the wider boxplots of part B that training the model "harder" leads to much greater variance of model predictions. The bagged model prediction is indicated by the centre line of the box plots. It can be seen that the bagged model has a similar trajectory in both parts A and B showing that model aggregation neutralises, or cancels out, much of the random variance in bootstrap model predictions caused by overfitting.

However, despite the reduced sensitivity of the bagged model to overfitting, it is still possible to discern a shallow optimum validation error with respect to training error where hidden units are present (see figures 4.1 to 4.3 for a clearer representation). For example, in the case of model 1 (figure C.1), where 5 or more hidden units have been used, an optimum exists where the stopping error is 1. As model

complexity is increased beyond the optimal level (ie training error is reduced), overfitting starts to impact performance of the aggregated model. This means that bagging has failed to completely eliminate the need to tune model complexity parameters in order to obtain the best performance on validation data. Thus, while bagging is clearly helpful to prediction accuracy, optimum validation error for a given model can only be achieved by stopping training prior to convergence.

In order determine the significance of the observed error rate optima in terms of model predictions, figures D.1 to D.7 plot the observed algal abundance versus time with the bagged model predictions superimposed. Part A and B of these diagrams plot the training and validation mode predictions respectively for the optimum ANN configurations identified with respect to hidden layer configuration, stopping error and training algorithm. Parts C and D plot training and validation mode predictions for ANN models trained to the "theoretical maximum model complexity" – that is, 20 hidden layer units and 0 stopping error (note that the use of a rotation validation method means that while the same data appears for both training and validation, the validation data is completely independent from that used for training). These plots show that allowing the maximum dimensionality of the ANN to be used in terms of both the number of hidden layer units and the stopping criteria leads to perfect, or near perfect model performance on training sets. By contrast, the models with optimum complexity are clearly not able to perfectly predict the training sets, with over and under predictions of actual biomass being evidenced and some missed predictions on specific events. This observation is consistent with the notion that penalising ANN complexity biases the model. The predictions of the models on validation set data show that allowing full exploitation of the ANN structure does not improve the model's predictions on independent data despite improving performance on training data. Indeed, predictive performance at 0 stopping error has been degraded in all cases with an increase in the number of false positive predictions. Thus it can be concluded that the subjective appraisal of bagged model predictions further emphasises the need to tune the stopping error of training for each model application, despite implementation of bagging, to prevent overfitting from reducing model accuracy on independent data.

This conclusion conflicts with results published by Cannon and Whitfield (2002) showing that, as the number of hidden units and training iterations are increased, the prediction ability of bagged ANNs "plateaus", meaning that bagging eliminates prediction risk posed by overfitting. There are important differences between the demonstration of Cannon and Whitfield (2002) and the present study;

1. Cannon and Whitfield (2002) developed models using much larger training databases drawn from daily observations over a period from 1965-1998. This gives a potential training set size of over 11000 records compared to between 72 and 606 records for the present study. It is generally known that the severity of overfitting is reduced when the size of the training set is large

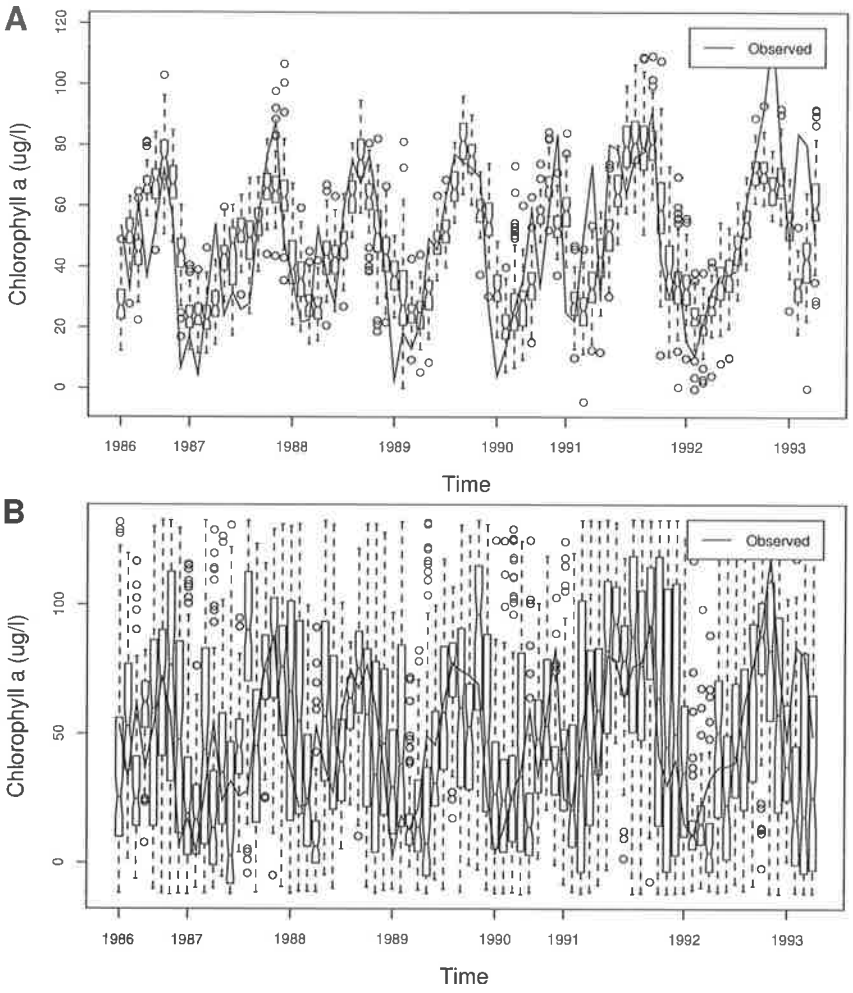

Figure 4.7: Model No. 1. Time series plots of observed and predicted algal abundance – 20 hidden unit ANN. **A** Stopping error = 1.0. **B** Stopping error = 0.

relative to the size of the parameter space of ANN models (Abu-Mostafa, 1989; Kung, 1993; Nejad and Gedeon, 1995).

2. Cannon and Whitfield (2002) gives no information regarding the training algorithm used in their demonstration. Lawrence and Giles (2000) concluded that inefficient training algorithms may reduce the propensity of ANNs to overfitting.

3. The ensemble (bagged) model of Cannon and Whitfield (2002) consisted of 50 member ANNs compared to 30 ANNs for the present study. While Breiman (1996a) suggests that errors due to an insufficient number of member models tends to become insignificant where there are > 25 replicates, it is conceivable that more replicates may improve performance of the bagged models when the member models are overfitting in the context of the present study.

In regards to the first two of these differences, the following points can be made;

- In the present study, performance of bagged models with the most training data – models 2 and 7 – appeared to be just as affected by overfitting as the remaining models. More data may be needed to determine whether training set size does influence the sensitivity of the bagged model to overfitting.

- The results clearly show that "penalising" ANNs by early stopping reduces overfitting. It follows that the use of inefficient training algorithms will have the same effect.

The effect of the number of member ANNs on the sensitivity of the bagged model to overfitting can be checked by a simple experiment. 70 more replicates of model 1 were trained increasing the total number of replicates to 100. SCG training was used and results were recorded for the stopping errors 0, 0.1, 0.5, 1.0, 2.0 and 20 hidden units. Figure 4.8 shows that for most training errors, the 100 member ensemble performs marginally better than the 30 member model. However, there is still a distinct optimum for the 100 member model. This result suggests that increasing the size of the ensemble will not overcome the effect of overfitting on bagged models in the context of the present study. Thus it seems likely that the failure to reproduce the results of Cannon and Whitfield (2002) is related to differences in the size of the training set and/or the type of training algorithm employed.

4.3.2.2 Effect of Bootstrapping on Model Performance

As shown in section 4.3.1.3, overfitting is characterised by a sharp increase in variance in members of an ANN ensemble. Possible sources of variance include;

- Random measurement errors of input and output variables.

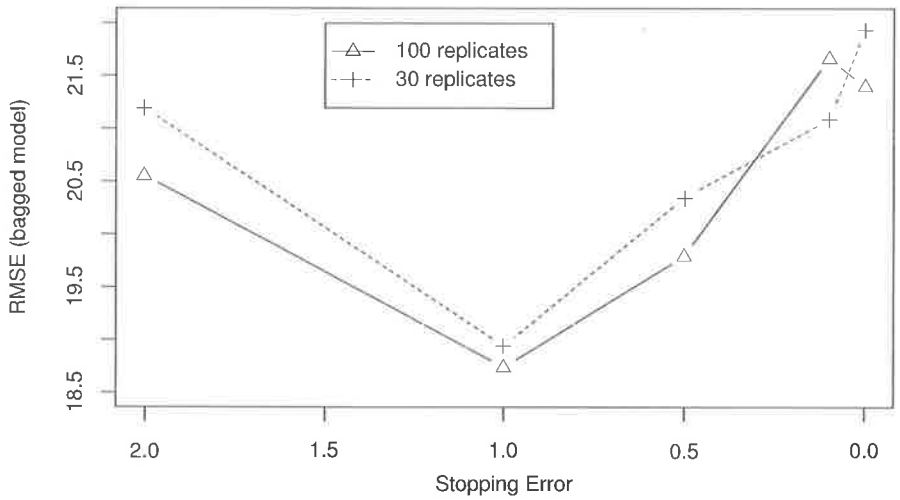

Figure 4.8: Model No. 1. Comparing validation error of bagged model with 30 and 100 member models.

- Fluctuations in ANN training caused by random initialisation of weights, rounding errors during training, entrapment at local optimum etc.
- Sampling errors when selecting training sets.

The effect of sampling errors is easily investigated by comparison of models with and without bootstrap sampling of training sets, since bootstrapping deliberately introduces a source of variance. This was done by retraining model 1 utilising a 20 hidden unit ANN and the SCG training algorithm. 20-fold blocked crossvalidation was used as previously, except that training sets were the complete training sample and not bootstrap subsamples meaning that each model replicate has identical training sets. 30 replicates of models trained to similar stopping errors used previously (ie 0, 0.1, 0.5, 1.0, 2.0).

Part A of figure 4.9 compares average standard deviation of predictions for models trained with and without bootstrap sampling of training data, while parts B and C compare unbagged and bagged validation set error respectively for the two sampling methods. It can be seen from part A that bootstrapping leads to considerably higher variance at higher training errors than the non-bootstrap model. However both models are characterised by sharply increasing variance at lower training errors, with similar variance being exhibited. It appears that when ANNs overfit training data, the sample variance introduced by bootstrapping contributes little to the observed increase in overall model variance. Thus, bootstrapping does not appear to be introducing a source of variance that worsens the effect of overfitting. Part B shows that, when left unaggregated, the non-bootstrapped models generally perform somewhat better than the bootstrapped models at higher stopping errors. This observation is reasonable in light of the fact that the non-bootstrap models display lower variance at high training errors. As training error is reduced, both

models show a similar sharp increase in validation error. Part C shows that when the models are aggregated, the bootstrap and non-bootstrap models have very similar optimum validation error rates at a stopping error of 1.0. However, the bootstrap model is considerably better behaved at lower training errors. Andersen et al. (2001) points out that bagging works best when the errors between member ANNs are uncorrelated and the correct responses are correlated. Thus, a likely explanation for the superiority of the bootstrap bagged model at low training error is that bootstrapping "decorrelates" some errors caused by overfitting.

4.3.2.3 Reservations

A drawback of bagging is that it imposes a greater processing cost on model development because of the manifold increase in the number of ANNs that are trained (a 30-fold increase in the present study). The increased costs include that time taken to train extra ANNs and the increased data preparation and analysis requirements. However it is likely that these problems will become less important in time as the power of computer hardware continues to advance. In this study, the longest recorded training time for a single bagging run using 30 bootstraps and 20-fold-crossvalidation was approximately 8 hours on an desktop PC equipped with a 733 MHz Intel Pentium III CPU and 256 megabytes of RAM. This cost is clearly justifiable in the context of the benefits achieved.

4.3.3 Effect of the Training Algorithm and Model Complexity

Figures E.1 to E.7 illustrate the effects of stopping error on model prediction error rates given the training algorithm and the number of hidden layer units for all 7 models, with part A showing training error and part B showing validation set error. Note that the single low outlier in many of these plots corresponds to the error rate for the bagged model, whilst the remainder of the points correspond to the unaggregated models. Part A of these plots shows that in most cases the SCG algorithm leads to slightly lower training error rates regardless of the number of hidden layer units used or the stopping error of training. Part B of these plots indicates, in most cases, that the distributions of validation error rates appear similar for the two algorithms. However, in the cases of models 2, 4, 6 and 7 (figures E.2, E.4, E.6 and E.7) it is apparent that the SCG algorithm achieves a slightly lower validation error rate for ANNs without hidden layer units. Similar interactions between the effects of stopping error and the number of hidden layer units are present to those observed in figures C.1 to C.7.

Figures 4.10 to 4.12 compare the effect of stopping error on validation performance of the 20 hidden unit bagged model over a range of stopping errors. It can be seen from these plots that the relative performance of the two training algorithms depends on the stopping error and the model. Comparing optimum

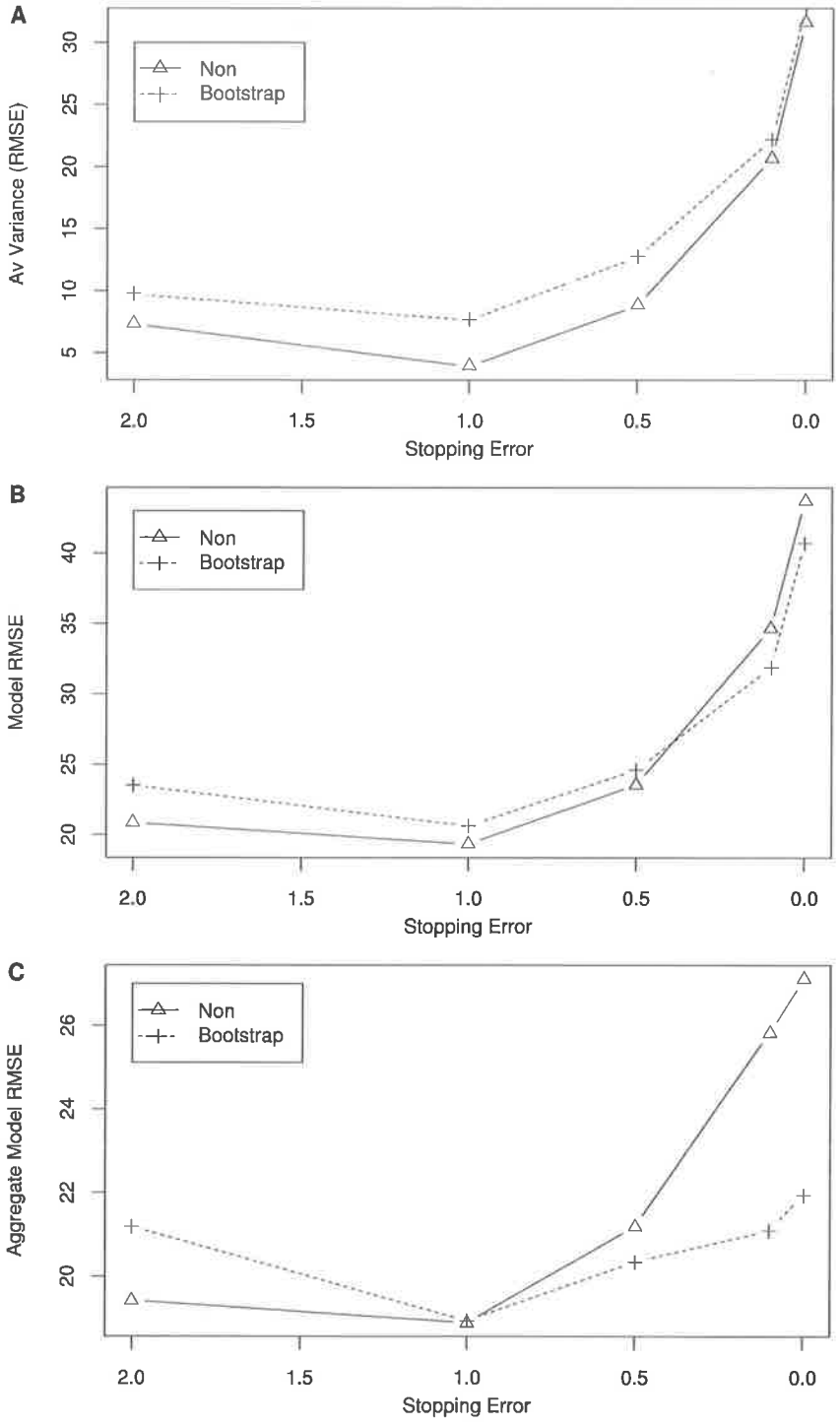

Figure 4.9: Model No. 1. Comparing the use of bootstrap and regular sampling of training set data – 20 hidden unit ANN. **A** Average standard deviation. **B** Validation error of unaggregated models. **C** Validation error of aggregate (averaged) model.

results, BP performs best for models 6 and 7, SCG performs best for models 1, 4 and 5 and similar performance was achieved for models 2 and 3. These optima where achieved at similar stopping errors for models 1, 3, 4, 5 and 7, while different optima were observed for models 2 and 6.

Evidently, it is difficult to judge the superiority of a training algorithm on the basis of validation performance. Clearly there is some interaction between the data and model designs. Also, in most cases, the differences in RMSE are small enough to be judged insignificant. This finding does not support the argument by Lawrence and Giles (2000), nor the empirical findings of Alpsan et al. (1995), that "simpler" training algorithms such as BP produce superior results on independent data than quasi-second order approaches such as SCG. Contrary to the suggestion of Lawrence and Giles (2000), BP appeared to be just as susceptible to overfitting as SCG as indicated by the sharply rising validation error with decreased training error (where hidden units are present). One possible reason for disagreements is that the studies of Alpsan et al. (1995); Lawrence and Giles (2000) draw conclusions on the basis of results from a single model compared to multiple models in the present study.

Contrastingly, in terms of training set error, it is clear that SCG has a slight, but consistently reproduced advantage over BP indicating that it is indeed a more efficient training algorithm. This conclusion is further emphasised by consideration of the time taken by each algorithm to reach the desired stopping error of training. Figure 4.13 plots the training time v the number of hidden units for 3 models trained to a stopping error of 0. The time axis represents the total CPU time taken to train a total of 600 ANNs required by the experimental design (ie 20-fold-crossvalidation repeat 30 times). The plots show that SCG is generally faster to train than BP. Furthermore, there appears to be an interaction between the ANN size and the effect of the training algorithm. The training time for BP scales linearly to model size whereas SCG appears to train slower with intermediate size networks of 5 or 10 hidden units than for ANNs with 0 or 20 units.

4.3.4 Validation Method

Table 4.4 compares validation error rates for optimal ANN configurations validated using the leave-one-out bootstrap and blocked bootstrapped 20-fold cross-validation. Models 1,3,4,5 and 6 have similar error rates for both validation methods. However, substantial improvements in error rates are observed for models 2 and 7 when the leave-one-out bootstrap method is used. These 2 models forecast variables in the Myponga Reservoir, which, as indicated by table 4.4, has significantly better data availability and higher sampling frequency than the other sites investigated in this experiment.

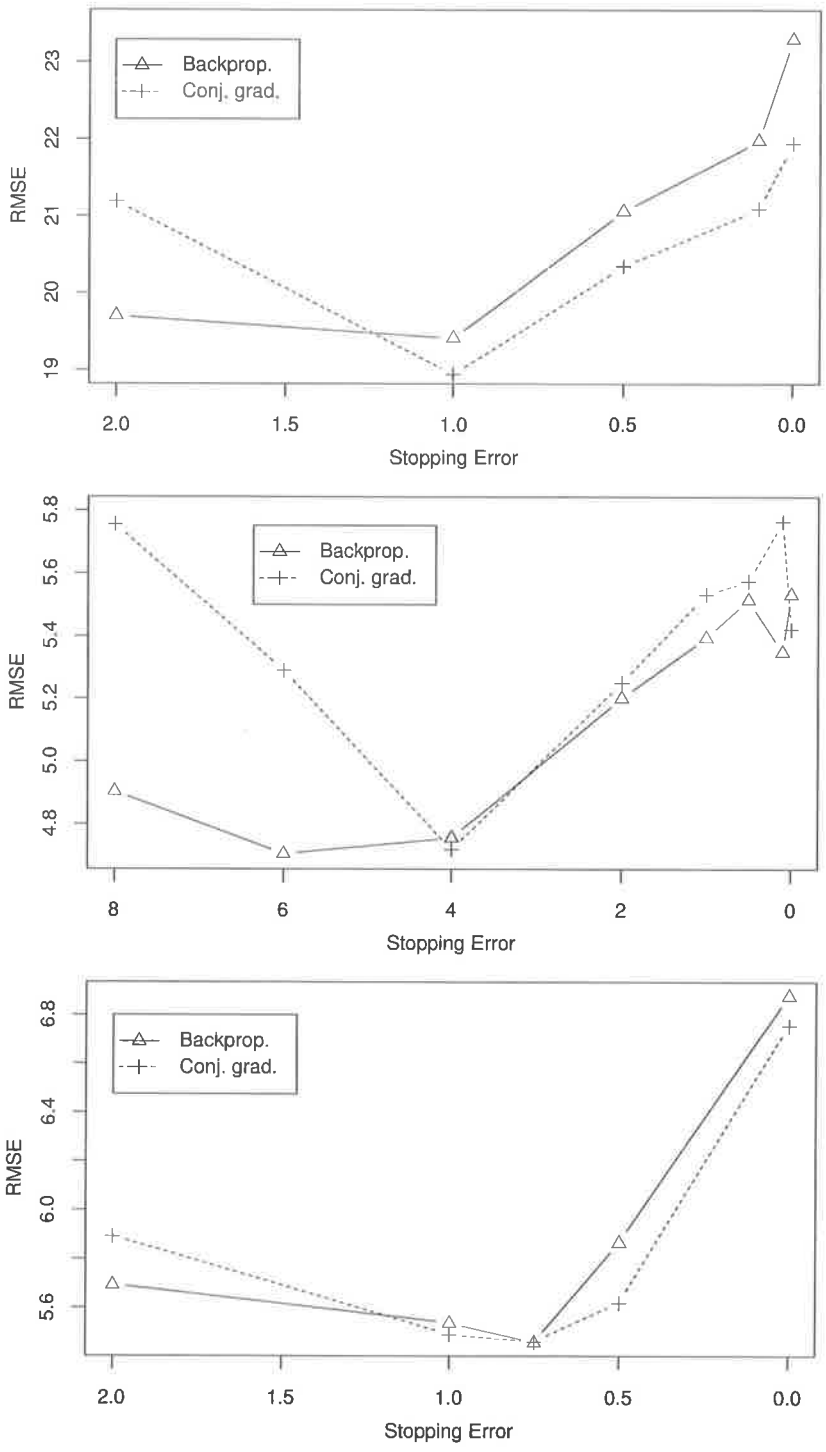

Figure 4.10: Comparison of BP and SCG for 20 hidden unit ANN. **A** RMSE v Stop error – model 1. **B** RMSE v Stop error – model 2. **C** RMSE v Stop error – model 3.

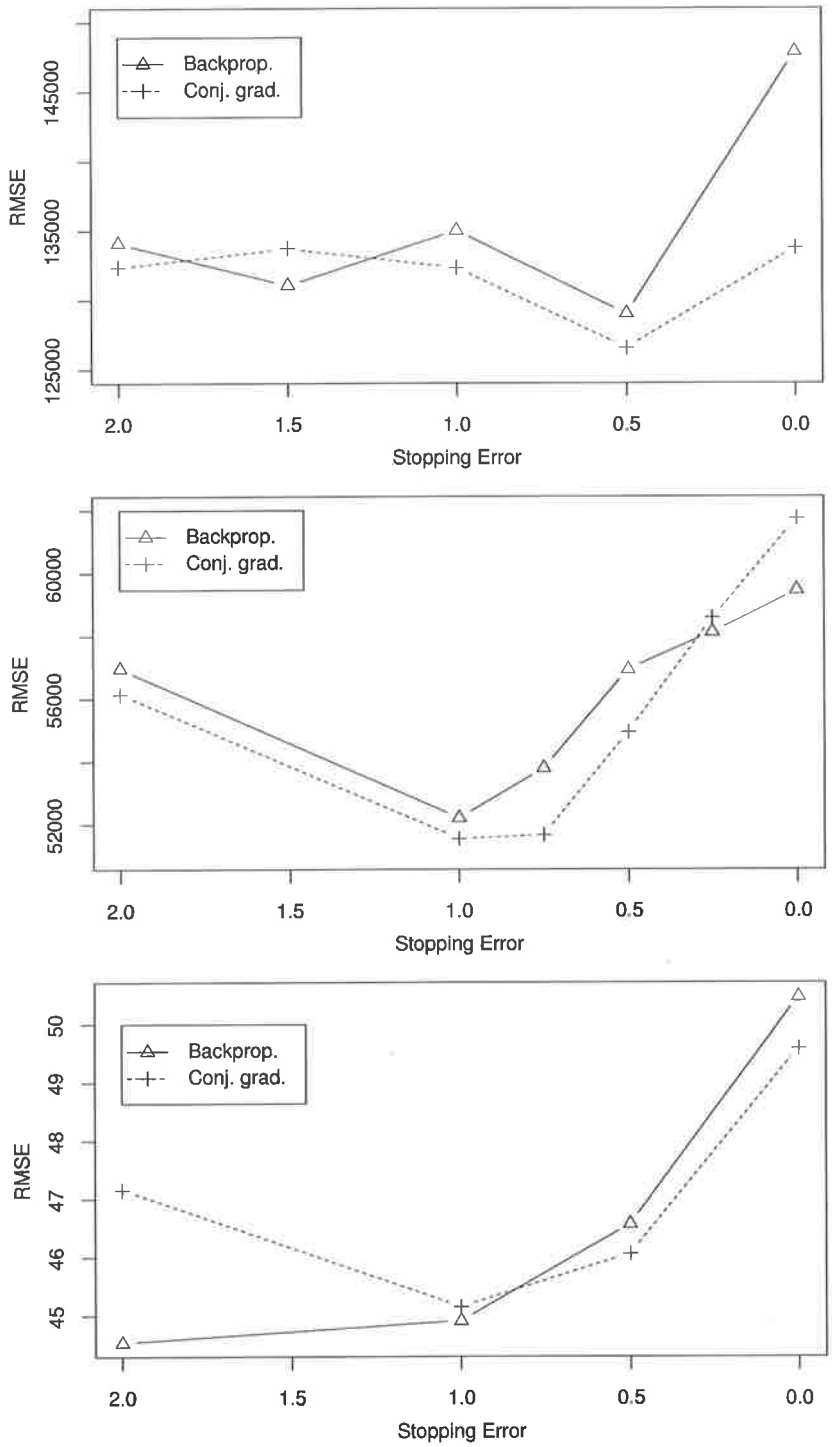

Figure 4.11: Comparison of BP and SCG for 20 hidden unit ANN. **A** RMSE v Stop error – model 4. **B** RMSE v Stop error – model 5. **C** RMSE v Stop error – model 6.

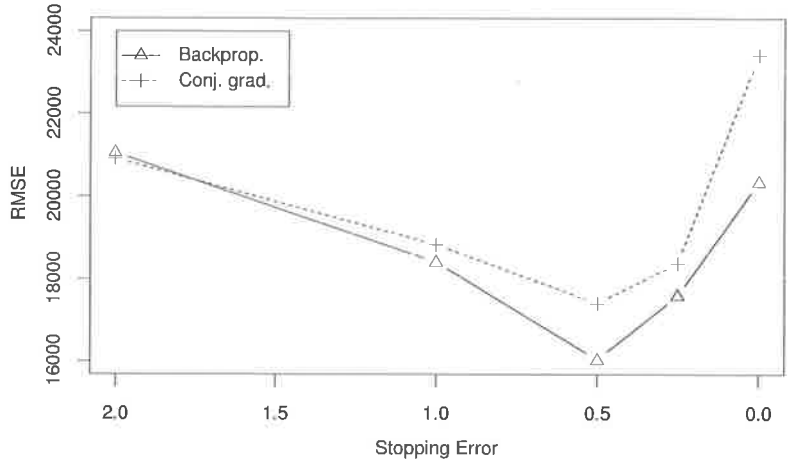

Figure 4.12: Comparison of BP and SCG for 20 hidden unit ANN. RMSE v Stop error – model 7.

Table 4.4: Comparing leave-one-out bootstrap and bootstrapped blocked 20-fold-crossvalidation.

Model	Site	Output	No. records	Validation	RMSE
1	Kasumig. (site 9)	Chlorophyll *a*	82	Bootstrap	19.29
				Cross-val.	18.94
2	Myponga	Chlorophyll *a*	606	Bootstrap	3.941
				Cross-val.	4.718
3	Soyang	Chlorophyll *a*	87	Bootstrap	5.483
				Cross-val.	5.463
4	Kasumig. (site 3)	*Microcystis a.*	125	Bootstrap	124100
				Cross-val.	126600
5	Kasumig. (site 9)	*Oscillatoria spp.*	72	Bootstrap	55060
				Cross-val.	51540
6	Kasumig. (site 3)	Chlorophyll *a*	102	Bootstrap	44.48
				Cross-val.	45.15
7	Myponga	*Scenedesmus spp.*	228	Bootstrap	13510
				Cross-val.	17410

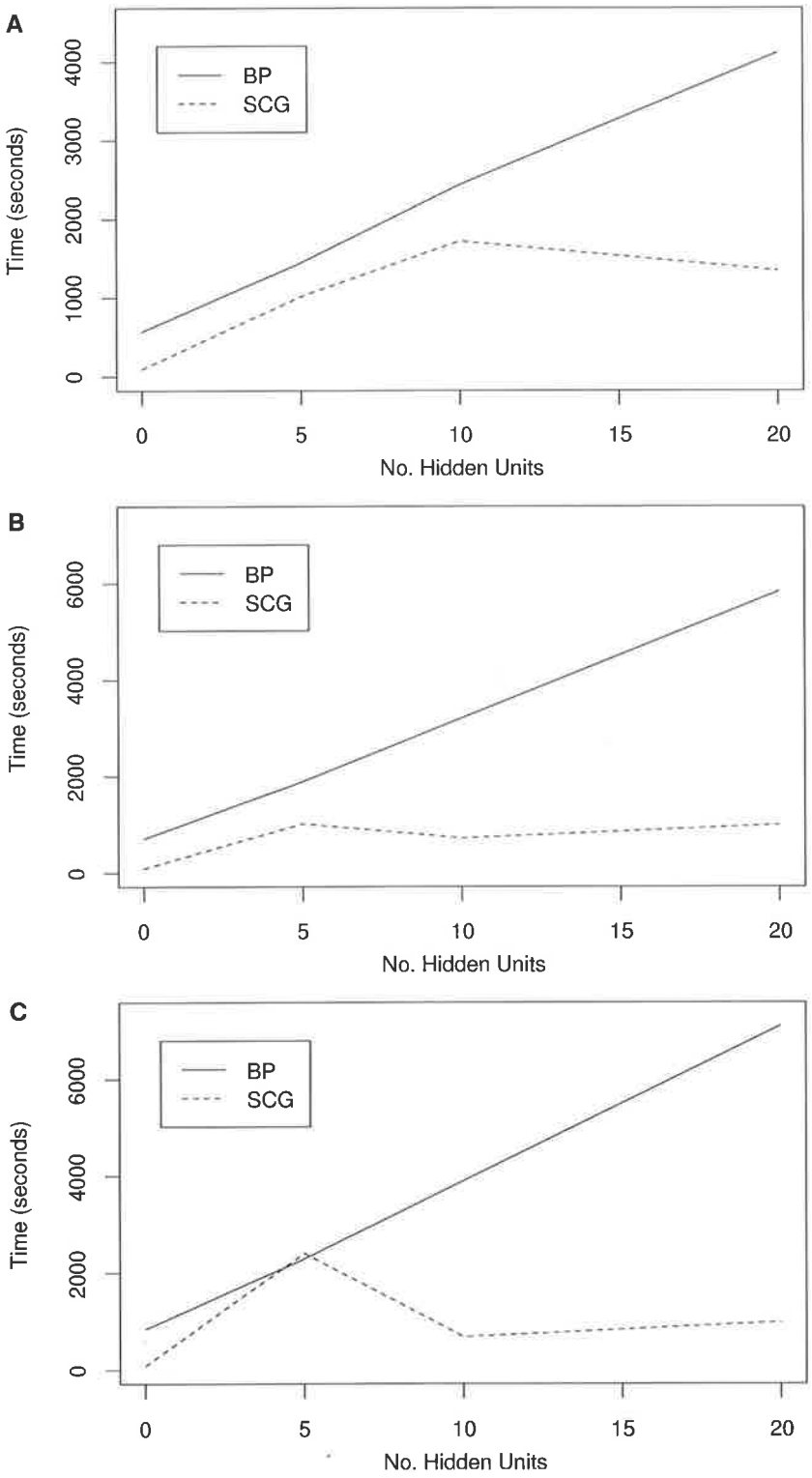

Figure 4.13: Total CPU time v No. Hidden units – a comparison of BP and SCG training algorithms. CPU = Intel Pentium III (733 MHz clock speed). **A** Model 1, **B** Model 2, **C** Model 3.

As explained in section 2.5.3, the 20-fold crossvalidation method used in this experiment employs the following techniques to reduce the effect of time-series correlations between training and validation sets;

- Rotation of 20 hold-out blocks contiguous in time as validation sets.
- A 90 day hold-out period after each validation set not used for training or performance estimation.

By comparison, the leave-one-out bootstrap estimator permits records that are 1 sample interval before or after a each validation record to exist in the training set. Furthermore, since it uses the "out-of-bag" sample for validation, no data is held out from possible training set selection meaning that it will have greater data availability for training. The results in table 4.4 support the conclusion that, at some threshold sampling density, allowing records close in time to be spread across training and validation sets improves the perceived model performance. However, there is no indication that the improved data availability in training sets afforded by the leave-one-out bootstrap method has any affect on model performance, since the models characterised by low data density and availability indicated similar performance using both methods.

The leave-one-out bootstrap estimator has the advantage that it is significantly less resource intensive to perform than the cross-validation method applied in this experiment, since only 30 ANNs have to be trained compared to 600 ANNs (ie 30 bootstrap samples * 20 train/test sets). Thus, it is more convenient to apply in situations where resources are limited or training times are likely to be long.

4.4 Conclusion

The evidence of the experimental results supports the following conclusions and recommendations regarding the application of ANNs to limnological modelling tasks.

4.4.1 Model Approximation

It was shown that both incremental BP and SCG are efficient at approximation tasks since they were both able to fully exploit ANN architectures in the context of the models trained. However, it can be recommended that SCG be used for ANN modelling applications since it is significantly faster than BP while offering similar generalisation characteristics. Additionally, SCG is more convenient for practitioners since there are no parameters such as learning rate or momentum that need optimisation.

4.4.2 Model Generalisation

A surprising result from this experiment is that, for all but 1 model, perceptrons achieved similar validation error rates to MLPs. This means that the ability of MLPs to generalise non-linear relationships rarely confer an advantage over inference methods constrained to linearity in the context of the models developed in this work. This finding contrasts with a large body of work showing superiority of MLPs for ecosystem modelling applications due to their ability to handle non-linear relationships. While further research is warranted to explore this finding, it is recommended that all future experimentation with ANN applications should use perceptrons as a control.

The results clearly show that training error is a far more powerful parameter than the hidden layer size for controlling MLP generalisation characteristics. Indeed, it appears that hidden layer configuration should never be considered in isolation since it has little effect on overfitting behaviour. Furthermore, stopping error is a more convenient parameter because it does not require that ANNs be trained to convergence, which may, in the context of some models and datasets, be computationally expensive. Thus a clear recommendation from this work is that future MLP modelling applications should specify excess hidden layer units and optimise stopping error to reduce generalisation error.

The MLPs trained in this experiment displayed the underfit – optimum – overfit pattern with decreasing training set error. In addition, it was shown that the overfitting phase of ANN performance is characterised by an increase in the variance component of prediction error. These two observations are consistent with widely held expectations from the literature. It was proposed that model variance be used as a goal function for optimising stopping error of training because it is closely correlated with validation error and, unlike error, does not require access to the target values for prediction of the validation set. This eliminates the need for double crossvalidation to determine an unbiased estimation of model performance. An overfitting index based on model variance was proposed as a universal, unit free measure of overfitting. More work is needed to further explore the properties of this quantity.

4.4.3 Model Aggregation

It was shown that bagged models have significantly reduced sensitivity to overfitting and lower validation error than the individual member ANNs of the ensemble. This effect is consistent with expectations from the literature. However, it was shown that bagging does not completely eliminate the need for tuning of training error since overfitting can still impact the performance of the aggregate model. It was shown that bootstrap sampling of training data does not contribute significantly to increased model variance of overfitted ANNs. Instead, it tends

to limit the effect of overfitting on ensemble models by possibly decorrelating prediction errors. Thus it is recommended that bagging should be applied in all future modelling work, but that for optimal model performance, early stopping of training should also be applied.

4.4.4 Model Validation

It was observed that the use of the leave-one-out bootstrap estimator leads to better perceived validation error rates than the blocked 20-fold cross-validation estimator when data availability and/or sampling density of the time-series are high. This is probably due to the former method not eliminating time-series correlations existing between training and validation data. Decisions regarding the choice of performance estimation method depend on the specific requirements of the application. However, it should be noted that the leave-one-out bootstrap estimator is less computationally expensive than the blocked 20-fold crossvalidation approach developed for this work.

Chapter 5

The Generic ANN Model

5.1 Introduction

The review in sections 2.3.1 and 2.4.2 showed that input layer designs for supervised ANN models predicting phytoplankton abundance generally considered one or more of the following factors;

- Theories regarding the causes of algal growth.
- Serial correlation of time-series data.
- Outcomes of data analysis or previous modelling work.
- Availability of variables in historical water quality monitoring database.

It can be concluded from the review in chapter 2 that input layer designs are, in practice, being implemented in an *ad-hoc* nature depending on data availability. However, it was shown in chapter 3 that variables expressing nitrogen and phosphorus bioavailability, water temperature and underwater light penetration are common to all six datasets available for the present study. These variables are widely used by modelling applications in the literature such as those reviewed in chapter 2. It was proposed by Wilson and Recknagel (2001) that these four variables comprise a *generic* ANN model; the benefits of which are:

- Compatibility with existing "typical" datasets.
- Potential rationalisation of future data collection efforts.
- Facilitation of direct comparisons between models/datasets.
- Compaction of ANN size leading to a reduction of overheads associated with data pre- and post-processing and model training.
- Aggregation of data from multiple sites to permit the training of increasingly generic "mega-models".

This chapter aims to validate the generic model structure in the context of the following;

- The six datasets reviewed in chapter 3.
- 21 output variables identified in chapter 3.
- The input-window model structure proposed in section 2.5.1.
- Bootstrap aggregation as suggested by Breiman (1994).
- Blocked 20-fold cross-validation as proposed in section 2.5.3.
- Findings with respect to model approximation, generalisation and complexity discovered in chapter 4.

In addition to validation of the generic model design, this chapter aims to explore a number of error measures for realistically quantifying and comparing ANN model performance where different ranges and units of the output variable are used.

5.2 Methods

5.2.1 ANN Models

The generic model design investigated in this chapter is described in section 3.4.1. As indicated, four model outputs were specified for each of the datasets, except Soyang, where one model was specified resulting in a total of 21 models (see table 3.20). The "generic ANN model template" is illustrated in figure 3.18 showing the four inputs and a "feedback" input comprising the output variable in each case. Table 3.19 shows the actual input layers used for each dataset. Time series dependencies were modelled using the TDNN structure, with input variables lagged the output by two weeks. The "input window" approach to representing model inputs, described in section 2.5.1, was utilised meaning that inputs to the ANN were average values of each of the driving variables falling between two and four weeks prior to the output date.

5.2.2 Model Inference

A three layer feed-forward multi-layer perceptron consisting of an input, hidden and output layer was used as the underlying structure for all models. As in chapter 4, ANNs with zero and 20 hidden layer units were compared to determine the importance of non-linear decision boundaries to all the modelled outputs. The methodology with respect to the architecture of neural processing, data conditioning, the maximum number of training epochs, validation and computation is identical to that used in the experiments in chapter 4 (see table 4.3).

Since the experimental evidence presented in chapter 4 showed the Scaled Conjugate Gradient method (Møller, 1993) to be an efficient learning algorithm, it was used for training all models. Model predictions were stabilised by bagging, with the bagged model ensemble consisting of 30 member models in each case. The models were further stabilised by stopping of training early. It was found, as discussed in chapter 4, that there was a correlation between the average standard deviation of model predictions and the bagged model error rate. Thus the stopping error chosen was the error at which the average standard deviation of predictions started to increase dramatically.

5.2.3 Model Validation

The blocked bootstrapped 20-fold-crossvalidation performance estimator, described in section 2.5.3, was utilised for all model outputs. Note that in every case, model performance refers to the bagged model ensemble, not to individual member models. In order to compare the model performance given variation in the range of the output values, a number of unit free error measures were investigated. These are described below.

5.2.3.1 Continuous Error Measures

Tables 5.1 to 5.6 compare performance of the optimised generic ANN models to similarly configured perceptron models (ie ANNs without a hidden layer). The error measures include;

- **RMSE**. This measure is the goal function of ANN training. Note that the square term tends to emphasise the importance of large discrepancies between modelled and observed values over small ones.
- $U1$ – Theil's inequality type 1 Theil (1961) is a standardised RMSE measure. A U1 error rate of 0 indicates perfect agreement, while 1 indicates perfect disagreement.

$$U1 = \frac{\sqrt{\frac{1}{n}\Sigma(y_o - y_p)^2}}{\sqrt{\frac{1}{n}\Sigma y_o^2} + \sqrt{\frac{1}{n}\Sigma y_p^2}} \tag{5.1}$$

  where y_o is the observed value and y_p is the model predicted value.

- $U2$ – Theil's inequality type 2 is a comparative error rate comparing the root means square error (RMSE) of the ANN with that of a naive model where the naive model is expressed as $y_t = y_{t-1}$. $U2 < 1$ indicate that the ANN model is performing better than the naive model. According to Armstrong

and Collopy (1992), U2 is a useful measure for time-series where large changes occur over the forecast horizon. :

$$U2 = \frac{\text{RMSE ANN}}{\text{RMSE NAIVE}} \tag{5.2}$$

- **Correlation coefficient** (R^2) is the proportion of covariance between the observed and predicted values.

$$R^2 = \left(\frac{\Sigma((y_p - \bar{y}_p)(y_o - \bar{y}_o))}{n\sigma_p\sigma_o}\right)^2 \tag{5.3}$$

where σ is the standard deviation and $\bar{y}$ is the mean.

U1, U2 and R^2 are unit free measures thus facilitating meaningful comparison between models with different ranges in the output variable.

5.2.3.2 Classification Error Measures

Tables G.1 to G.21 compare the performance of the ANN models as classifiers. Classification errors were calculated as "confusion matrices" given the models' ability to classify "bloom" and "non-bloom" events. These matrices contain counts of each of 4 different classification outcomes – true positive predictions (TPP), true negative predictions (TNP), false positive predictions (FPP) and false negative predictions (FNP). These attributes were calculated for 5 different thresholds for defining bloom events for each model, since it is known that the value of the threshold can have a large effect on perceived performance. The values of the thresholds are data specific and were chosen such that there were approximately equal numbers of records in each of the classifications. The statistics presented were calculated as follows;

- **Prevalence** – the proportion of positive (ie "bloom") classifications in the observed data.
- **Sensitivity** – the conditional probability that a bloom case is classified by the model.

$$\textbf{Sensitivity} = \frac{TPP}{TPP + FNP} \tag{5.4}$$

- **Specificity** – the conditional probability that a non-bloom case is classified by the model.

$$\textbf{Specificity} = \frac{TNP}{TNP + FPP} \tag{5.5}$$

- **Positive predictive power** – the conditional probability that a bloom prediction is actually present.

$$\mathbf{PPP} = \frac{TPP}{TPP + FPP} \tag{5.6}$$

- **Kappa** – an estimation of the proportion of agreement between the model and the observed data after removal of the proportion of agreement due to chance. For example, $\kappa = 0.5$ suggests 50% better classification accuracy than a naive model making random classifications. κ is a robust measure of classification accuracy that allows valid comparisons between different datasets since it provides reasonable weighting against conditions of very low or high prevalence. Agreement is generally considered weak where $\kappa < 0.4$, moderate where $0.4 < \kappa < 0.75$ and strong where $\kappa > 0.75$.

$$\kappa = \frac{(TPP+TNP)-(((TPP+FNP)(TPP+FPP)+(FPP+TNP)(FNP+TNP))/N}{N-(((TPP+FNP)(TPP+FPP)+(FPP+TNP)(FNP+TNP))/N} \tag{5.7}$$

5.2.4 Computational Platform

All experiments were run using the same software and hardware used in chapter 4 (section 4.2.4). Thus MySQL was used for data storage, SNNS v 4.1 was used for ANN simulation and all communication between the database and application layers was handled by LakeNet. As previously, standard desktop PCs utilising Intel Pentium III and AMD Athlon XP CPUs were used for data processing and ANN training.

5.3 Validation Set Performance

5.3.1 Model Performance Evaluation

Figures F.1 to F.6 illustrate time-series plots of observed values and bagged model predictions of outputs on validation data. Note that individual observations have been joined by interpolated lines to emphasise the trajectory of the modelled and observed values through time. The following section contrasts the "subjective" performance of the ANN models according to the time-series plots with the continuous error measures documented in tables 5.1 to 5.6 and the classification error rates documented in tables G.1 to G.21.

Table 5.1: Generic model error rates. Lake Biwa.

Output	Hid.	RMSE	U1	U2	R^2
Chlorophyll *a*	0	6.13	0.292	0.812	0.071
	20	6.27	0.293	0.831	0.061
Euglena americana	0	1194	0.627	0.775	0.010
	20	1048	0.467	0.680	0.233
Melosira granulata	0	682	0.503	0.840	0.147
	20	712	0.497	0.877	0.137
Pediastrum biwae	0	310	0.412	0.741	0.405
	20	284	0.406	0.680	0.454

Table 5.2: Generic model error rates. Burrinjuck Dam.

Output	Hid.	RMSE	U1	U2	R^2
Chlorophyll *a*	0	20.6	0.370	0.534	0.512
	20	19.3	0.326	0.501	0.568
Chlorophyta	0	4150	0.570	0.754	0.004
	20	4050	0.547	0.736	0.010
Cyanophyta	0	100000	0.622	0.986	0.035
	20	78800	0.518	0.777	0.195
Diatoms	0	1940	0.426	0.867	0.300
	20	2040	0.455	0.912	0.236

Table 5.3: Generic model error rates. Darling River.

Output	Hid.	RMSE	U1	U2	R^2
Total phytoplankton	0	22500	0.381	1.049	0.293
	20	24100	0.431	1.126	0.186
Chlorophyta	0	5000	0.450	1.067	0.274
	20	5210	0.500	1.112	0.212
Flagellates	0	1870	0.375	0.993	0.312
	20	1840	0.363	0.975	0.340
Cyanophyta	0	5110	0.452	1.055	0.236
	20	5520	0.509	1.140	0.110

Table 5.4: Generic model error rates. Lake Kasumigaura.

Output	Hid.	RMSE	U1	U2	R^2
Chlorophyll *a*	0	43.4	0.248	0.825	0.190
	20	45.3	0.257	0.860	0.157
Oscillatoria spp.	0	50600	0.589	0.733	0.085
	20	56200	0.595	0.814	0.045
Microcystis aeruginosa	0	89100	0.322	0.753	0.570
	20	82400	0.301	0.697	0.626
Gomphosphaeria spp.	0	24100	0.419	0.844	0.426
	20	25700	0.458	0.901	0.348

Table 5.5: Generic model error rates. Myponga Reservoir.

Output	Hid.	RMSE	U1	U2	R^2
Chlorophyll *a*	0	4.55	0.257	0.977	0.469
	20	4.69	0.265	1.007	0.436
Ankistrodesmus spp.	0	3390	0.642	0.949	0.252
	20	3360	0.538	0.942	0.118
Dictyosphaerium spp.	0	1520	0.589	0.959	0.011
	20	1460	0.570	0.922	0.029
Scenedesmus spp.	0	17500	0.445	1.204	0.327
	20	19200	0.507	1.324	0.199

Table 5.6: Generic model error rates. Lake Soyang.

Output	Hid.	RMSE	U1	U2	R^2
Chlorophyll *a*	0	1.77	0.385	0.810	0.305
	20	1.79	0.389	0.823	0.281

5.3.1.1 Lake Biwa

The chlorophyll *a* model (figure F.1 part A) features mixed measured performance with poor R^2 and κ (0.071 and 0.3 respectively), but good U1 and U2 error rates (0.293 and 0.831 respectively). The time-series plot shows that the model fails to capture most of the observed dynamics of this variable. The good U2 result can be explained by bad performance of the no-change model due to the short term dynamics in the observed data. However, the good U1 result is difficult to explain in the context of the poor model performance evident from the the other error measures and the time-series plots. Table 5.1 shows that the ANN model performs somewhat better than the perceptron model according to all error measures indicating that generalisation of non-linear relationships is advantageous.

The *Euglena americana* (figure F.1 part B) model generally performs somewhat better than the chlorophyll *a* model. It has a mixed measured performance with poor R^2 and U1 (0.233 and 0.467 respectively) but good U2 and κ (0.680 and ≤ 0.61). The time-series plot shows an annual pattern with short lived blooms occurring in spring and an absence of biomass for the remainder of the year. The ANN model successfully predicted the timing of most events, but failed to predict the 1987 and 1988 blooms. In general the ANN tended to underpredict observed peak magnitudes of bloom events. Table G.2 shows a curious balance of results with respect to κ where values of ≤ 0.32 are exhibited for all but the highest threshold of 1000 cells/ml where $\kappa = 0.61$. At this threshold, both sensitivity and specificity are good with values of 0.79 and 0.92 respectively. However, at lower thresholds model performance is marred by low specificity indicating a poor resistance to false positive predictions. The time-series plot confirms these observations. Table 5.1 shows that the ANN performs considerably better than the perceptron.

Mixed measured performance was also observed for *Melosira granulata* (figure F.1 part C) with poor R^2 and U1 (0.137 and 0.497 respectively), but moderate U2 and κ (0.877 and ≤ 0.46 respectively). Once again the good U2 result can be attributed to very poor no-change model performance caused by extreme short term dynamics. With respect to κ, table G.3 shows that, unlike the *Euglena americana* model, performance is better at low thresholds (≤ 46 cells/ml) than higher thresholds. The time-series plot appears to support this finding, as it is clear that the ANN model predicts the presence/absence of this species reasonably well. However, the poorer performance at high thresholds (in particularly, low sensitivity indicating false negative predictions) is reflected by the failure to predict the timing of events in 1985, 1988 and 1989. The ANN performs better according to U1, whereas RMSE, U2 and R^2 indicate better performance by the perceptron for this output.

The *Pediastrum Biwae* model achieved moderate measured performance for this dataset with $R^2 = 0.454$, U1 = 0.406, U2 = 0.831 and $\kappa \leq 0.44$. The time-series plot (figure F.1 part D) shows that the ANN is able to predict the timing

of the bloom events in 1984 – 1987 and a minor event in 1989. Furthermore, the model resists false positive predictions in other years. However, the magnitudes of predictions are characterised by under-predictions in 1984 and 1987 and over-predictions in 1985 and 1986. The U2 error rate indicates a reasonable improvement in RMSE over the no-change model. The classification statistics in table G.4 show that the model performs best at intermediate threshold values where reasonably high sensitivity and specificity values are observed. However, at the highest threshold (150 cells/ml), the model has high specificity but poor sensitivity indicating a tendency for false negative predictions. This may be due to a failure to predict the 1987 event and the commencement of the 1984 event. Table 5.1 shows the ANN delivers slight but consistently better performance than the perceptron for this output.

5.3.1.2 Burrinjuck Dam

The chlorophyll *a* model achieved the best measured performance for this dataset with $R^2 = 0.568$, U1 = 0.326, U2 = 0.501 and $\kappa \leq 0.40$. Figure F.2 part A shows that the time-series is dominated by 2 severe blooms in 1980 and 1983 and is relatively featureless for the remaining years. The ANN model predicts the timing and magnitude of the 1983 event perfectly and under-predicts the 1980 event. The small scale dynamics in 1981 and 1982 are also well forecast by the model. A slight false prediction may be observed in 1994. This model had a U2 error rate of 0.501 indicating very much better performance than the no-change model. The classification statistics in table G.5 show that the model performs best at the second highest threshold (10.55 μg/L). Note that the maximum threshold value of 16 is low relative to the maximum observed value of ≈ 225 because of the low prevalence of very severe bloom events. Table 5.2 shows that the ANN delivers consistently better performance than the perceptron for this output.

The chlorophyta model (figure F.2 part B) had relatively poor measured performance with $R^2 = 0.010$, U1 = 0.547 and $\kappa \leq 0.29$. However, U2 was reasonable with a value of 0.736 reflecting poor no-change model performance. The time-series plot shows that the poor error measures are justified as the model is clearly unable to predict the algal bloom events. Table 5.2 shows that the ANN performs slightly, but consistently better than the perceptron for this output.

Like chlorophyll *a*, figure F.2 part C shows that cyanophyta dynamics are dominated by 3 extreme events in excess of $1 * 10^5$ cells/ml, whilst the remainder of the time-series is featureless. However, this model has a much poorer measured performance than the chlorophyll *a* model with $R^2 = 0.195$, U1=0.518 and $\kappa \leq 0.32$. Once again, U2 was reasonable (0.777). The time-series plot shows that the model predicts the 1996 event quite well, but under-predicts the 1980 event and overpredicts the 1982 event. Additionally there is a false positive prediction extending over much of 1983. The classification statistics in table G.7 show that the model performs best at the highest threshold values where reasonable

sensitivity and specificity values of ≥ 0.73 are observed. However, the very low prevalence of extreme bloom events means that the maximum threshold of 10000 cells/ml is very much lower than the maximum observed value of $\approx 7 * 10^5$ cells/ml. Table 5.2 shows that the ANN performs considerably better than the perceptron for this output with $\approx$ 20% performance improvement according to most indices.

The Diatoms model (figure F.2 part D) achieves moderate measured performance with $R^2 = 0.236$, U1 = 0.455, U2 = 0.912 and $\kappa \leq 0.48$. The time-series plot shows that the model predicts the major bloom events in most years, but with errors in timing and magnitude. Also there is a significant false positive prediction evident in 1986-1987. The classification statistics for this model (table G.8) indicate that the model performs best at low to moderate thresholds of 300 and 700 cells/ml, at which the model has a sensitivity of 0.9 indicating a high probability of correctly classifying bloom events, although specificity is moderate at 0.64 indicating a significant risk of false positives. At the highest threshold of 2000 cells/ml the specificity dramatically improves, but sensitivity drops to 0.50 indicating a tendency to false negative predictions These conclusions are supported by the time-series plots. Table 5.2 indicates that the perceptron performs better than the ANN for this output indicating that constraint to linearity is most appropriate for this model.

5.3.1.3 Darling River

The total phytoplankton model (figure F.3 part A) achieved moderate to poor measured performance with $R^2 = 0.186$, U1 = 0.431, U2 = 1.126 and $\kappa \leq 0.42$. The time-series plot shows that the ANN fails to predict algal abundances higher than 40000 cells/ml and only predicts some of the lower level dynamics. The classification statistics in table G.9 show that model performs best at the relatively low threshold of 21000 cells/ml. At a higher threshold, the model has a good specificity, but is compromised by poor sensitivity indicating a resistance to false positive predictions but a tendency to false negatives. These conclusions support observations of the time-series plots. Also, a U2 error rate > 1 indicates that the ANN performs significantly worse than the no-change model. As discussed below in section 5.3.1.7, this is a result of the high sampling density causing relatively good no-change model RMSE error rates. Table 5.3 shows that the perceptron performs somewhat better than the ANN for this output.

The chlorophyta model (figure F.3 part B) had similar measured performance to the total phytoplankton model, with $R^2 = 0.212$, U1 = 0.500, U2 = 1.112 and $\kappa \leq 0.42$. The time-series plot shows that the dynamics of chlorophyta are highly correlated with total phytoplankton, with peak events coinciding in 1980-81, 1982, 1985-86 and 1987-88. Clearly, chlorophyta contribute significantly to the total algal cell count. The observed performance of the ANN was almost identical to that of the total algal cell count model described above. Thus the

model was unable to model larger scale dynamics, but performed reasonably well at modelling lower level events. Once again the U2 error rate indicated that the ANN had higher RMSE than the no-change model. Table 5.3 shows that the perceptron performs significantly better than the ANN for this output.

The cyanophyta model (figure F.3 part C) had similar measured performance to both the total phytoplankton and the chlorophyta models ($R^2 = 0.110$, U1 = 0.509, U2 = 1.140 and $\kappa \leq 0.27$). The time-series plots shows that the dynamics of cyanophyta in the Darling River are closely correlated with both chlorophyta and the total cell count. This indicates that, like chlorophyta, cyanophyta is a significant fraction of the river microflora. However, the time-series plots also show that the generic ANN model performs worse than the chlorophyta or total phytoplankton models being less able to accurately model low or high level dynamics in the time-series. The classification error rate data (table G.11) shows that model predictions are either characterised by high sensitivity but low specificity, or low sensitivity but high specificity. Table 5.3 shows that the perceptron performs significantly better than the ANN for this output.

The flagellates model for the Darling River is the best performing with $R^2 = 0.340$, U1=0.363 and $\kappa \leq 0.56$. Figure F.3 part D shows good general correspondence between observations and model predictions. The model appears to be particularly highly accurate at forecasting the low level dynamics in algal abundance occurring from 1989 onwards. However, it can be seen that the model is not able to forecast the extreme events – particularly those $>$ 5000 cells/ml occurring in 1981, 1982 and 1985 – 1987. This inability to meet extremes may explain why table G.12 shows that κ is higher for intermediate bloom thresholds than for the highest threshold. At intermediate thresholds of 800 and 1900 cells/ml, predictions are characterised by reasonable sensitivity and specificity. However, at the highest threshold of 3500 cells/ml, the sensitivity drops to 0.44 indicating a tendency for false negative predictions. These observations reflect findings from the time-series plots. Table 5.3 shows that the ANN performs slightly better than the perceptron according to all indices.

5.3.1.4 Lake Kasumigaura

The chlorophyll *a* model (figure F.4 part A) has mixed measured performance with poor R^2 of 0.157, but reasonably good remaining measurements (U1=0.257, U2 = 0.860, $\kappa \leq 0.41$). The time-series plot shows a regular periodicity to the data with peak events in the summer and autumn months of most years. The model is clearly able to capture the periodicity, but does not appear to predict the year to year variations in concentration particularly well. For example, the model does not predict the higher peak values in 1983, 1984, 1985 and 1986 compared to the rest of the time-series. The classification statistics (table G.13) show that the model only achieves reasonable κ values at the lowest threshold value (30 μg/L) indicating that at higher thresholds the model does not perform a great deal better

than chance. However, even at the lowest threshold, the model has extremely good sensitivity (0.98) but poor specificity (0.33). The intermediate to high thresholds are classified by low sensitivity which is reflective of an inability to meet the peak events. Table 5.4 shows that the perceptron model performs slightly better than the ANN according to all indices.

The *Gomphosphaeria spp.* model (figure F.4 part B) measures relatively poorly with $R^2 = 0.348$, U1 = 0.458, U2 = 0.901 and $\kappa \leq 0.34$. The time-series plot shows that dynamics of this genera are are largely featureless with occasional explosive events in 1983, 1987, 1990 and 1992. The ANN model is capable of predicting to some extent each of the dominant events. However, the magnitudes are generally somewhat under-predicted and there are several significant false positive predictions. Note that the extremely low prevalence of cell counts > 0 in this time-series meant that only 4 thresholds were defined. The classification statistics (table G.14) show that best performance was achieved at the highest threshold of 35000 cells/ml where very high specificity, but low sensitivity is observed. This indicates that the model tends to resist false positives at this threshold, but has a tendency to make false negative predictions. Table 5.4 shows that the perceptron model performs somewhat better than the ANN for this output.

The ANN forecasting *Microcystis aeruginosa* in Lake Kasumigaura is shown by the error measures to be one of the best performing models, with $R^2 = 0.626$, U1 = 0.301, U2 = 0.697 and $\kappa \leq 0.69$. Figure F.4 part C shows that the ANN is capable of forecasting the timing an magnitude of the bloom events observed in 1983, 1984, 1985 and 1992 with reasonable accuracy, although there appears to be a slight delay between the prediction and the observation. The model fails to predict the event in 1986, overpredicts in 1988 and underpredicts in 1989. However, false positive predictions in other years appear to be well resisted by the model. The classification statistics show best performance for the 3 intermediate threshold values (2000, 13000 and 65000 cells/ml) with high κ, sensitivity and specificity values. Sensitivity is reduced at the highest threshold of 160000 indicating an increased chance of false negative predictions. Table 5.4 shows that the ANN model performs significantly better than the perceptron for this output.

The *Oscillatoria spp.* model (figure F.4 part C) measures poorly with $R^2 = 0.045$, U1=0.595, U2=0.814 and $\kappa \leq 0.31$. The time-series plot shows that *Oscillatoria spp.* abundance is dominated by a single intense bloom ($\approx 5 * 10^5$ cells/ml) in late 1987 and some lower level dynamics in 1992-93. The model predicts the timing of the 1987 event well, but significantly underestimates the magnitude. The low level events in 1992 and 1993 are captured to some extent. A significant false positive prediction occurs in early 1983. The classification statistics show that the model performs best at the highest threshold value (40000 cells/ml), although performance at this threshold is characterised by low sensitivity indicating a likelihood of false positive predictions. Table 5.4 shows that the perceptron model performs significantly better than the ANN for this output.

5.3.1.5 Myponga Reservoir

The ANN forecasting chlorophyll *a* achieved reasonable measured performance with $R^2 = 0.436$, U1 = 0.247, U2 = 1.007 and $\kappa \leq 0.51$. Figure F.5 part A shows close correspondence between model output and observed values with every major event being predicted by the model and no significant false positive predictions. The model predicts magnitudes reasonably well, although there is generally an underprediction of peak events. The κ values (table G.17) show that the model performed well for most threshold values, although best performance was observed at intermediate thresholds (5.5 and 8.4 μg/L) where high sensitivity, specificity and PPP values were observed. At the highest threshold (12 μg/L), the sensitivity falls to 0.47 indicating an increased likelihood of false negative predictions. Interestingly, a U2 value of ≈ 1 indicates that the ANN does not achieve a better RMSE on validation data than the no-change model. The time-series plots indicate a slight delay between the onset of an observed bloom event and the model prediction indicating that the model is predicting some events too late. This delay may be the reason that the ANN performance is no better than the no-change model, which would have a similar delay in predicting bloom events. Table 5.5 shows that the perceptron model performs slightly better than the ANN for this output.

The *Ankistrodesmus spp.* model (figure F.5 part B) has very poor measured performance with $R^2 = 0.118$, U1 = 0.538, U2 = 0.942 and $\kappa \leq 0$! These numbers are reflected in the time-series plot that clearly shows that the ANN model has failed to generalise any of the observed dynamics in the time-series. Similarly, the *Dictyosphaerium spp.* model (figure F.5 part C) also measures very poorly with $R^2 = 0.029$, U1=0.570, U2=0.922 and $\kappa \leq 0.18$. Like the *Ankistrodesmus spp.* model, these measures are reflected in the time-series plot by a clear failure of the model to capture the observed dynamics. Table 5.5 shows that the ANN performs somewhat better than the perceptron for these outputs.

The *Scenedesmus spp.* model (figure F.5 part D) is somewhat better performing with $R^2 = 0.029$, U1= 0.507, U2=1.324 and $\kappa \leq 0.46$. The time-series plot indicates that *Scenedesmus spp.* are absent from Myponga Reservoir until 1990, after which they become an annual feature. The biggest peaks are observed at the end of 1991 and in 1993. Note that sampling of this variable was sporadic as indicated by the intense short term dynamics separated by long straight lines. The plot shows that the ANN captures the dynamics to a certain extent being able to predict the timing of most events. However the 1993 event is under-predicted whilst the 1994 event is significantly over-predicted. The classification statistics in table G.20 show that the model performs best at the highest threshold value of 10000 cells/ml with reasonable sensitivity and specificity values. Table 5.5 shows that the perceptron model performs significantly better than the ANN for this output.

5.3.1.6 Lake Soyang

The chlorophyll *a* model (figure F.6) has moderate performance with $R^2 = 0.281$, U1 = 0.389, U2 = 0.823 and $\kappa \leq 0.40$. The time-series plot shows that sampling of this lake intensified in 1995 as indicated by the increasing short term dynamics in the observed trace. The ANN model appears to perform well from 1995 onwards being able predict peak events on an annual basis, although the magnitude of the 1995 event is significantly under-predicted. Prior to 1995 the model appears to perform much worse in that it does not appear to model the observed dynamics. The classification statistics in table G.21 indicate that the model performs best at an intermediate threshold of 1.45 μg/L. Table 5.6 shows that the perceptron model performs slightly better than the ANN for this output.

5.3.1.7 Summary of Model Performance

Tables 5.1 to 5.6 show that R^2 values range from a minimum of ≈ 0.01 for 3 model outputs/datasets to a maximum of 0.626 for the ANN model forecasting *Microcystis aeruginosa* in Lake Kasumigaura. 14 out of 21 models had R^2 values of < 0.3 indicating that, for these models, over 70% of the prediction variance did not correspond with variance in the validation data. Only two models achieved R^2 values of > 0.5 – models predicting chlorophyll *a* concentration in Burrinjuck Dam and *Microcystis aeruginosa* in Lake Kasumigaura. A generally negative correlation was observed between U1 error rates and R^2, with higher U1 error tending to correspond to lower R^2 values. A minimum disagreement value of 0.248 was observed for the perceptron forecasting chlorophyll *a* in Lake Kasumigaura and a maximum of disagreement value of 0.642 was noted for the perceptron forecasting *Ankistrodesmus spp.* in Myponga Reservoir. In most cases, U1 fell between 0.3 and 0.5.

Tables G.1 to G.21 show that the Kappa statistic κ varied according to the threshold value, although there was no clear trend observed between κ and threshold. As with the continuous error measures, κ also varied according to dataset and output. In general, models shown to be better performing by U1 and R^2 values tended to have higher κ values as well. 13 out of 21 models achieved values of ≥ 0.4 for at least one threshold value indicating moderate agreement. Of these, 10 models achieved such a value for at least 1 of the 2 highest thresholds indicating that the model is a reasonable predictor of more extreme bloom events. Thus the remaining 11 models could be considered relatively unreliable for classifying such events.

Sensitivity of the models tends to drop as the bloom threshold rises, while the specificity rises. This means that with a higher threshold, models are less likely to correctly classify bloom cases and more likely to correctly classify non-bloom cases. The PPP (positive predictive power) of the models, like sensitivity, tended to drop as the threshold rises. This indicates that the likelihood of a positive

prediction being matched by an observation decreases with increased threshold. For many models, PPP dropped below 0.5 at the highest threshold indicating that positive predictions of the model were more likely to be incorrect than otherwise. It can be observed that where $\kappa \geq 0.4$, sensitivity, specificity and PPP values are simultaneously high, whereas if one of these statistics is ≤ 0.5, κ tends to be less than 0.4.

It can be observed that a subjective appraisal of the time-series plots corresponds to a certain degree with expectations of performance based on the U1 and R^2 error rates presented in tables 5.1 to 5.6 and the classification error rates presented in tables G.1 to G.21. Thus models with relatively low U1, high R^2 and good classification statistics tend to show correspondence between the traces for the observed and modelled data. Conversely, high U1 and low R^2 and κ values are generally associated with poor correspondences between the two traces. However, it can be observed that all models had clearly evident false negative and false positive predictions and that there were often discrepancies in the timing of bloom events.

U2 error rates are ≤ 1 for most models developed for the Biwa, Burrinjuck, Kasumigaura and Soyang datasets indicating that, in these cases, ANNs perform better on validation data than the naive "no-change" model. However, most ANNs trained on the Darling and Myponga datasets are characterised by U2 error rates ≥ 1.

In general, it can be observed that there is little correlation between U2 and the other unit-free error measures. The cause for this observation is that U2 depends on the RMSE of both the ANN and the naive "no-change" model (see equation 5.2). The performance of the no-change model is influenced by the degree of autocorrelation in the output variable which is, in turn, affected by both sampling density and the magnitude of short term dynamics experienced in the time-series. High sampling density and/or low short dynamics leads to increased autocorrelation and thus good no-change model performance. This in turn is likely to lead to increased U2 error rates. Conversely lower sampling density and/or high short term dynamics would be characterised by less autocorrelation, poor no-change model performance and thus lower U2 error rates. Table 3.21 compares the sampling intervals of the data retrieved to train and validate each model. It can be seen that the datasets for which the highest U2 error rates where observed, Myponga and Darling, also had the lowest median and mean sampling intervals with median values of ≈ 7 days compared to 15-30 days for the remaining datasets[1].

[1]Table 3.21 shows that the chlorophyll *a* model for Lake Soyang was also trained with data having a relatively low median sampling interval. However, the sampling density for this dataset abruptly changed approximately half way through the time-series leading to a higher mean sampling interval.

Tables 5.1 to 5.6 show that ANNs achieved superior performance to perceptrons for 10 out of 21 models according to the continuous error rates, while the perceptrons perform best for the remaining 11. This means that for approximately half of the models trained, there was a distinct advantage conferred by the ability to generalise non-linear relationships between inputs and outputs compared to a model constrained to linearity. This result contrasts with the results presented in chapter 4 where the perceptron generalised as well as, or better than, the ANN model for nearly all models. It can be observed that a hidden layer was advantageous for outputs for which the best generalisation was achieved in the case of the Biwa, Burrinjuck, Darling and Kasumigaura datasets.

In summary, the error measures and time-series plots show the following;

- The generic model structure generally achieves poor to moderately good performance.
- There were no cases where model performance was very good, since the error measures and time-series plots clearly indicate conditions of substantial prediction error in every case.
- There is little evidence of interaction between model performance and the dataset, since ANNs exhibiting both good and poor generalisation characteristics were observed for each site despite variations in sampling variability and density illustrated in table 3.21.
- There is little evidence of an interaction between output type (ie chlorophyll *a*, functional group, or species) and model performance.

5.3.2 Effect of Forecast Interval

As described in section 5.3.1.7, U2 error rates indicate that the comparative advantage of the ANN model over the no-change model tends to diminish as the sampling frequency of the output variable increases. Indeed, the results presented in tables 5.1 to 5.6 indicate that the no-change model would, in general, be a better guide to future algal abundance than the generic ANN model in cases where the median sample interval is $\approx$ 7 days. However, the comparison between the two models for these datasets may be unfairly biased against the ANN, because the ANN is constrained to making a 2 week forecast, whereas the no-change model defined makes use of the most recent observation of the output variable. Thus in the case of the Darling and Myponga datasets, the no-change model may have a forecast interval of 5 to 6 days compared with a 14 day forecast for the ANN.

To investigate the effect of forecast interval on U2 error rates, the forecast interval was reduced from 2 to 1 week(s) for all ANN models developed for the Darling, Myponga and Soyang datasets. Note that, according to table 3.21, a reduction in forecast interval is not supported by the sampling density of the Biwa, Burrinjuck

Table 5.7: Generic model error rates; Darling River – Comparing 7, 14 day forecasts.

Output	F.cast	No. obs	RMSE	U1	U2	R^2
total phytoplankton	7	510	22900	0.376	1.074	0.270
	14	508	24100	0.431	1.126	0.186
Chlorophyceae	7	510	4880	0.434	1.043	0.299
	14	508	5210	0.500	1.112	0.212
Flagellates	7	510	1710	0.343	0.897	0.410
	14	508	1840	0.363	0.975	0.340
Cyanophyce	7	510	5340	0.473	1.108	0.166
	14	508	5520	0.509	1.140	0.110

Table 5.8: Generic model error rates; Myponga Reservoir – Comparing 7, 14 day forecasts.

Output	F.cast	No. obs	RMSE	U1	U2	R^2
Chlorophyll *a*	7	421	4.34	0.247	1.062	0.476
	14	428	4.69	0.265	1.007	0.436
Ankistrodesmus spp.	7	103	1688	0.449	0.941	0.019
	14	108	3360	0.538	0.942	0.118
Dictyosphaerium spp.	7	221	1621	0.540	0.849	0.084
	14	213	1460	0.570	0.922	0.029
Scenedesmus spp.	7	146	14669	0.486	1.140	0.323
	14	136	19200	0.507	1.324	0.199

and Kasumigaura datasets without interpolation of existing data. The model design, inference and validation methodology was identical to that of the 2 week models (see section 5.2).

It can be seen from comparisons of continuous error measures in tables 5.7 to 5.9 that reduction of the forecast interval improves generalisation performance in most cases. Similarly, tables H.1 to H.9 show that classification error rates have also been improved. However, there are no cases where U2 error rates have been reduced to < 1 by the 1 week structure where they were > 1 for the 2 week structure. Thus it appears that with respect to a number of outputs, the generic ANN is likely to be out-performed by the no-change model when the median sampling interval of the dataset is ≈ 7 days.

Table 5.9: Generic model error rates; Lake Soyang – Comparing 7, 14 day forecasts.

Output	F.cast	No. obs	RMSE	U1	U2	R^2
Chlorophyll *a*	7	200	1.782	0.412	0.875	0.297
	14	222	1.79	0.389	0.823	0.281

5.3.3 Comparison with ANN Models from the Literature

All of the models developed in the present study are novel in terms of the model design, inference and validation method and the datasets used for training. However, there are a number of outputs/datasets modelled in the present study that have also been modelled by other authors using some machine learning method. Table 5.10 presents a comparison of these models with the generic models made on the basis of visual appraisal of time-series plots of predictions. The "best timing" and "best mag." columns show which model is able to best meet the timing and magnitude of bloom events respectively in the observed data, where "pub" indicates the published model, "generic" indicates the generic model developed in the present study and "equal" indicates that there is little difference in the performance of the two models.

It can be observed that the published models are generally considered to perform as well as, or better than, the generic models reported in the present study. However, table 5.10 highlights a number of key differences between previously published machine learning models and the generic ANN models;

- The generic model makes 14 day ahead forecasts, whereas all published models reviewed make same day predictions, except for Walter et al. (2001) where 7 day ahead forecasts are made.
- All models reviewed consider at least double the number of inputs of the generic ANN. They typically have access to data regarding zooplankton availability, solar radiation, micronutrient availability, pH etc. These extra variables are potential driving variables for algal growth.
- All models reviewed consider only 2 or 3 validation years instead of all the available data. Thus, it is possible that validation data has been "hand-picked" to show the best possible performance.

Thus it can be concluded on the basis of this evidence that;

- Choosing "case-specific" input layers that consider all possible driving variables leads to better generalisation,

  and/or

Table 5.10: Comparison of performance of ANN models in literature with generic ANN model.

Publication	Dataset	Output	Mod. type	F.cast	No. inputs	Val. years	Best timing	Best mag.
Bobbin and Recknagel (2003)	Kasumigaura	chlorophyll *a*	GP*	0	12	1986, 93	equal	equal
		Microcystis aer.	GP	0	12	1986, 93	pub	pub
		Oscillatoria spp.	GP	0	12	1986, 93	pub	pub
Recknagel et al. (1997)	Biwa	*Melosira gra.*	ANN	0	10	1986-87	equal	pub.
	Kasumigaura	*Microcystis aer.*	ANN	0	11	1986, 93	pub	pub
		Oscillatoria spp.	ANN	0	11	1986, 93	equal	equal
Recknagel et al. (1998)	Kasumigaura	*Gomphosphaeria spp.*	ANN	0	12	1986, 93	pub	pub
		Microcystis a.	ANN	0	12	1986, 93	pub	pub
		Oscillatoria spp.	ANN	0	12	1986, 93	pub	pub
Walter et al. (2001)	Burrinjuck	chlorophyll *a*	rec. ANN**	7 days	10	1979-82	generic	pub

* Genetic Programming
** Recurrent ANN

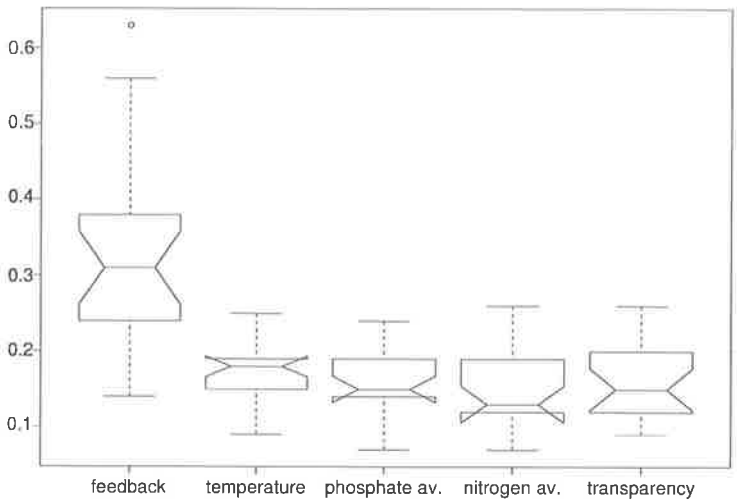

Figure 5.1: Comparison of absolute total sensitivity for each input – all models.

- Decreasing the temporal difference, as defined by the forecast interval, between input and output layers leads to better generalisation.

5.4 Sensitivity Analyses

Tables I.1 to I.21 show statistics regarding the sensitivity of each of the 21 models to the input variables. The "sens." column lists the proportion of the total average absolute sensitivity of the bagged model with respect to each input variable (see section 2.5.4 for a discussion on the sensitivity analysis procedure used). This value can be interpreted as the relative importance of inputs with respect to the driving the output variable. The correlation column shows the correlation coefficient calculated between the value of the input perturbation and the model response[2]. The sign of this value indicates whether the general relationship between the relevant input and output variable is positive or negative. The magnitude of this value reflects the likely linearity or "complexity" of the relationship. Values close to 1 indicate that the model's response to an input was relatively uniform with respect to the perturbation indicating a relatively linear, "simple" relationship with other modelled variables. Values close to 0 indicate that the model's response was highly variable indicating a non-linear, "complex" relationship. There are three likely sources of complexity;

- The relevant input has a non-linear relationship with the output variable.

[2]As explained in section 2.5.4, the sensitivity analysis procedure determined the model's response to 8 discrete levels, or perturbations of each input variable.

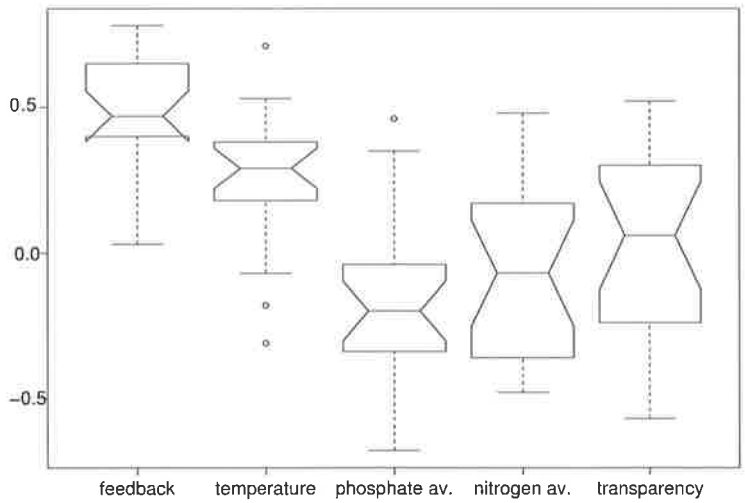

Figure 5.2: Comparison of R values for each input – all models.

- There is a non-linear interaction between the input variables with respect to their relationship with the output variables. For example, the model's response to high nutrient levels may be different at low and high temperatures.
- Variance between member models in the model ensemble caused by sampling errors or overfitting.

Figures 5.1 and 5.2 use boxplots to depict the distribution of sensitivity and R values presented in tables I.1 to I.21 with respect to each of the inputs of the generic ANN model[3]. Figure 5.1 shows that, in general, ANNs were most sensitive to the "feedback" or lagged output variable. On average, this accounted for 30% of the total observed sensitivity of the model. The models are significantly less sensitive to the remaining 4 model inputs (ie temperature, P, N, secchi). The boxplots indicate somewhat similar ranges of sensitivity for these inputs, although it may be tentatively concluded that water temperature and nitrogen availability are the most and least important inputs respectively. The data presented in tables I.1 to I.21 shows that model responses tend to be case specific – particularly for the less sensitive inputs.

Figure 5.2 shows that the feedback input and to a somewhat lesser extent, water temperature, had generally high positive R values indicating that increases of these inputs causes an increased prediction of algal abundance in the forecast period. Also, the fact that the absolute value of the R values were highest for these 2 inputs indicates that the relationship between these variables and the model output is likely to be a straightforward positive correlation. However, the lower R values

[3]The R values for turbidity in the case of the Darling River and Myponga Reservoir models in figure 5.2 were inversed, so that positive R values indicated an increase in algal biomass with higher transparency.

observed for temperature indicate that model response to this input are likely to be characterised by a greater degree of complexity than to feedback.

There were 3 exceptions with respect to the observed relationships with water temperature – flagellates (Darling river), *Gomphosphaeria spp.* (Lake Kasumigaura) and *Ankistrodesmus spp.* (Myponga Reservoir) all showed a negative response to this variable. This means that increasing temperature will lead to reduced predictions of algal abundance in the forecast period and, logically, that decreased temperature will lead to increased predicted values. The sensitivity analysis of the *Ankistrodesmus spp.* model can probably be disregarded, as the error rate data and time-series plots show a failure to generalise. Examination of the time-series plots of abundance for flagellates in figure F.3 part D and *Gomphosphaeria spp.* in figure F.4 part B reveals that most bloom events commence in late winter or spring and tend to dissipate in summer.

It can be observed that most models (16 out of 21) define a negative relationship between phosphorus availability and the output variable. This means that, for these models, increasing phosphorus availability will lead to reduced predictions of algal biomass in the forecast period. This observation is somewhat counter-intuitive, as it would be expected that the opposite relationship would be more likely since phosphorus availability has been identified as having pivotal role in precipitating eutrophication events. However, it is important to note that R values were close to 0 for a number of models indicating a complex relationship that may be dependent on other factors. An exception to these general findings was the model predicting chlorophyll *a* concentration in Lake Kasumigaura, which had an R value of $+0.46$ indicating a positive correlation.

The boxplots depicting the distributions of R values for nitrogen availability and secchi disk depth are wider than those for feedback and temperature and they extend over both positive and negative values. Tables 5.1 to 5.6 confirm that the R values for these inputs tends to be case specific. The fact that the median values indicated by the boxplots are close to 0 indicates that many models define complex relationships with respect to these two variables. One interpretation of this observation is that the model's responses to either of these two variables is dependent on the values of the other variables.

All Darling River models show a negative relationship with flow. Thus, as flow is increased, algal abundance in the forecast period of the model is reduced. This finding corresponds well with similar findings from ANN models of phytoplankton abundance in rivers reported by Recknagel et al. (1997); Maier et al. (1998) and Jeong et al. (2001).

5.5 Discussion and Conclusions

5.5.1 Performance of the Generic ANN Model

The evidence of the experimental results presented in this chapter show that the generic ANN model successfully generalises on independent validation datasets, since;

- The time-series plots show correspondence between the observations and predictions with respect to the timing and magnitude of some or most bloom events.
- In nearly all cases, the κ statistics indicate that the ANN is able to classify "bloom" and "non-bloom" events in independent data better than a naive random classifier.
- The continuous error rates show some level of agreement between predictions and observations in nearly all cases.
- A prediction advantage was conferred by the presence of a hidden layer for half of all ANNs trained indicating that non-linear interactions between modelled variables was being generalised.

However, despite these observations, the results indicate a number of shortcomings with respect to the performance of the generic ANN model.

- The success of the model clearly varies according to the model output and the dataset, with some models performing very poorly. Thus there can be no guarantee of a minimum performance level.
- No model achieves better than moderate performance. The time-series plots and error rate statistics clearly show significant prediction errors in every case.
- In conditions where the sampling frequency of the monitoring datasets is relatively high (ie $\approx$ every 7 days), a naive $y_t = y_{t-1}$ model performs better than the generic ANN in most cases.

Despite the observed variations in the performance of the model depending on the output and dataset, no clear trends were evident with respect to either of these factors. Thus it is not clear whether the generic model is better suited to a certain type of output. Nor is it clear that data availability, either in terms of the total number of records or the sampling frequency of the output, has a consistent effect on model performance.

However, the results from section 5.3.2 show that model performance is sensitive to the forecast interval defined as the time lag between inputs and outputs. Clearly, as the forecast interval is decreased, the model performance increases. The same

conclusion could be made on the basis of comparisons with ANN model applications in the literature (see section 5.3.3). Indeed, it has been shown by Maier et al. (1998) and Recknagel et al. (1998) that ANN models consideration of input variables with multiple lags is beneficial to performance.

Comparing the generic ANN with ANN models from the literature also points to potentially the most profound shortcoming of the model – that is, that it is too simplistic. By only including 5 to 6 input variables, it only considers a small subset of the possible forcing functions of phytoplankton growth in lakes and rivers. Thus the model provides no clear account of many factors likely to influence phytoplankton growth and community structures. For example;

- Top down control of algal abundance through grazing by zooplankton Hosper (1998); Gragnani et al. (1999).
- The effect of nutrients other than nitrogen and phosphorus, such as silica and trace elements.
- The effect of high pH on the abundance of certain species of cyanobacteria Shapiro (1990); Reynolds (1984).
- The effect of vertical and horizontal spatial variability on lake conditions. The generic model assumes uniform conditions and yet there known to be profound effects on phytoplankton communities caused by;
  - Thermal stratification imposing a physical barrier between regions of high nutrient and light availability Ganf and Oliver (1982); Reynolds et al. (1984); Burns (1994)
  - The onset of "over-buoyancy" of certain species of blooming cyanobacteria leading to intense concentration of abundance in surface waters Reynolds (1987).
  - Horizontal variability caused by wind, currents, lake morphometry, physical barriers etc.
- Relatively slow growth rates of cyanobacteria Reynolds (1987), meaning that longer term temporal links between input and output variables may need to be considered.
- Competition and/or mutualism exhibited between different species of phytoplankton.
- Human induced biomanipulations of lake ecosystems caused by herbicide dosing, introduction of fish, regulation of residence time, etc.
- The morphometry of the lake – the ecology of shallow, mixed lakes is different from deep stratified lakes.
- More detailed consideration of internal nutrient cycling, release from sediments etc.

Reduction of the size of the input layer to a "generic" model was an explicitly stated design decision in the case of the present study. Since the results show that the generic model is highly case-specific in its success, it may be concluded that the *ad-hoc* approach to model design evident in the literature may be justified. Chapter 6 investigates the hypothesis that ANN model performance can be improved by using a dataset and output specific approach to model identification.

5.5.2 Error Measures

General agreement between subjective performance evaluation, R^2, U1 and κ indicates that the unit free error measures are useful for quantifying model performance and for comparison of performance between model outputs and datasets. The classification error measures permit further characterisation of malpredictions by identifying conditions when error is dominated by false positive or false negative predictions. Furthermore, classification error rates showed that perceived performance varies according to the definition of the threshold at which a "bloom" condition is defined. This may be important should a model be implemented in an operational capacity, since resource managers may be specifically interested in the model's performance within certain critical ranges of the output variable. The U2 and the κ error rates serve a different function from the other error measures used in that they identify the model's performance compared to a "naive" predictor. The use of such comparative error measures provides a further degree of objectivity when assessing the models' performance.

5.5.3 Sensitivity Analysis

The sensitivity analyses method used allows quantification of both the relative importance of an input to the model and the complexity of the relationship that an input has with the other modelled variables. It should be noted that trust in the outcomes of sensitivity analyses is dependent on the quality of model generalisation. In general, it was found that the generic ANN model was mostly driven by the lagged output variable, indicating that previous algal abundance is the most important factor in determining current algal abundance. This is consistent with the structure of a typical time-series model. Also it was found that, in general, the models have a positive sensitivity to water temperature indicating that algal growth rates increase with increasing temperature. This is consistent with the observations that algal abundance (particularly cyanobacteria) is greater in the summer and autumn months.

Relationships with other modelled variables were shown to be more complex. This may be because the response of phytoplankton to nutrient availability and transparency is dependent on both water temperature and existing algal abundance. Interestingly, many models displayed a negative response to phosphorus

availability, indicating that increased growth follows a reduction in P levels. A hypothesis for this observation is that the model is considering consumption, where low P levels are indicative of high consumption and thus high growth rates.

Chapter 6

Identification of Lake Specific ANN Models

6.1 Introduction

It was concluded in chapter 5 that the generic ANN model has a number of shortcomings in terms of predictive ability. Specifically;

- There are significant prediction errors in every case according to both subjective and objective analyses.
- Performance is worse than the naive $y_t = y_{t-1}$ model when the sampling frequency of the output variable was ≈ 7 days.

It was hypothesised that a likely reason for the observed shortcomings in accuracy is that the generic model is too simplistic on two counts;

1. It considers a small subset of variables likely to be deterministically or correlatively linked with algal abundance.
2. It only considers a narrow range of lag data from 2 to 4 weeks prior to the output date.

With respect to the first issue, a range of variables present in existing datasets hypothesised to be linked to algal growth in some way were proposed in section 5.5.1. While some of these variables are not deterministically linked to specific phytoplankton abundance, Scardi (2001) has demonstrated that "co-predictors", that is, variables correlatively rather than causatively linked, can also lead to improved ANN models. With respect to the second of the above issues, Recknagel et al. (1998) and Maier et al. (1998) have shown that including inputs with a range of lag times relative to the output variable is useful to modelling outcomes. Specifically, given the slow growth rate of many species of cyanobacteria (Reynolds,

1987), it is reasonable to hypothesise that longer term lags than those used in the generic model may be predictive of algal abundance in some cases.

Since the set of variables that are *potentially* predictive of algal biomass is large, it is probable that ANNs configured with all the available inputs and a range of input lag times will be very large indeed. As a consequence, it may be hypothesised that model inference in the context of a large, case specific, ANN may be impacted by the *curse of dimensionality* (Bellman, 1961) – that is, that the exponential increase in the number of potential solutions with the increase in model size will make it increasingly difficult for the ANN to identify general relationships in the data. Indeed, Maier and Dandy (2000) point out that model identification is a particular problem when dealing with time-series applications, since there is effectively no upper limit to the potential number of lag times that may be utilised for a given input.

To combat the problem of model selection in the context of a large number of candidate input variables, a number of feature selection approaches have been devised, which have been briefly reviewed in section 2.3.1. For the present chapter, it was elected to compare the outcomes of a *supervised* and a *non-supervised* feature selection method, where the supervised method is guided by model error rates on independent data and the non-supervised method is guided by sensitivity analyses as an indicator of the model's internal structure after training. The non-supervised approach has the advantage that no access to independent data is required for model selection making it more data efficient to validate in a real world application. The two feature selection methods compared are;

- Data strip-mining (Embrechts et al., 2001). This method starts with all available inputs and performs a backwards elimination of inputs using input sensitivity as the goal function.
- Forward selection (Olden and Jackson, 2000). This method starts with a minimally configured input layer and iteratively adds variables using generalisation error as the goal function.

In investigating these model selection approaches, it is expected that the following hypotheses can be tested;

- Feature selection eliminates irrelevant variables leading to lower model error rates compared to complete model input layers.

  and

- The ability to select "case-specific" ANN models will bring improved performance over the generic models developed in chapter 5 since they can form links with a greater range of relevant variables.

The methods section describes the two feature selection methods employed and the the following sections describe the outcomes of experiments with models

comprising the same 21 output variables from the 6 datasets used for training the generic ANN models in chapter 5.

6.2 Methods

6.2.1 Data Strip-Mining

The strip-mining approach to model identification, explained by Embrechts et al. (2001), is straightforward. Figure 6.1 summarises the method. The starting model represents a set of variables hypothesised to be predictive of the output variable in some way. An ANN model is trained and a bootstrapped sensitivity analysis performed to estimate the relative importance of the input variables in predicting the output. Inputs deemed to be irrelevant to the model by the sensitivity analysis are dropped leaving a reduced feature set. The process of ANN training, sensitivity analysis and feature reduction is iterated until no inputs are shown to be irrelevant.

The advantages of this approach to model identification are;

- It is "model-free" in that selection is not biased by *a-priori* assumptions of relevance. Thus it is consistent with the spirit of machine learning.
- The goal function of the optimisation task is relative input sensitivity, not model error rate. This means that no access to target variables is needed eliminating the need for double cross-validation[1].
- Bootstrapping of the sensitivity analysis eliminates uncorrelated errors caused by sampling errors, ANN initialisation, local optima and other errors caused by random chance.

Whilst this methodology is straightforward, it is clear that outcomes are dependent on the following;

- The choice of variables in the initial "starting model".
- The accuracy of the sensitivity analysis technique at estimating the relative importance of input variables to an ANN model.
- The criteria used for determining whether or not an input is relevant to the model.

[1]Double cross-validation, as explained in section 2.3.4, involves the use of two independent datasets apart from training during model inference. The first set, the validation or *tuning* set, is used to calibrate model parameters, while the second set is used to estimate model performance on population data

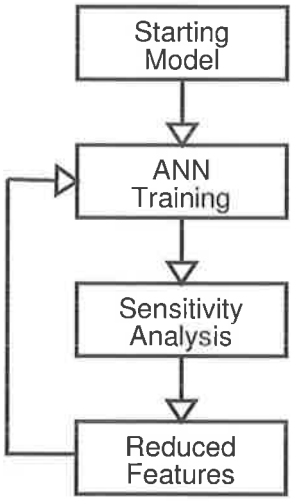

Figure 6.1: Methodology for model identification by strip mining.

6.2.1.1 The Initial Model

The choice of variables for inclusion as inputs in the initial model is constrained by the availability of data with matching measurement dates to the output variable. Data availability for each of the six datasets is reviewed in chapter 3. The identification procedure followed to determine initial models is outlined in section 3.4.2. The input layer designs of the initial model are illustrated in tables J.1 to J.6. The outputs are the same as those predicted by the generic model in chapter 4 (see table 3.20).

6.2.1.2 Feature Set Reduction

As in chapter 5, sensitivity analysis of trained ANN models was used to determine the relative importance of input variables in predicting the output. The sensitivity analysis method is described in detail in section 2.5.4. As explained previously, the sensitivity analysis method used is more comprehensive than previously used approaches, as it seeks to account for;

- Non-linear and non-monotonic relationships between inputs and outputs.
- Non-linear interactions between the effects of inputs on the outputs.

For each model, 50 ANNs trained with bootstrap sub-samples of the datasets were used to generate the sensitivity data. The input sensitivities of the bagged model was taken to be the average input sensitivities of the 50 model ensemble generated during the bootstrapping aggregation bagging procedure. This replication provides greater confidence in the outcome since uncorrelated errors caused by sampling errors, ANN initialisation etc are cancelled out.

Two dummy variables drawn from random uniform and Gaussian distributions respectively were included in each of the starting models. It was assumed that inputs with bagged sensitivity $\leq$ bagged sensitivity of the *least* sensitive dummy input variable is irrelevant to the ANN model. Thus feature set reduction was achieved by discarding these inputs.

6.2.2 ANN Model Identification by Modified Forward Selection

A dataset specific model was constructed for each dataset/output by means of a simplified forward selection procedure. Model selection by forward selection is carried out by adding every available model input in turn and selecting the most statistically significant for inclusion into the model (Olden and Jackson, 2000). Since the large number of models and potential input variables makes this process very computationally expensive in the context of the present study, it was elected to perform a simplified approach as follows;

1. The generic 5 input model (ie N, P, water temperature, secchi depth and a feedback input) was trained and validated for each dataset/output and the error rates observed (note that flow was also included in the generic model for the Darling River dataset).
2. The additional inputs available for each dataset were grouped according to the mechanism of their likely ecological impact. These groupings are listed in table 6.1.
3. Each of the input groupings was added to the generic model in turn. The models were trained and the error rates observed. Thus the performance of the generic+input 1, generic+input 2 ... generic+input n models was obtained to determine the performance impact of each input grouping.
4. A "combination" model including all the input groups that *improved* performance relative to the generic model on its own was constructed, trained and validated.
5. The outcome of this process is then the best performing model according to a vote of the 5 error measures observed.

The choice and number of inputs comprising each of the groupings outlined in table 6.1 is different for each dataset due to variations in data availability. Tables J.1 to J.6 outline the composition of each of the input groups for each dataset. Note that, due to constraints of time and computing resources, not all variables are allocated to input groups – instead only the variables deemed to be most likely to be relevant to determining dynamics of the output variable are included in the forward selection process.

Table 6.1: Functional groups of input variables for forward selection.

Dataset	Input grouping	Abbr.	Ecological Mechanism
Biwa	7–60 day lag generic inputs	+lag	long term dynamics
	3 species	+spe	competition and/or mutualism
	pH	+pH	cyanobacterial abundance
	weather	+wea	light availability, nutrient inflow, water column stability
	Silica	+Si	nutrient for diatoms
Burrinjuck	7–60 day lag generic inputs	+lag	long term dynamics
	3 algal functional groups	+spe	competition and/or mutualism
	weather	+wea	light availability, nutrient inflow, water column stability
	inflow	+inf	nutrient enrichment
	stratification	+str	separation of light / nutrient availability
	water depth	+dep	nutrient concentration, water column stability
Darling	7–60 day lag generic inputs	+lag	long term dynamics
	3 algal functional groups	+spe	competition and/or mutualism
	pH	+pH	cyanobacterial abundance
	Silica	+Si	nutrient for diatoms
Lake Kasumigaura	7–60 day lag generic inputs	+lag	long term dynamics
	3 algal genera	+spe	competition and/or mutualism
	pH	+wea	cyanobacterial abundance
	zooplankton	+zoo	grazing of phytoplankton
	weather	+wea	light availability, nutrient inflow, water column stability
	Silica	+Si	nutrient for diatoms
Myponga	7–60 day lag generic inputs	+lag	long term dynamics
	heavy metals	+hea	presence of algicide, redox environment
Soyang	7–60 day lag generic inputs	+lag	long term dynamics
	pH	+pH	cyanobacterial abundance
	weather	+wea	light availability, nutrient inflow, water column stability
	inflow	+inf	nutrient enrichment

It is important to note the differences between this approach and conventional "forward-selection" approaches generally applied in the context of multiple linear regression;

- The goal function in this case is validation set error rate, not some measure of input significance. This means that error estimates of the final model arising from this procedure may be optimistically biased, since the validation set data was effectively used in the model selection process.
- The grouping of inputs means that, as a model selection process, this method is somewhat "course". Better results may be achievable if inputs are added discretely to reduce the risk of irrelevant inputs being added to the model.
- The mandatory inclusion of the generic model as the "platform" from which the model selection process is initiated adds a deterministic component to the model selection process, since it is always assumed that these 5 inputs are fundamental in driving algal dynamics.

6.2.3 Model Inference, Validation and Computation

As in chapters 4 and 5, all models were based on three layer feed-forward multilayer perceptrons consisting of an input, hidden and output layers. As in chapter 4, ANNs with zero and 20 hidden layer units were compared to determine the importance of non-linear decision boundaries to all the modelled outputs. The methodology with respect to the architecture of neural processing, data conditioning, the maximum number of training epochs, validation and computation is identical to that used in the experiments in chapter 4 (see table 4.3). The Scaled Conjugate Gradient training algorithm was used for training all models. As in previous chapters, models were stabilised by bootstrap aggregation and by stopping training prior to convergence.

The comprehensive nature of the sensitivity analysis causes a very large quantity of data to be generated with each model replicate. This is because a model output is calculated for each input, record and input perturbation. For example, in the case of the Darling River models, there are 60 inputs, 388 sample dates and 8 input perturbations. This results in $60 * 388 * 8 = 186240$ records each replication. If the blocked 20-fold-crossvalidation method is used in combination with bagging, 50 replicates would result in $50 * 186240 = 9312000$ records to be stored and analysed making the experimental procedure data intensive. Thus, in order to "keep the lid on" the processing task, the leave-one-out bootstrap validation procedure (see section 2.5.3) with 50 replications was used instead of 20-fold-crossvalidation. In the case of the Darling River models, this brings the total number of records generated by the sensitivity analysis down to a more manageable level, since 50 ANNs are trained instead of $50 * 20 = 1000$ ANNs.

6.2.4 Experimental Treatments

The aim of the experiments conducted in this chapter is to contrast the modelling outcomes of the data strip-mining and forward selection model selection methods. The performance of all models is compared using the five error measures discussed in section 5.2.3, that is, RMSE, U1, U2, R^2 and the classification error. For the purposes of this experiment, classification error is expressed as average κ calculated for all classification thresholds defined for each output in chapter 5.

The outcomes of the data-strip mining procedure is, for each output, a series of ANN models with incrementally smaller input layer sizes until no input is shown to have equal or lower sensitivity than the two dummy variables. For the purposes of this experiment, the following models arising from this series are compared;

- The *starting model* – the ANN prior to removal of redundant inputs.
- The *first pass model* – the ANN after a single pass of the strip-mining procedure (ie removal of inputs identified to have lower sensitivity than the dummy variables).
- The *final pass model* – the ANN following iteration of the strip mining procedure until its conclusion (ie no inputs identified as having lower sensitivity than the dummy variables).

Performance of these three model types is compared with a generic 5-input model used as a control (defined in chapter 5) and the "combo" model identified by the forward selection procedure outlined above. This comparison is carried out for 21 model outputs from the 6 datasets outlined in table 3.24. Note that all results quoted are for models where dummy input variables have been removed and the model retrained and validated.

6.3 Experimental Results – Data Strip-Mining

Table 6.2 shows the number of feature reduction (or stripping) passes that was carried out for each of the modelled outputs before the minimum input sensitivity was greater than the sensitivity of the least sensitive dummy input. The number of stripping passes varied according to the model from a minimum of 2 for 5 different model outputs to a maximum of 16 in the case of *Scenedesmus spp.* in Myponga Reservoir.

6.3.1 Model Error Rates

Tables K.1 to K.6 compare the number of inputs variables and validation set error rates, using the *leave-1-out bootstrap estimator* and with the two dummy inputs

Table 6.2: Input summary periods for different sampling densities.

Dataset	Output	No. strips
Biwa	chlorophyll *a*	2
	Euglena americana	13
	Melosira granulata	4
	Pediastrum biwae	5
Burrinjuck	chlorophyll *a*	4
	Chlorophyta	9
	Cyanophyta	6
	Diatoms	2
Darling	total phytoplankton	3
	chlorophyta	4
	cyanophyta	5
	flagellates	2
Kasumigaura	Chlorophyll *a*	8
	Gomphosphaeria spp.	5
	Microcystis aeruginosa	5
	Oscillatoria spp.	8
Myponga	Chlorophyll *a*	2
	Ankistrodesmus spp.	4
	Dictyosphaerium spp.	2
	Scenedesmus spp.	16
Soyang	Chlorophyll *a*	5

removed, for all outputs given the starting models, the first pass model, the final pass model and the generic model. Note that the error rate in boldface type is the best error rate achieved by the input stripping process (with the generic model being excluded from the comparison).

6.3.1.1 Lake Biwa

In the case of the Lake Biwa models (table K.1), the starting model (ie all inputs) produced the best performing ANN for predicting chlorophyll *a* according to all error rates except average κ, which indicated the final model to be best. *Euglena americana* and *Melosira granulata* were best predicted by the "single pass" model, while *Pediastrum biwae* was best predicted by the "final pass" model.

The choice of input layer had a large effect on error rates for the *Euglena americana* model, with an $\approx 30\%$ difference between RMSE between the best and worst input layer. Significant effects of input layer choice were also observed for the chlorophyll *a* and *Pediastrum biwae* models with a 10-20% difference between the worst and best performing for many error rates. However, the *Melosira granulata* model appeared less affected by the choice of input layer. In general, least 2 of the "specific" type input layers brought better performance than the generic input layer for all model outputs.

6.3.1.2 Burrinjuck Dam

The error rates for the Burrinjuck Dam models (table K.2) show disagreement between different error measures in deciding the best input layer for all 4 outputs. In the case of the chlorophyll *a*, cyanophyta and diatom models, 4 out of 5 error measures indicate the first pass input layer as being the best performing. The final pass input layer was the best for the chlorophyta models according to 3 output of 5 error measures, with U2 and average κ preferring the starting model.

There was a 20-30% difference in performance between the best and worst performing input layer for all the modelled variables. Also, at least one of the specific input layer types performed better than the generic model for all outputs.

6.3.1.3 Darling River

The error rates for the Darling River models (table K.3) show disagreement between the error measures in deciding the best input layer for 3 out of 4 outputs. 3 out of 5 measures indicate that the starting model is best for the chlorophyta and cyanophyta models, while RMSE and average κ indicate that the first pass input layer is best. All error measures except $U2$ indicate that the starting model

is best for the flagellates model, while all 5 error measures indicate that the first pass model is best for predicting total phytoplankton.

Interestingly, there was no error measure that indicated that the final pass model was best for any of the modelled variables. In general the difference in performance between the starting and first pass models was relatively small ($< 10\%$) in all cases except according the the average κ measure that bigger differences in the case of chlorophyta, cyanophyta and flagellate models. By comparison, the final pass models were considerably worse performing. The specific input layers brought better performance than the generic inputs for all models according to all error measures with the exception of the flagellates model, where according to RMSE the generic model is best performing.

6.3.1.4 Lake Kasumigaura

The error rate comparison for Lake Kasumigaura (table K.4) show disagreement between the error measures in deciding the best input layer for all modelled variables. 4 out of 5 error measures showed the first pass models to be best for predicting chlorophyll *a*, *Gomphosphaeria spp.* and *Oscillatoria spp.*, with the dissenting measures being R^2 for chlorophyll *a* and RMSE for the remaining 2 outputs. 3 out of 5 error measures showed the starting model to be best for predicting *Microcystis aeruginosa*, with RMSE preferring the final pass model and average κ preferring the first strip model.

As with the Darling River models, the error measures overwhelmingly favoured the starting and first pass models, with only 2 error measures (out of 20) preferring the final pass model. The differences between indicated performance between input layers varied according to the output and the error measure, with significant differences in performance ($> 10\%$) observed between the worst and best models. Interestingly, the generic model achieved better measured performance than all specific input layers according to RMSE. However, the superiority of the generic model was not unanimous according to the remaining 4 error measures.

6.3.1.5 Myponga Reservoir

The error rate comparison for Myponga reservoir (table K.5) shows that the starting input layer brings best performance to the chlorophyll *a* model according to all 5 error measures. The first pass input layer is unanimously voted as best for predicting both *Dictyosphaerium spp.* and *Scenedesmus spp.*, while the final pass is favoured for the *Ankistrodesmus spp.* model by 4 out of 5 error measures (with RMSE preferring the starting model).

The differences in performance between the 3 specific input layers were relatively small according to most error measures for the *Ankistrodesmus spp.*, *Dictyosphaerium spp.* and *Scenedesmus spp.*. However, it is clear that the starting

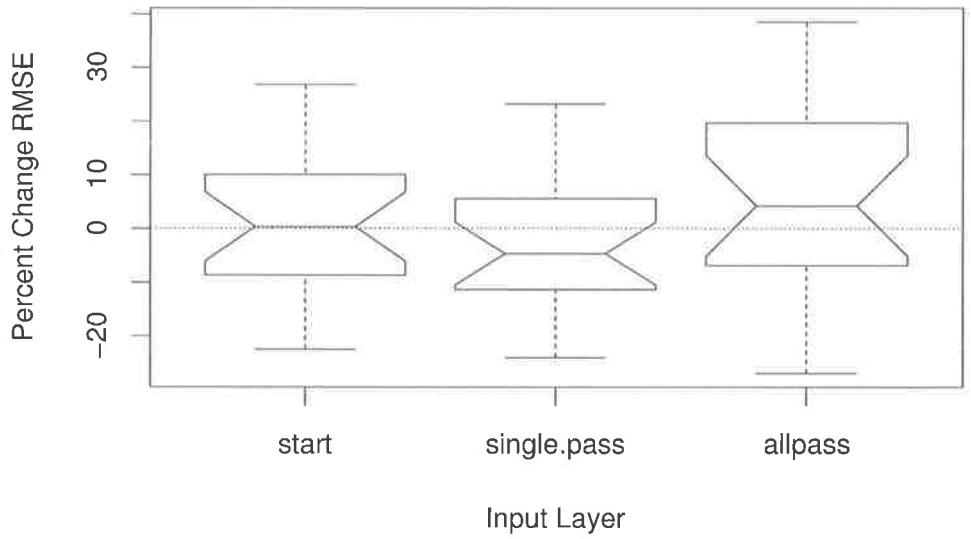

Figure 6.2: RMSE – grouped by data strip-mining model type.

model is superior by a significant margin for prediction of chlorophyll *a*. The specific model types generally do better than the generic model at predicting chlorophyll *a* and *Dictyosphaerium*. However, it is clear the generic model is considerably better at predicting *Ankistrodesmus spp.* and is similar in its ability to predict *Scenedesmus spp.*.

6.3.1.6 Lake Soyang

The error rate comparison for Lake Soyang (table K.6) shows that the first pass model is best according to all error measures except average κ, which prefers the starting model. The results clearly show that the generic model performs somewhat better than the first pass model, although the margin is small according to most error measures.

6.3.1.7 Effect of Input Layer

The box and whisker plots illustrated in figures 6.2 to 6.6 compare percentage change in model error rates from the generic model for the starting, single pass and final pass models with each plot illustrating the comparison for each of the 5 error measures used respectively. The results are grouped by input layer only to gain an overall comparison of the effect of the different input layers on model performance without considering the effect the output or dataset. Note that, for RMSE, U1 and U2, a lower number compared to the generic model indicates improved performance, whereas for R^2 and average κ, a higher number indicates improved performance.

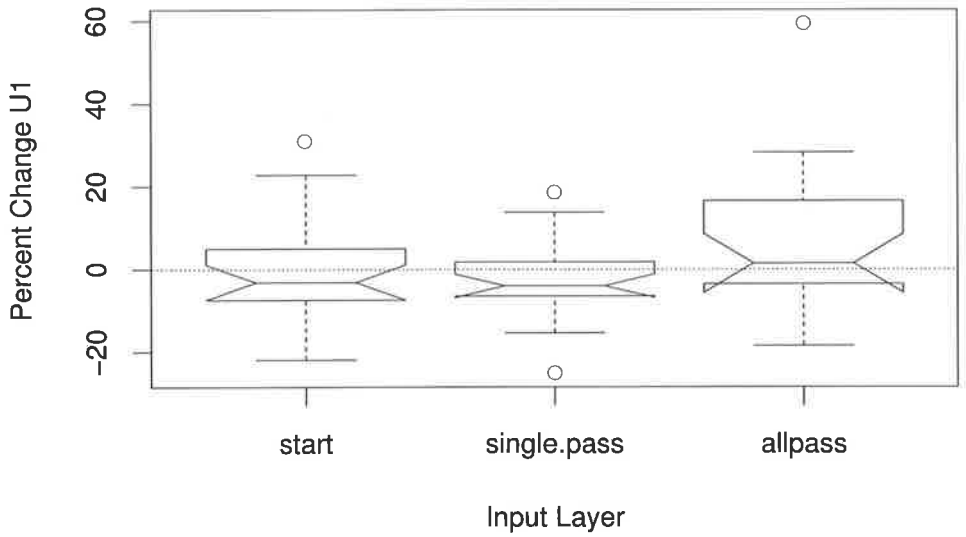

Figure 6.3: U1 – grouped by data strip-mining model type.

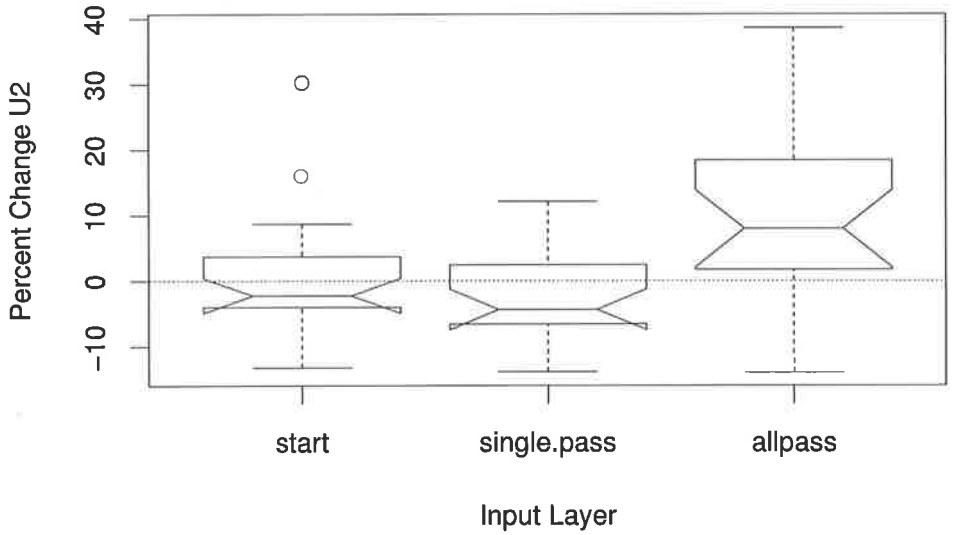

Figure 6.4: U2 – grouped by data strip-mining model type.

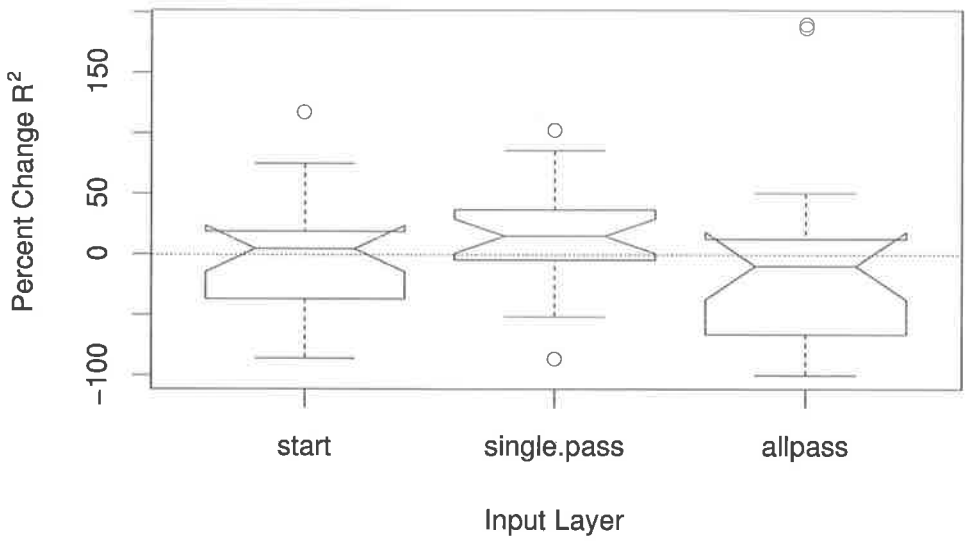

Figure 6.5: R^2 – grouped by data strip-mining model type.

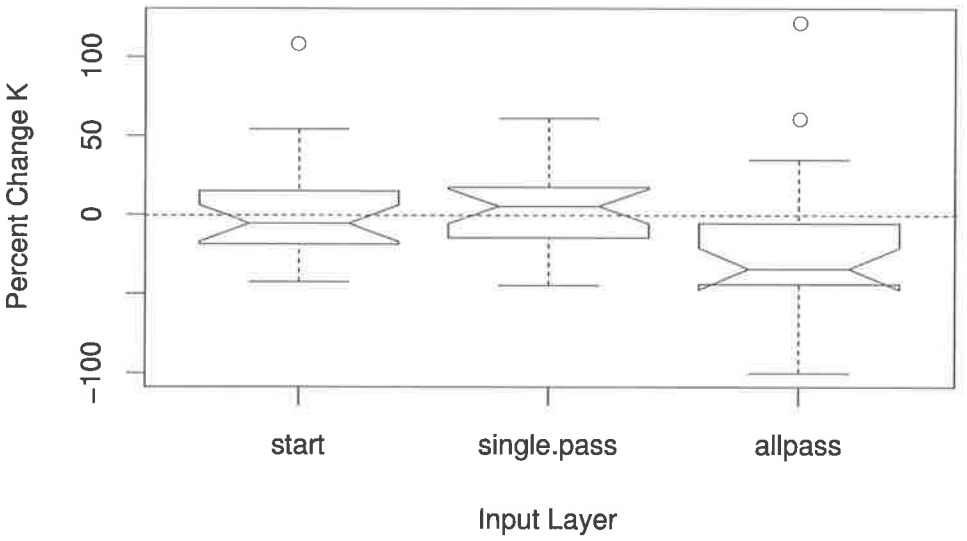

Figure 6.6: Av. κ – grouped by data strip-mining model type.

Figure 6.2 shows that, with respect to RMSE, the start model on average performed similarly to the generic model, while the single pass model performed better and the final pass model performed worse. With respect to the U1, U2 and R^2 error measures, (figures 6.3, 6.4 and 6.5 respectively), both start and single pass models perform better than the generic model on average whereas the final pass model performs worse. For average κ (figure 6.6), only the single pass model performs better than the generic model on average, while both start and all pass models perform worse.

The height of the boxplots illustrated in figures 6.2 to 6.6 show considerable variability in the results which can be attributed to interactions with the effects of the dataset and model output. This finding is reflected by the observations with respect to error rates made from tables K.1 to K.6 above. In the case of RMSE, U1 and R^2, the notches in all boxplots overlap meaning that the input layers are not significantly difference in terms of performance at the 95% confidence level. However, for U2 and average κ, it is clear that the final pass model performs significantly worse than the start and single pass models since the notches of the former box do not overlap those of the latter two. While the single pass model is consistently voted by the 5 error measures to be better performing than the generic model, it must be pointed out that in every case the box part of the box and whisker plot still straddles the dotted zero line meaning that at least 25% of all single pass models have worse performance. Thus it cannot be concluded categorically that a single pass of the input stripping procedure improves model performance.

In summary, the evidence of figures 6.2 to 6.6 and tables K.1 to K.6 supports the following conclusions;

- The single pass model tends to perform better than the generic, start and final pass model types. However, based on the current analysis, it cannot be concluded that such an effect is statistically significant.
- Allowing the input stripping process to iterate until no more inputs are indicated to be insignificant degrades error rates relative to other model types.
- There is considerable interaction between the effectiveness of the data strip mining procedure and the dataset / output.

6.3.2 Model Structure

Tables K.1 to K.6 show that, for each output, the stripping procedure significantly reduced the number of input variables in the model. Iterating the stripping process until all inputs were more sensitive than the dummy variables resulted in smaller input layers than the generic input layer for 11 out of the 21 models. In the case of models predicting *Oscillatoria spp.* in Lake Kasumigaura and cyanophyta

in Burrinjuck Dam, the "final" model had only a single input. The final model predicting total phytoplankton model in the Darling river was left with no inputs.

6.4 Forward Selection

Tables L.1 to L.6 compare performance of the different lake specific input layers constructed in the course of the forward selection exercise. The error rates in bold indicate the best performance for a given output, error measure and dataset. The last row of error rates for each model output is that of the "combo model" – that is, the model constructed using the generic inputs and the input groupings that lead to an improvement in performance over the generic model. Note that in cases where only one of the input groupings leads to improved results there is no combo model.

6.4.1 Model Error Rates

6.4.1.1 Lake Biwa

The results for lake Biwa (table L.1) show that according to 18 out of 20 error measures, a lake specific input layer leads to better performance than the generic model. In the case of the model predicting chlorophyll *a*, all 5 error measures showed that the input layer combining both species abundances and lag data was the best performing. The *Euglena americana* and *Melosira granulata* outputs were shown to be best predicted by the +species +lag combination by a vote of 3 and 4 out of 5 error measures respectively. The *Pediastrum biwae* model was shown to be best predicted by the +lag model according to 3 error measures, with U1 preferring the +species +lag model and average κ indicating the generic model. The combination models were not voted best by any of the error measures for any of the model outputs. In the case of models predicting *Melosira granulata* and *Pediastrum biwae*, many error rates indicated worse performance on the combination model than the generic model, despite the individual input groups leading to better performance than the generic model on their own.

6.4.1.2 Burrinjuck Dam

Table L.2 shows that the +lag models are voted best for prediction of chlorophyll *a* and cyanophyta by 4 error measures out of 5 in both cases. In each case, the dissenting error measure was average κ, which preferred the +species +lag model in the case of chlorophyll *a* and the +stratification +depth model in the case of cyanophyta. The +species model were indicated the best for prediction of chlorophyta and diatoms by all error measures except average κ again, with

the latter error measure voting for +lag in the case of the chlorophyta model and the combination model in the case of diatoms. The best specific models were generally only 5-10% better performing than the generic model in many cases. Furthermore, there were many specific models that actually performed worse than the generic model indicating that the Burrinjuck dataset is sensitive to the presence of irrelevant inputs. As with Lake Biwa, the combination model performed relatively poorly, with only a single vote out of 20 as best model.

6.4.1.3 Darling River

Table L.3 shows that the combination model was best for prediction of total phytoplankton by all 5 error measures. The +pH model was voted best for prediction of chlorophyta by 4 out of 5 error measures, although the margin of difference between it and the generic model is negligible with respect to all error measures. The +pH model and the +species model are equally as effective at predicting cyanophyta with 3 votes a piece (identical R^2 values were achieved). Both of these models show a modest but consistent improvement over the generic model. The +species input layer is best for prediction of flagellates according to 4 error measures, although the margin in performance between it and the generic and a number of other specific models is relatively small. Apart from the total phytoplankton model, the combination model did not appear to perform well relative to either the generic model or the other specific models.

6.4.1.4 Lake Kasumigaura

Table L.4 shows that the combination model is best for predicting chlorophyll *a* concentration by 4 out of 5 error measures (with average κ preferring the +zooplankton model). The margin in performance between the combination model and the generic model was significant $> 10\%$ according to most error measures. *Gomphosphaeria spp.* was best predicted by the +Si model according to 4 out of 5 error measures. This model was significantly better than the generic model and all other specific models except for the combination model (which included +Si). The *Microcystis aeruginosa* model was best predicted by the +lag model by a large margin according to 4 error measures, with average κ indicating the generic model to be best. *Oscillatoria spp.* is voted best by 3 error measures, with the remaining 2 error measures favouring the generic model. The combination input layer performed relatively well compared to the generic input layer for models predicting chlorophyll *a*, *Gomphosphaeria spp.* and *Microcystis aeruginosa*. However, the combination model for *Oscillatoria spp.* was significantly degraded relative to the generic model.

6.4.1.5 Myponga reservoir

Table L.5 shows that the combination model was best for three model outputs. The combination model was unanimously voted best for predicting chlorophyll *a*, with moderate margins of between 10 and 15% improvement in performance over the generic model. *Ankistrodesmus spp.* was best predicted by the +heavy metals input layer, although performance of all 3 models was very similar. Note that no combination model was trained for *Ankistrodesmus spp.* since the +lag model failed to register an improvement in performance over the generic model. The *Dictyosphaerium* model clearly favoured by the combination input layer gaining all 5 votes. As with chlorophyll *a*, there was a solid improvement in performance over the generic model. Very similar performance for all 3 models predicting *Scenedesmus spp.* was observed, although the combination model was indicated best by 4 out of 5 error measures.

6.4.1.6 Lake Soyang

Table L.6 shows that very similar performance at predicting chlorophyll *a* was achieved by all 6 input layers. The +pH model was considered best by 3 out of 5 error measures, with $U1$ indicating +weather to be best and average κ indicating the combination model to be best.

6.4.1.7 Effect of Input Layer

The box and whisker plots illustrated in figures 6.7 to 6.11 compare percentage change in model error rates from the generic model for the +species, +lag, +species+lag, +pH, +weather, +Si and combo models, with each successive plot showing the comparison for each of the error measures. The error rates in these plots are grouped by input layer only to gain an overall comparison of the effect of the different input layers on model performance without considering the effect the output or dataset.

Figures 6.7 to 6.11 show that, with respect to most error measures, the +lag, +strat, +heavy and the +combo models tended to perform better than the generic model. In these cases, the entire box is placed below the zero line indicating that at least 75% of the observed models achieved better validation set performance than the generic model. As shown in table 6.1, the +lag and +combo inputs are available for all datasets and outputs, while the +strat input group is exclusive to Burrinjuck and +heavy is exclusive to Myponga. Conversely, it is clear that the +inflow models (exclusive to Burrinjuck and Soyang) suffered reduced performance compared to the control. The remaining models did not appear to perform significantly differently on average from the generic model, since the boxplots straddle the zero line. However, in many cases, wide boxplots indicated a high degree of variability

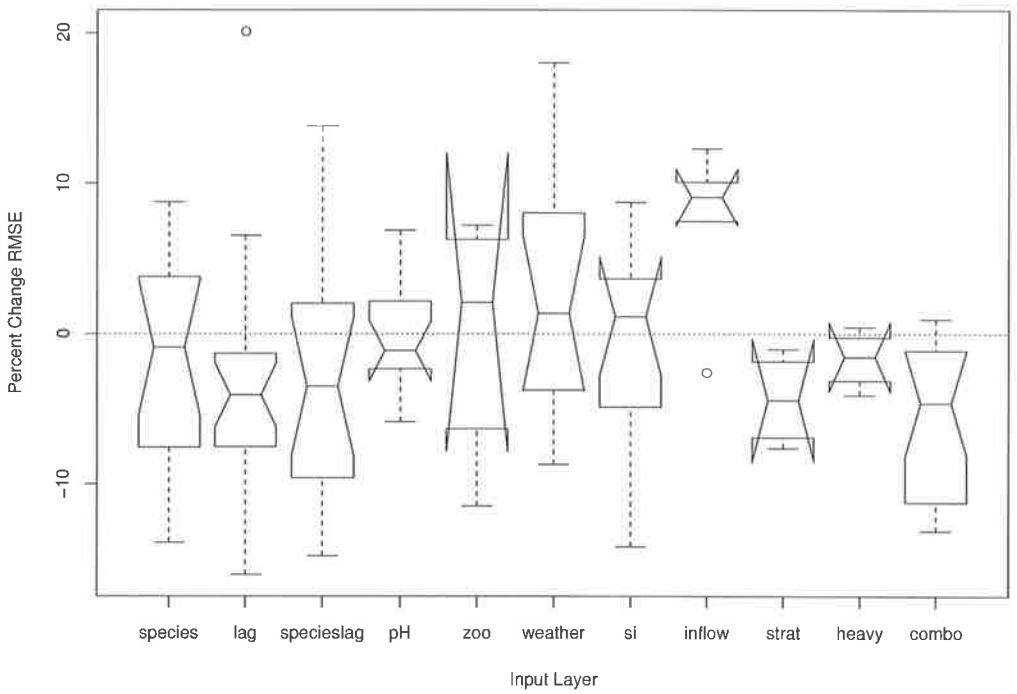

Figure 6.7: RMSE – grouped by forward selection model type.

in performance relative to the generic model that may be attributed to interactions with the dataset or the output variable.

Based on the evidence presented in tables L.1 to L.6 and figures 6.7 to 6.11, the following conclusions can be drawn from the forward selection experiment;

- Inclusion of long 7-67 day lags in addition to the short 7-x day lags brings a general improvement to the generic 5 variable ANN model.
- Inclusion of stratification and data for metal ion concentration variables in addition to the 5 input generic model appears to significantly improve ANN models for those datasets where data is available.
- Input groups proven to increase performance individually also significantly improve model performance when combined into a single model. However, this leads to a model that is dataset and output specific rather than generic.
- Inclusion of inflow variables appears to be highly detrimental to model performance where this data is available.
- A great deal of variability exists with respect to the effect of some input groupings such as zooplankton, weather and Si. Clearly, the presence of these groupings strongly interacts with the dataset or output variable.

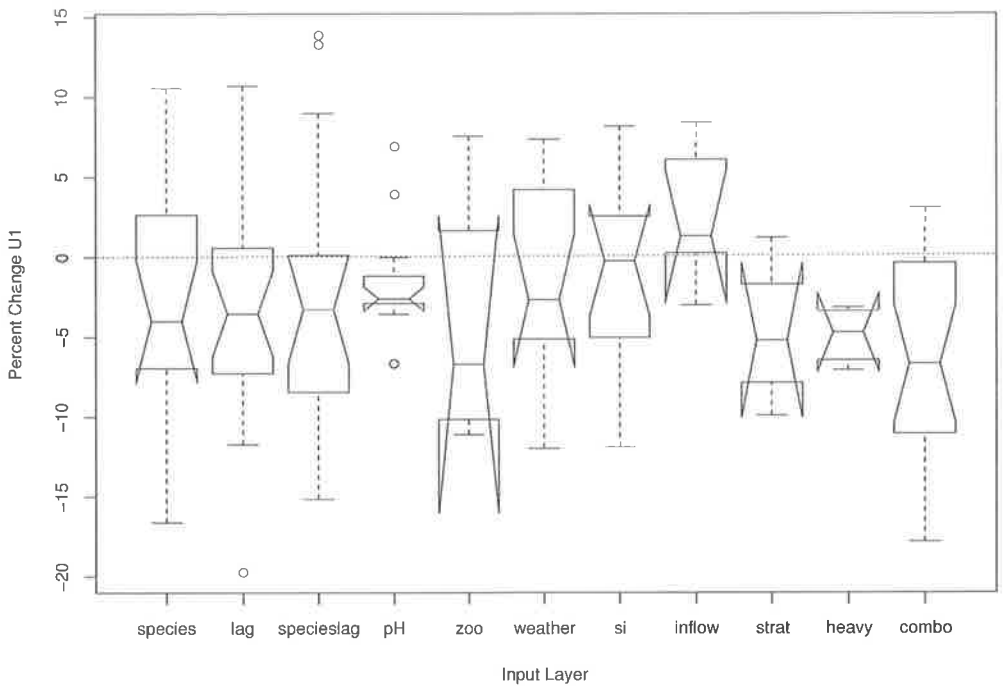

Figure 6.8: U1 – grouped by forward selection model type.

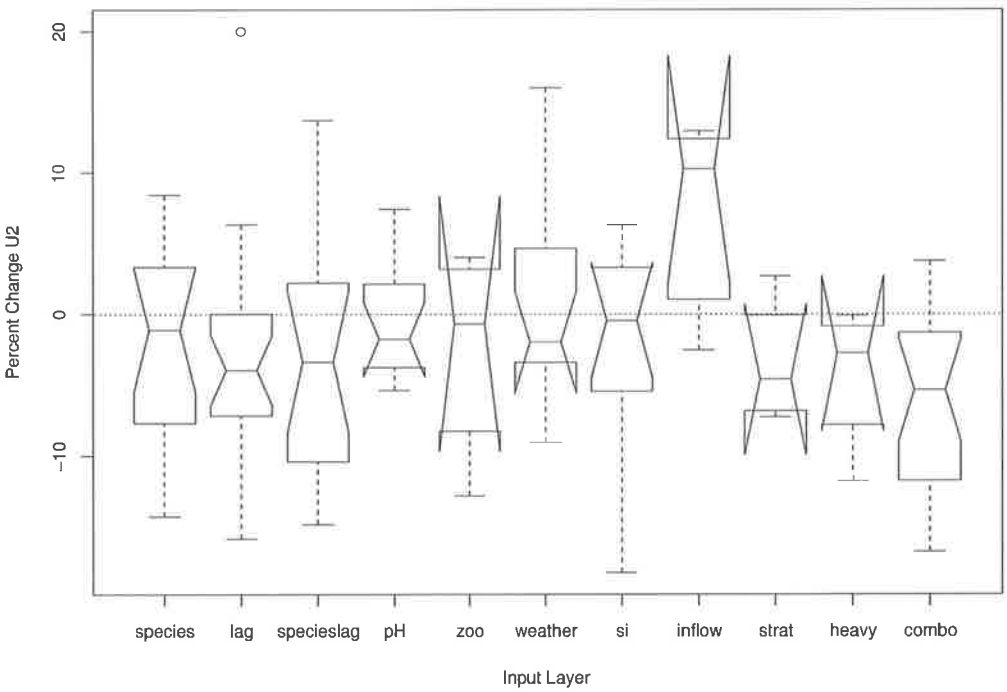

Figure 6.9: U2 – grouped by forward selection model type.

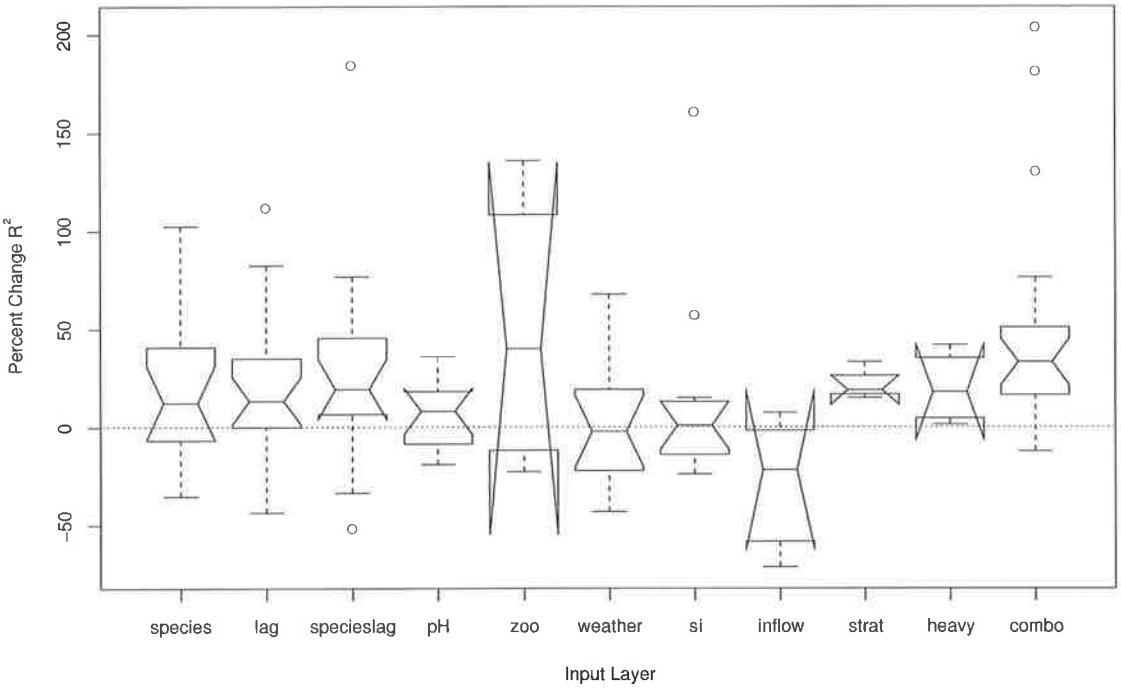

Figure 6.10: R^2 – grouped by forward selection model type.

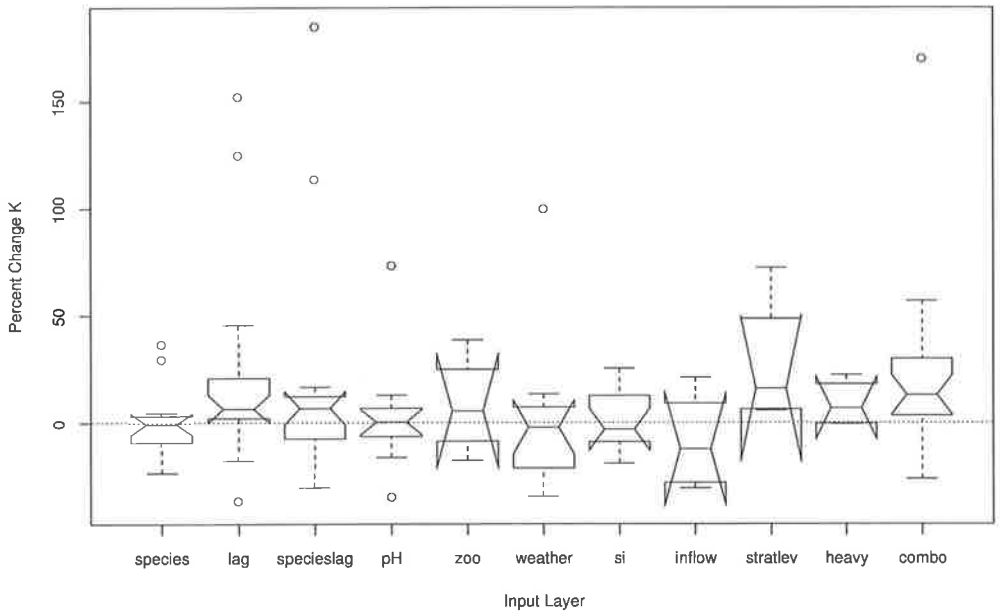

Figure 6.11: Av. κ – grouped by forward selection model type.

6.5 Comparing Performance of Model Selection Approaches

The box and whisker plots in figures 6.12 to 6.16 compare percentage change in model error rates from the generic model for the first pass data strip-mining model, the +lag, the +lag +heavy +stratification model (labelled "enhanced") and the combo model (labelled "specific"), with each successive plot showing the comparison for each of the error measures. The error rates in these plots are grouped by input layer to gain an overall comparison of the effect of the different input layers on model performance without considering the effect the output or dataset.

These plots show that all four of the models have median performance levels that are better than the generic 5 input model according to all error measures. The most improvement is achieved by the combo or specific model which clearly performs significantly better than the generic model. It also performs better than the other model types with a higher median improvement relative to the generic model according all error measures except RMSE. In general, the median improvement over the generic model for the specific model is in the order of 5-10% for RMSE, U1, U2 and average κ and 40-50% for R^2.

The single pass, +lag and enhanced models appear to perform very similarly in terms of median improvement over the generic model, with similar median improvements and overlapping notches. The improvement was approximately 5% for RMSE, U1, U2 and R^2 and 10-20% for R^2. The box-plots of the single pass model were wider than those for the +lag and enhanced models indicating greater variability in performance. This was particularly the case for RMSE. Also, the boxplots of the single pass model crossed the zero line for all error measures indicating a consistent agreement that at least 25% of this model type performed worse than the generic model. Thus it can be concluded that the data-strip mining method is somewhat less reliable in its outcomes than the models identified by forward selection.

6.6 Validation Set Performance of the Specific Model

Since it has been demonstrated that the forward selection approach identifies consistently better performing ANN models than the data strip mining approach, this section will discuss the performance of the specific models identified by forward selection. Figures M.1 to M.6 illustrate time-series plots of observed values and bagged specific model predictions of outputs on validation data. As with figures plotting predictions of generic models in chapter 5 (see figures F.1 to F.6), observations are joined by interpolated lines to emphasise the trajectory of the modelled and observed values through time. Unlike figures F.1 to F.6, the plots

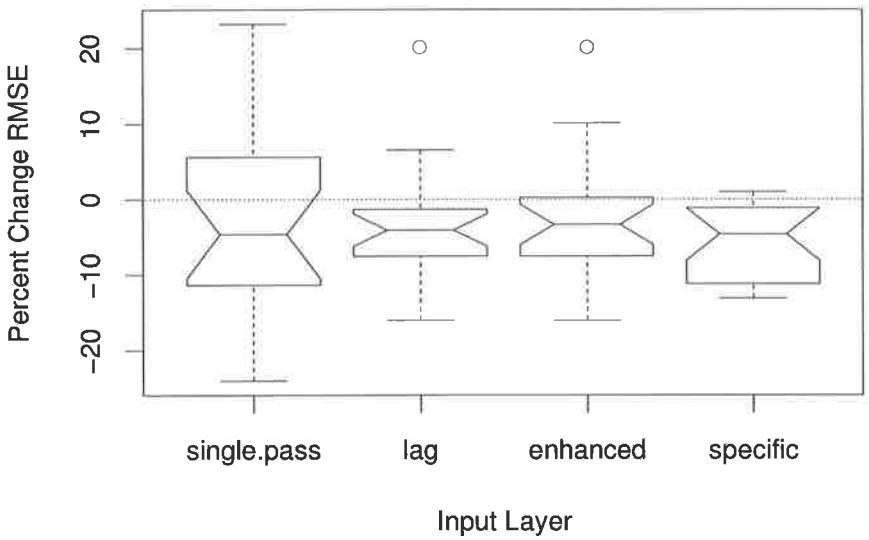

Figure 6.12: RMSE – grouped by model type.

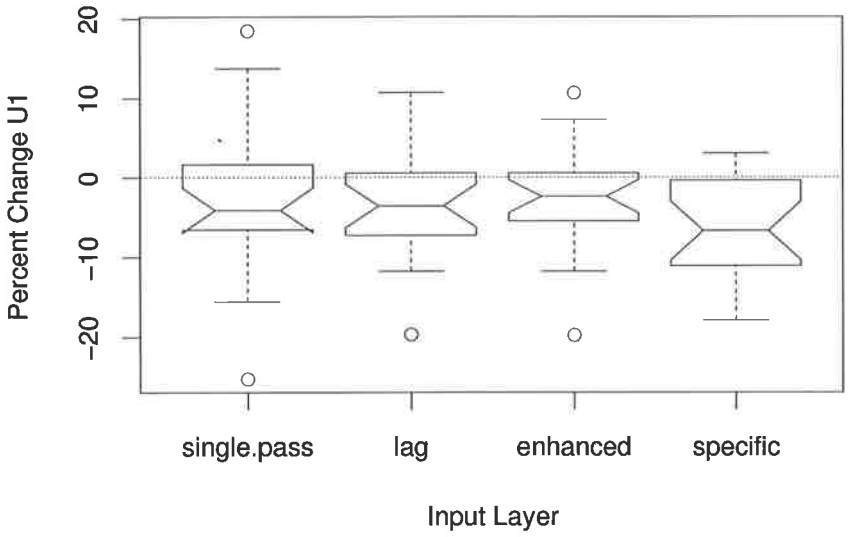

Figure 6.13: U1 – grouped by model type.

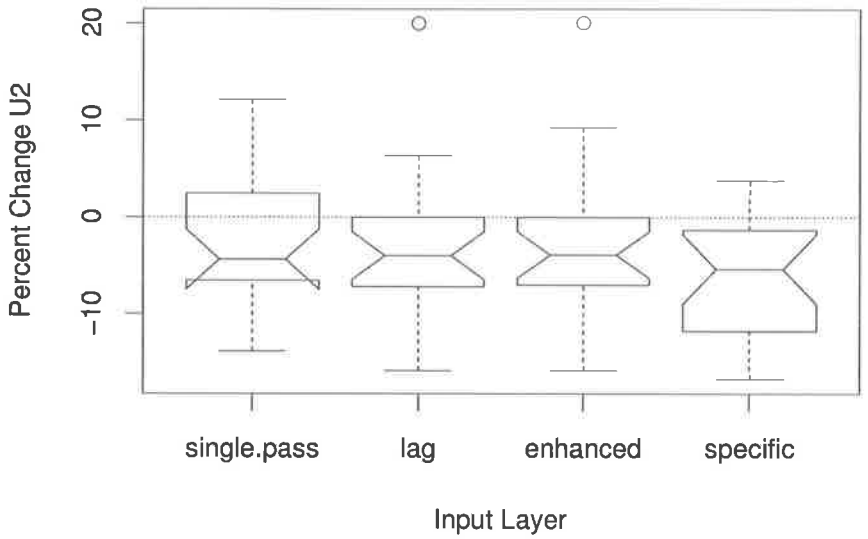

Figure 6.14: U2 – grouped by model type.

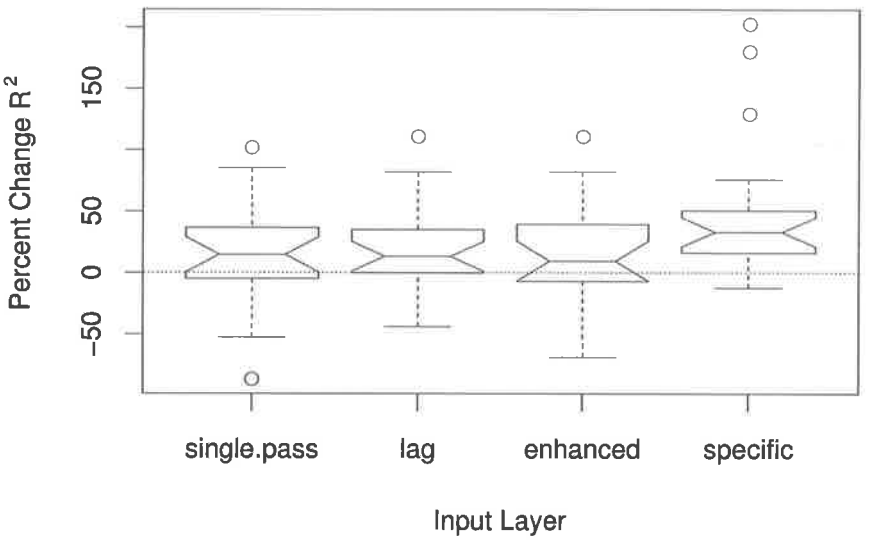

Figure 6.15: R2 – grouped by model type.

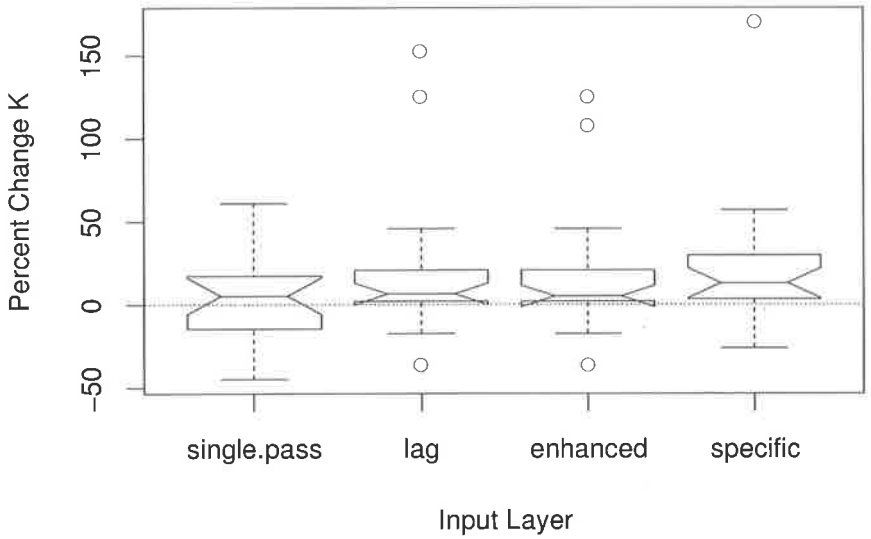

Figure 6.16: K – grouped by model type.

show model output where the leave-one-out bootstrap validation method has been employed instead of 20-fold cross-validation. The following section contrasts the subjective performance of the ANN models according to the time-series plots with the error rates for the combo models documented in tables summarised in table 6.3.

6.6.1 Model Performance Evaluation

6.6.1.1 Lake Biwa

The chlorophyll *a* model (figure M.1 part A) features mixed measured performance with good U1 and U2 (0.273 and 0.834 respectively) but poor R^2 and κ (0.2 and 0.324 respectively). The time-series plot shows that the ANN model captures the observed dynamics over time of the observed values, but that there tends to be under-predictions of extreme events (eg, in 1985, 1988) and some false positive predictions (eg 1986).

The *Euglena americana* model (figure M.1 part B) has poor measured performance (U1 = 0.558, R^2 = 0.123, κ = 0.286), although the model clearly performs better than the no-change model as indicated by a U2 error of 0.827. The time-series plot indicates this species to be characterised by explosive bloom events in spring and an absence of biomass the remainder of the year. The model appears able to match the timing of bloom events very well and there are no significant false positive predictions. The magnitudes of most events, except 1985 and 1990, appear well matched leading to the conclusion that the error measures are perhaps unrealistically harsh in this case.

Table 6.3: Error rates of specific (combo) model.

Output	Inputs	RMSE	U1	U2	R^2	Av κ
Biwa						
Chlorophyll *a*	16	6.07	0.273	0.834	0.200	0.324
Euglena americana	16	1780	0.558	0.827	0.123	0.286
Melosira granulata	18	562	0.422	0.882	0.290	0.426
Pediastrum biwae	14	557	0.604	0.953	0.094	0.324
Burrinjuck						
Chlorophyll *a*	23	18.9	0.383	0.608	0.431	0.338
Chlorophyta	16	3540	0.487	0.774	0.098	0.192
Cyanophyta	14	68400	0.588	0.778	0.122	0.334
Diatoms	20	1660	0.350	0.927	0.480	0.434
Darling						
Total phytoplankton	12	19300	0.312	0.901	0.480	0.568
Chlorophyta	8	4380	0.367	0.937	0.435	0.508
Cyanophyta	19	4850	0.402	0.998	0.326	0.468
Flagellates	10	1680	0.308	0.880	0.465	0.448
Kasumigaura						
Chlorophyll *a*	22	46.9	0.255	0.824	0.272	0.306
Gomphosphaeria spp.	15	23700	0.523	0.919	0.194	0.268
Microcystis aeruginosa	17	87500	0.307	0.704	0.599	0.536
Oscillatoria spp.	9	45200	0.502	0.940	0.222	0.348
Myponga						
Chlorophyll *a*	18	3.52	0.187	0.901	0.695	0.654
Ankistrodesmus spp.	14	2250	0.509	0.829	0.050	0.306
Dictyosphaerium spp.	20	1650	0.448	0.944	0.217	0.226
Scenedesmus spp.	20	10300	0.238	0.734	0.756	0.476
Soyang						
Chlorophyll *a*	6	2.09	0.367	0.946	0.274	0.278

The *Melosira granulata* model (figure M.1 part C) also has reasonably poor measured performance according to U1 and R^2 (0.422 and 0.290 respectively), but κ and U2 indicate reasonable predictive ability compared to naive models (0.882 and 0.426 respectively). The time-series plot shows that while the model correctly predicts the timing and magnitude of most bloom events reasonably well, there are several false positive predictions (1984, 1989) and a failure to meet the magnitude of some events (eg 1986, 1988, 1991). Like *Euglena americana*, this species can be considered reasonably difficult to model, since it is characterised by occasional explosive events with very little apparent autocorrelation of values.

The *Pediastrum biwae* model (figure M.1 part D) is not flattered by U1, R^2 and κ values of 0.604, 0.094 and 0.324 respectively, although U2 offers some redemption with a value of 0.953 indicating better performance than a no-change model. The time-series plot shows that the model accurately predicts the timing of the 4 major events occurring in 1984, 1985, 1986 and 1987. However, it fails to predict the magnitude of the 1984 and 1985 blooms. There are no significant false positive predictions by this model.

6.6.1.2 Burrinjuck Dam

The model predicting Chlorophyll *a* (figure M.2 part A) has mediocre measured performance according to U1, R^2 and κ with values of 0.383, 0.431 and 0.338 respectively. However, U2 = 0.608 shows that the model is significantly better than a naive no-change model. The time-series plot indicates good performance with the major events in 1979-80, 1982 and 1983 being well matched and a lack of false positive predictions in the remainder of the time-series. However, the model fails to predict the peak of the 1980 event.

Chlorophyta (figure M.2 part B) is not well handled by the model as indicated by both poor measured performance and poorly matched predictions with observations in the time-series plot. U1, R^2 and κ (0.487, 0.098 and 0.192) highlight the inability of the ANN to correctly model this output. However, the highly dynamic nature of the observed cell counts over time means that the no-change model also performs very poorly resulting in a relatively good U2 error rate of 0.774. The time-series plot shows that many of the large scale dynamics are not captured (eg 1983) although some smaller scaled events later in the time-series appear matched quite well (1992 – 1996). The presence of long straight lines in the observations in the years 1979, 1983 – 1991 indicate long intervals between measurements at these times. It appears that more frequent and regular sampling from 1992 onwards had a positive effect on model predictions during this time.

The cyanophyta model (figure M.2 part C) again achieves poor measured performance according to U1, R^2 and κ (0.588, 0.122 and 0.334 respectively) but good U2 performance (0.778) indicating better RMSE than the corresponding no-change model. The time-series plot shows that the "shape" of the cyanophyta

dynamics over time is very similar to that of chlorophyll *a* (part A) with significant bloom events in 1979-80, 1982 and 1983 and very little activity for the rest of the time-series. This suggests that, in terms of abundance, Burrinjuck Dam is dominated by cyanobacteria. The model matches the 3 observed events reasonably well in terms of timing, but fails to predicting the magnitude of the 1980 event. Also, there is a significant false-positive prediction occurring over much of 1983.

The diatoms model (figure M.2 part D) achieves reasonable measured performance with U1, U2, R^2 and κ values of 0.35, 0.927, 0.48 and 0.434 respectively. Similarly, the time-series plot shows good performance, with the dynamics of algal abundance well predicted – particularly in terms of short term dynamics. Despite this, there are significant prediction errors that evidently preclude a better measured performance such as a false positive prediction in 1984 and false negative predictions in 1980 and 1995.

6.6.1.3 Darling River

The total phytoplankton model (figure M.3 part A) is characterised with good measured performance (U1 = 0.312, U2 = 0.901, R^2 = 0.480 and κ = 0.568). The time-series plot indicates that the model is highly accurate at predicting the observed dynamics with little deviation between the model prediction and the observation. However, it is possible to note underprediction of extreme events in 1980, 1981 and 1988. Also, a slight delay is evident in the model's prediction of the onset and decay of the 1980-81 bloom event.

The chlorophyta model (figure M.3 part B) achieves reasonable, but not excellent, measured performance (U1 = 0.367, U2 = 0.962, R^2 = 0.435, κ = 0.508). The time-series plot indicates that the model captures all the major dynamics in the observed variable over time. However there is underprediction of the peak events in 1980-81 and 1988. Also, as with the total phytoplankton model, the ANN appears to have a slight time delay in prediction of the more significant events.

The cyanophyta model (figure M.3 part C) achieves mediocre measured performance according to U1, U2 and R^2 (0.402, 0.998 and 0.326 respectively) but a good κ value of 0.468 suggests reasonable classification performance. However, cursorary examination of the time-series plot suggests very good performance, as all major dynamics appear well matched and there are no false positive predictions. However, more careful examination reveals under prediction of the peak events in 1981 and a delay in predicting the onset and decay of the 1981 event.

The flagellates model (figure M.3 part D) has reasonably good measured performance with U1 = 0.308, U2 = 0.880, R^2 = 0.465 and κ = 0.448. This is reflected by the time-series plot which shows that the model appears to predict the many and large short term dynamics of this variable highly accurately. However, as with the other models for this dataset, some peak events, such as 1982 and 1987-88 are under-predicted.

6.6.1.4 Lake Kasumigaura

The chlorophyll *a* model (figure M.4 part A) performs well according to U1 and U2 (0.255, 0.824) but poorly according to R^2 and κ (0.272, 0.306). The time-series plot shows that chlorophyll *a* dynamics are characterised by an annual periodicity with peak values in spring autumn and troughs in winter months. These events are punctuated by many shorter term dynamics. The ANN model appears to capture the annual periodicity well, with the onset and decay of chlorophyll *a* being well timed. However, shorter term dynamics are not always predicted – for example, the bloom events in 1983 and 1986.

The *Gomphosphaeria spp.* model (figure M.4 part B) has quite poor measured performance with U1 = 0.523, R^2 = 0.194 and κ = 0.268. However, a U2 value of 0.919 indicates that it still performs better than the no-change model. The time-series plot shows that all bloom events are predicted to some extent by the model. However, the 1983, 1986 and 1990 events are under-predicted and their onset is predicted late. There are no significant false positive predictions.

The *Microcystis aeruginosa* model (figure M.4 part C) has relatively good measured performance (U1 = 0.309, U2 = 0.704, R^2 = 0.599, κ = 0.536). The time-series plot shows that the timing and magnitudes of bloom events are predicted reasonably well, although there is underprediction of the severe 1986 bloom. The model makes a false positive prediction in 1991 when it incorrectly estimates that a bloom characterised by cell counts of 200,000 cells per ml will occur.

The *Oscillatoria spp.* model (figure M.4 part D) has poor measured performance according to U1, R^2 and κ (0.502, 0.222, 0.348), but a U2 value of 0.940 shows that the ANN still makes better predictions than the no-change model. The time-series plot shows Oscillatoria to be a particularly difficult output to model, since there is zero abundance for much of the time-series punctuated by occasional bloom events. In particular, the time-series is dominated by a single, very severe, bloom in 1987-88. Despite the poor measured performance, the plot shows that all events are predicted, with the latter smaller events from 1990 onwards being handled reasonably well. The 1987 event is predicted a month late and the magnitude is under-estimated. Also, there is a significant false positive prediction for 1983.

6.6.1.5 Myponga Reservoir

The chlorophyll *a* model (figure M.5 part A) measures very well according to U1, R^2 and κ (0.187, 0.695, 0.654). However, a U2 value of 0.901, while indicating better performance than the naive model, does not reflect the excellence of the other measures. The time-series plot mirrors the measured performance showing that the model is highly accurate in its ability to forecast the dynamics of this variable.

By contrast, the *Ankistrodesmus spp.* model (figure M.5 part B) is very poor performing with U1, R^2 and κ values of 0.509, 0.050 and 0.306 respectively. The U2 value of 0.829 shows that the ANN still performs well relative to the no-change model. The time-series plot shows that the ANN captures some of the short term, smaller dynamics, but fails to correctly model larger events in 1989 and 1990.

Similarly, the *Dictyosphaerium spp.* model (figure M.5 part C) measures poorly with U1 = 0.448, U2 = 0.944, R^2 = 0.217 and κ = 0.226. The time-series plot reflects the sporadic sampling of this variable with periods of very high short term dynamics indispersed with long straight lines indicating a lack of sampling. The ANN is able to capture some of the dynamics, but under-predicts peak events and makes a number of false positive predictions.

The *Scenedesmus spp.* model (figure M.5 part D) is one of the best performing ANNs with U1 = 0.238, U2 = 0.734, R^2 = 0.756 and κ = 0.476. The time-series plot reflects the good measured performance of the model showing that all events are very accurately forecast and no significant false positive predictions.

6.6.1.6 Lake Soyang

The chlorophyll *a* model (figure M.6) has reasonably poor measured performance with U1 = 0.379, U2 = 0.963, R^2 = 0.236 and κ = 0.294. The time-series plot shows that the sampling regime abruptly changes in 1995 at which point it becomes much more dense. Performance before 1993 appears mediocre, with most dynamics not well modelled. In 1993 and 1994, the model successfully forecasts the major dynamics, although it is characterised by early predictions. From 1995 onwards, the performance of the model appears to improve dramatically with the very short term dynamics being handled well and a lack of significant false positive predictions.

6.6.1.7 Summary of Model Performance

Table 6.3 shows that U1 values ranged from a minimum of 0.187 for the model predicting chlorophyll *a* in the Myponga reservoir, to a maximum of 0.604 for *Pediastrum biwae* in Lake Biwa. Overall, 4 models had U1 values < 0.3, 11 were < 0.4 and 6 were > 0.5. R^2 values ranged from a minimum of 0.050 for *Ankistrodesmus spp.* (Myponga) to a maximum of 0.756 for *Scenedesmus spp.* (Myponga again). 12 models had R^2 values of < 0.3 indicating, in these cases, that over 70% of prediction variance does not correspond with variance in validation data. This is a slight improvement over the generic model, where 14 models had R^2 values < 0.3. 3 models had good R^2 values > 0.5. Average κ ranged from a minimum of 0.192 for chlorophyta in Burrinjuck Dam to a maximum of 0.654 for chlorophyll *a* in Myponga Reservoir. 9 out of the 21 models had average κ values

> 0.4 indicating that, for these models, good performance was achieved relative to a random classifier for a range of threshold values.

In general, measured performance according to U1, R^2 and average κ appeared to be partially correlated with subjective visual appraisal of the time-series plots, with the good performing models appearing to forecast the timing and magnitude of bloom events well and to be resistant to false positive predictions. However, models ranked harshly by the error measures, such as *Pediastrum biwae* and *Euglena americana* in lake Biwa for example, did not appear to be particularly bad performing according to the time-series plots. Also, some models, such as chlorophyll *a* in Soyang and chlorophyta in Burrinjuck, were shown by the prediction plots to perform well over part, but not all, of the time-series. This indicates that the error measures, while useful as a general guide to performance, cannot be substituted for a time-series plot in these cases for characterising model performance.

U2 values were < 1 for all specific model outputs. This indicates that the specific ANN models perform better than the naive no-change model, suggesting that the ANN is able to model the processes of growth and decay over time to some extent. As was found for the generic models, U2 does not always correlate well with other error measures or subjective performance of the time-series plots – a corollary of being dependent on the RMSE of the naive model in addition to the ANN. These results are a significant improvement over the generic models described in chapter 5, where U2 values > 1 were observed for 3 outputs in the Darling dataset and 2 outputs in the Myponga dataset.

6.6.2 Interaction of the effects of Input Layer, Hidden Layer and Validation Method on Model Performance

It was concluded from experiments carried out in chapters 4 and 5 that;

- Model error estimations may be affected by the type of validation method used.
- A perceptron (ie 0 hidden unit ANN) should be used as a control for all ANN modelling exercises to determine the effect of non-linear processing properties on model performance.

To determine the effect of validation method and the presence / absence of a hidden layer on the performance of the specific ANN model identified by forward selection, a factorial experiment was run with the following treatments;

- Model input layer – specific (ie combo) and generic.
- Hidden layer – 0 and 20 units.

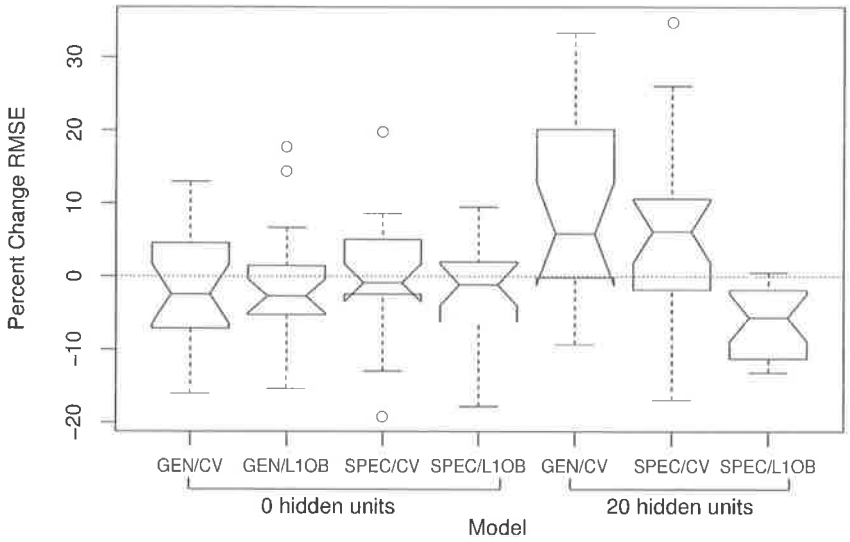

Figure 6.17: RMSE – comparing input layer, hidden layer, validation mode.

- Validation model – 20-fold blocked cross-validation and the leave-one-out bootstrap.

Thus, there was effectively 8 model treatments (2 input layer types * 2 hidden layer configurations * 2 validation modes).

Model inference was carried out as previously and error rates were calculated for the bagged model comprised of 50 member ANNs. The same 21 output/dataset combinations were used for training the ANNs as previously. The error rates for the treatments were compared by calculating the percentage change in error from an ANN control configuration for each model output/dataset. The designated control model was the generic 5 input ANN model for each output with a 20 unit hidden layer and validated by means of the leave-one-out bootstrap estimator. The box and whisker plots in figures 6.17 to 6.21 compare the percentage change in error rate from the control model for each of the respective error measures. The x-axis denotes the model treatment, where "GEN" and "SPEC" refer to generic and specific models respectively and "CV" and "L1OB" refer to 20-fold blocked cross-validation and the leave-one-out bootstrap respectively. Note that the box and whisker plots only show 7 treatments, since one was used as a control against which to compare the results.

It can be seen from figure 6.17, that in terms of RMSE, there appears to be an interaction between the effect of the hidden layer configuration and the effects of the other treatments. As was observed in chapter 4, the presence of a hidden layer makes model error rates sensitive to the effects of the remaining treatments, whereas for the perceptron models, the other treatments have little effect. In general, the perceptron models have similar RMSE to the control model. The 20 unit ANN models, on the other hand, respond to both validation method and

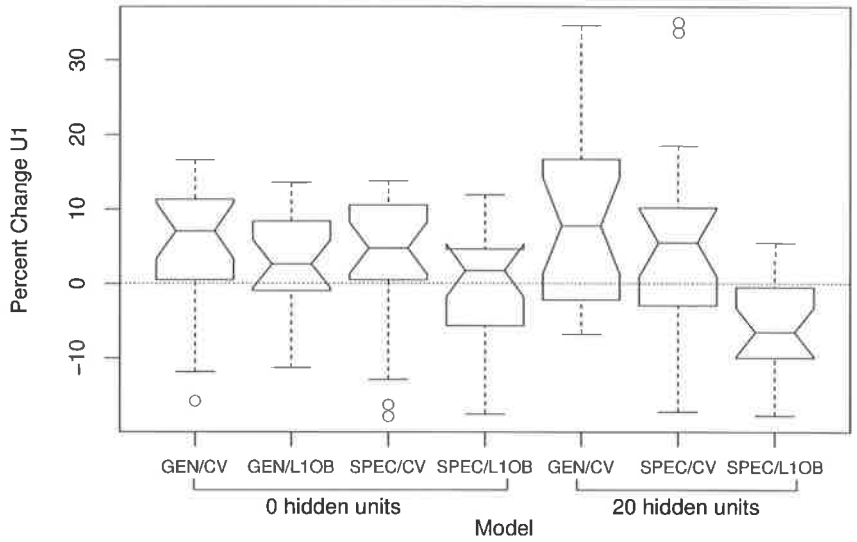

Figure 6.18: U1 – comparing input layer, hidden layer, validation mode.

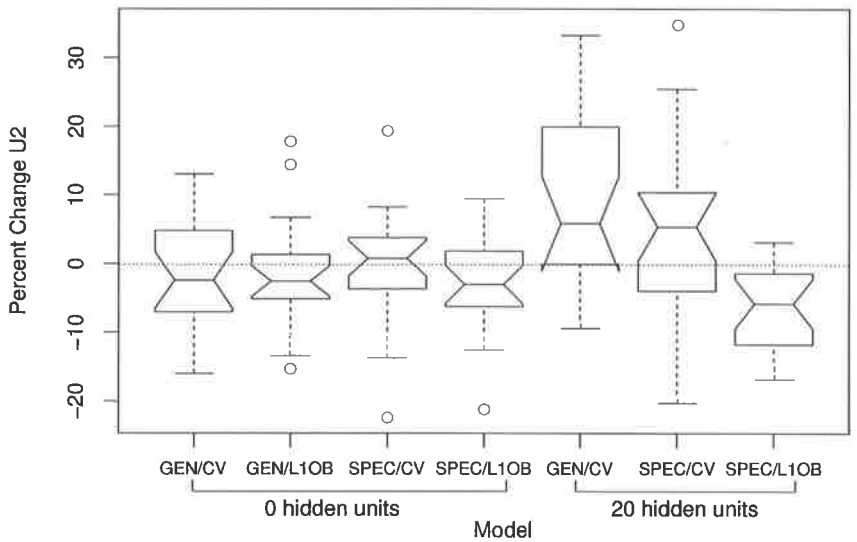

Figure 6.19: U2 – comparing input layer, hidden layer, validation mode.

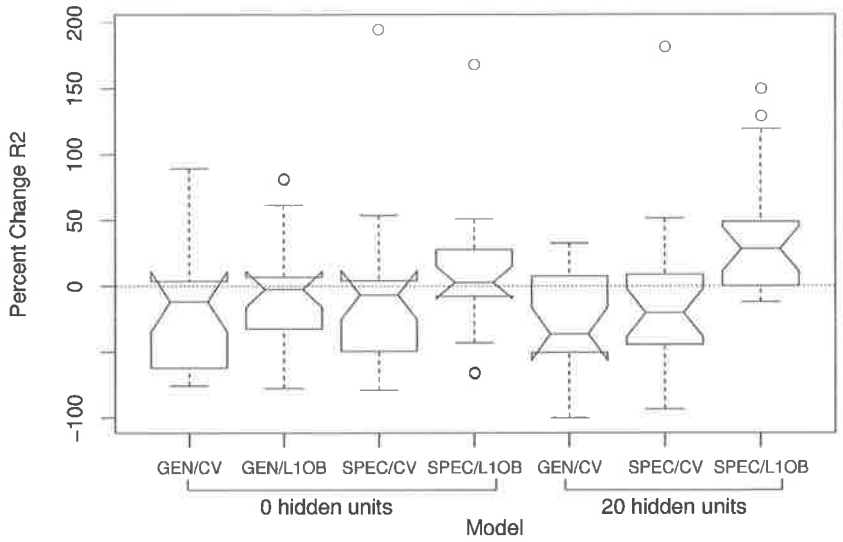

Figure 6.20: R^2 – comparing input layer, hidden layer, validation mode.

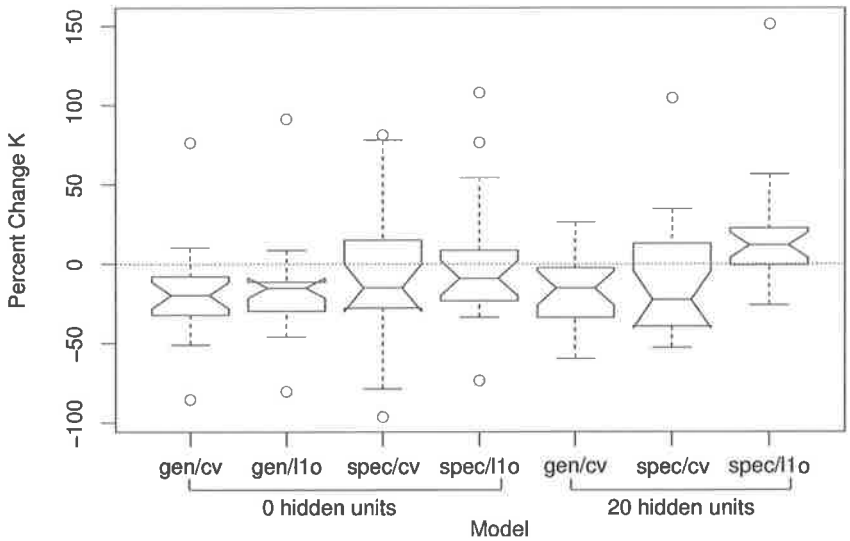

Figure 6.21: Av. κ – comparing input layer, hidden layer, validation mode.

input layer. Clearly the L1OB validation models return lower RMSE than the CV models and the specific models tend to perform better than the generic models.

This general pattern of experience, where the perceptrons appear relatively unaffected by either validation method or input layer type, is also reflected by the remaining four error measures portrayed in figures 6.18 to 6.21. However, there are some noteworthy features of these remaining plots. The U1 plot (figure 6.18) shows that all models, excepting the 20 unit specific L1OB model, have higher error rates than the control. Similarly, the results for average κ in figure 6.21 show that, relative to the control, the 20 unit specific L1OB model is the only model that has better classification than the control.

In summary;

- The use of a perceptron makes model performance insensitive to the effects of both input layer and validation method.
- Where a hidden layer is present, the following effects can be observed;
  - CV reports higher error rates than L1OB.
  - The specific input layer leads to better model performance than the generic input layer.
- The ANN only performs consistently better than the perceptron in the context of the specific input layer and the leave-one-out bootstrap validation method.

6.7 Discussion

6.7.1 Data Strip Mining

The outcomes of the data strip-mining experiments partially matched expectations. It was hypothesised that the final-pass model should have better performance than both the generic model and the starting model because;

- It may choose from a broader range of variables than is available for the generic model increasing the chances of discovering general relationships between input and output variables.
- The "stripping" of redundant inputs reduces the search space of possible solutions increasing the probability that a general model will be discovered.

However, it was found that, while a single pass of the input stripping procedure reduced error somewhat, continued iteration until all redundant inputs were removed increased error rates over both the starting model and the generic (5 input) control model for most outputs (the exception being *Pediastrum biwae* in Lake

Biwa). Clearly, continued stripping was causing relevant variables to be removed and/or redundant variables to be retained. Possible explanations for this behaviour include;

- Inference of the starting model is impacted by redundant input variables leading to inappropriate ranking of input relevance by the sensitivity analysis.
- The sensitivity analysis procedure itself does not accurately rank relevance of input variables.

If the former scenario is true, it suggests that a "catch 22" situation exists with respect to model selection in the context of a large number of potential inputs – the presence of too many inputs reduces the ability of sensitivity analyses to discern which variables are redundant and need to be dropped from the model. One way to alleviate this situation is to drop inputs deemed to be most redundant according to domain experts so as to improve the chance that the importance of remaining inputs may be correctly ranked. However, this adds to the user effort required to create ANN models. If the latter case is true, further investigation is needed to identify accurate techniques for gauging input relevance. It should be noted, however, that the sensitivity analysis procedure carried out in the context of the present study is very rigorous in comparison to procedures documented in the literature.

The results of this study are not entirely consistent with experimental results presented by Embrechts et al. (2001), where significantly improved performance resulted from "convergence" of data strip-mining on a minimally sized model. This is in spite of the fact that the methodology of the present study, with respect to model aggregation, sensitivity analysis and the use of early stopping to stabilise ANN learning, is similar to that prior study. However, there are a number of of differences between the two studies that may be the cause for the discrepancies in outcomes;

- ANN models of Embrechts et al. (2001) are developed for a bio-informatic application of in-silico drug design. The data used are molecular structure descriptors. This data is likely to be less noisy and more precise that that used for the present CEO-informatics ANN application making it easier to discover truly relevant relationships between variables.
- The present study compares the effects of data strip-mining on ANN modelling for 21 outputs from 6 datasets. By contrast, Embrechts et al. (2001) demonstrates the procedure on a single model. Thus it can be argued that the present study provides a somewhat larger sample size from which to judge the success of the technique.
- The models developed by the present study are complicated by consideration of time-series interactions as opposed to the steady state models of Embrechts et al. (2001).

- The criteria for feature reduction used by Embrechts et al. (2001) – that is, where sensitivity of an input falls below that of a dummy input, could be seen to be more aggressive in nature than that used in the present study, where an input was only removed when it's sensitivity fell below that of 2 dummy inputs.

6.7.2 Forward Selection

It can be concluded that the forward selection method consistently identified the best performing models, since the results show that the "combo" models consistently achieve lower validation error rates than the generic and single-pass strip mining models. Two possible reasons for this outcome include;

- Forward selection, as it was practised in the present study, is supervised by observation of the effect of adding each input grouping on generalisation performance. Data strip-mining is not supervised in this way.
- Forward selection starts with a small starting model known to be valid and grows it by adding inputs. Data strip-mining on the other hand starts with a large model whose inference is possible impaired by redundant inputs.

Given these observations, an interesting study would be to perform an unsupervised form of forward selection by retaining added inputs on the basis of sensitivity relative to dummy inputs instead of the effect on generalisation error. This may yield a superior non-supervised model selection approach that is not hampered from the beginning by a swathe of redundant input variables and, unlike the present approach, does not need access to target variables in independent datasets for the purposes of supervision.

In terms of performance, the specific "combo" model identified by forward selection achieved impressive performance gains over the generic model predictions described in chapter 4. There were improvements in all error rates and generally better subjective performance[2]. Of particular note is the fact that U2 errors were reduced to < 1 for all models, including those from Darling and Myponga, meaning that at all times the ANN was capable of better predictions than the no-change model.

6.7.3 Insights Regarding Time Series ANN Modelling

6.7.3.1 Modelling Time Series Interactions

The results for the forward selection experiments clearly showed that inclusion of the long ($t-7$ to $t-67$ day) input windows for the 5 generic inputs leads

[2] As shown in section 6.6.2, this may also be due to differences in the validation method.

to a general improvement in model performance over the generic control. This result is not unexpected, since, as Cortez et al. (2002) states, time-series models assume that prior patterns will be repeated in the future. Increasing the length of the input window evidently increases the likelihood that repeating patterns will be identified and learned by the ANN. Furthermore, it suggests that further experiments with input window lengths and lag periods may lead to even better results. This possibility was remarked on by Maier and Dandy (2000), who pointed out that selection of the correct lag times has an important influence the performance of time-series ANN models.

As explained in section 2.5.1, the way in which lags are expressed to the ANN in the present study is novel. All previous ANN applications have defined lags as discrete observations at a time in the past compared to the output variable, whereas the present study defines lags as the summary of a sliding window in time. While the input window technique was specifically intended as a means of casting an irregularly sampled time-series into a forecasting ANN structure, it has the benefit that it provides scope for experimentation not for just lag length and window length, but the window summary method as well. In the present study, the ANNs were trained on the average of the input variable records falling within the input window. However, it may be beneficial to use some other summary statistic such as the median, variance, trend, or even the outcome of some process equation to drive the model. Similarly, this approach to data representation could be extended to the output layer as well so that the model can be trained to predict trends, variance or some other statistic describing the distribution of the dependent variable within the output window.

6.7.3.2 The "Curse of Dimensionality"

The experimental results show that the "curse of dimensionality" does indeed impact the ability of the ANN to identify general relationships given data. In general, it was found that both strip-mining and forward selection approaches identified models that performed better than the complete model (ie the starting model for the strip-mining experiments). However, it should be conceded that the effect was relatively benign since the performance improvement wrought by model selection is generally small (approximately 5-10% lower error measures on average). Also, the fact that the generic and the full input layers had similar performance on average and the extreme nature of many of the models in terms of the ratio of the number of inputs to the number of training records, underscores the robustness of the ANN approach with respect to input dimensionality.

A number of ANN applications in the literature have presented results showing the effect of removing redundant input variables from ANN models. It is generally noted that increasing the size of the input layer makes the ANN easier to train (eg, Levine et al. (1996)). This is expected, since with more inputs, there are more solutions present in the search space that satisfy the training set. Also, as expected,

many authors find that ANN generalisation is improved by dropping irrelevant inputs. For example, Lee et al. (2003) performed backwards elimination of available variables to select inputs for an ANN forecasting chlorophyll *a* concentration one week in advance in a Hong Kong bay. Despite having 10 input variables describing a variety of environmental conditions deemed by domain experts to influence algal growth, it was found that the best model only considered lag chlorophyll *a*. Similarly, experiments reported by Maier et al. (1998); Aoki et al. (1999); Ball et al. (1998); Schaap and Bouten (1996) and Schleiter et al. (1999) found that taking steps to ensure the most parsimonious input layers improved generalisation performance of ANNs trained to model a range of environmental and ecological variables. However, what is noteworthy about most of these results is that the authors commented on the relatively minor nature of the improvement. Indeed, Walley and Fontama (1998), in training ANNs predict biodiversity in unpolluted river sites in Britain, found that stepwise removal of redundant input variables did not improve performance over maximally sized input sets.

Clearly, while the "curse of dimensionality" does have real implications for the task of identifying input variables for ANN models, it can be concluded that the effect appears to be relatively modest. This means that Scardi (2001)'s proposition that the robustness of ANNs with respect to redundant inputs provides scope for experimentation with input layer design in the hope that better performance can be achieved appears to be a valid one. The present results back this assertion, since it was found that in some cases, the best model discovered was the full model despite the undoubted presence of redundant inputs (eg, the model predicting diatoms in Lake Burrinjuck). The dataset and output specific nature of the results suggests that judgements about the value of elaborate input selection methods need to be made on a case by case basis, since there did not appear to be any clear patterns emerging linking effects of input reduction with factors such as data availability, sample density, or output variable.

6.7.3.3 Effect of Validation Method

The results show that the L1OB estimator tends to be a more optimistic observer of ANN model performance than CV. One possible reason for the differences is that L1OB is more data efficient than CV. As explained in section 2.5.3, CV entails the holding out of discrete blocks of samples in time from the pool of data available for bootstrap selection of training sets. As a result, training sets using L1OB will be on average 5-10% larger than when using the CV estimator leading to somewhat more accurate model inference.

Another possible reason for the differences is that use of the L1OB estimator brings better representation of "clusters" of data across training and validation sets. This is because L1OB causes validation records to be spread approximately evenly throughout the time-series on each bootstrap training run. As a result, it is likely that validation and training set records will be neighbours in terms

of their temporal position within the sample. Thus trends or non-stationarities in the dataset will be effectively represented in both training and validation sets decreasing the likelihood that the ANN will have to extrapolate when making predictions on validation data, since it will be more likely to be "in range" of the training set. This outcome is not unexpected, as Maier and Dandy (2000) pointed to i) the presence of non-stationarities in data as being a major impediment to the development of ANN time-series models and ii) the need for adequate representation of "classes" of data in both training and validation sets.

It could be argued that use of L1OB leads to unrealistic validation in the context of a time-series model because, when in training mode, the ANN may be exposed to training records in the near future relative to validation records. This is akin to being able to travel forwards in time 2 months to sample data for the purposes of training a model predicting ahead 1 month. So, a question arising is, how important is this effect in distorting estimates of modelling outcomes?

It can be argued that the answer to this question depends on the intended use of the model. Clearly, if an ANN model is applied in a real time forecasting application, best performance will be achieved if it is regularly retrained with the most up to date data to reduce the probability that trends cause the model to be "out of range" when making forecasts. In this context, the L1OB estimation method may be more realistic than a conventional cross-validation approach, since it allows better representation of trends in the data during training than where data is held out in large, contiguous blocks.

To date, all published ANN applications to time-series modelling of algal abundance hold out blocks of data 1-2 years in length for validation. The results of the present study would suggest this leads to overly pessimistic estimations of model performance. The present study highlights the need for further empirical assessment of the realism of different validation techniques in the context of real time forecasting applications.

6.7.3.4 The Effect of Hidden Layer Configuration

The results showing the interaction between the effects of hidden layer configuration, validation method and the input layer design are unexpected, because, under most conditions, the ANNs do not perform better in validation mode than the perceptrons. Indeed, the only conditions under which ANNs are clearly superior is when validated by L1OB and when configured with the larger, specific, input layers. What is most surprising is that ANNs never perform better when validated by the CV approach. Since all applications to time-series modelling of phytoplankton referred to in the literature use a simple form of cross-validation, as opposed to L1OB, the results suggest that these authors may have achieved better fits using a model inference approach constrained to linear decision boundaries! Note that these conclusions do not apply to performance on training sets, where,

without exception, the ANNs were found to be far superior to perceptrons (see chapter 4 for comparisons). Thus, the results show that ANNs are limited by their *generalisation* rather than *approximation* capabilities.

The fact that ANNs perform better than perceptrons with larger input layers is expected, since it is known that performance of linear regression estimation degrades as the dimensionality of the data increases. The reason for the interaction between the validation method and the hidden layer configuration is not so clear. A possible explanation is that ANNs may be generalising learned relationships over local rather than global time scales. Thus, records in the validation set close in time to records in the training set are better predicted by the ANN than by the perceptron. Since, as discussed previously, L1OB allows records in the validation set to be close in time to records in the training set, it takes advantage of this "local" generalisation. CV, by comparison, results in greater time separation between training and validation set records meaning that the ANN must rely instead on "global" generalisation. If this is the case, it would be reasonable to suggest that the time-series datasets used are characterised by trends or non-stationarities that make global generalisation by ANNs difficult.

While the superiority of ANNs over linear methods for modelling ecological variables has been widely demonstrated, several studies have hinted at limitations of the ANN approach under certain conditions. Hwarng and Ang (2001) showed that single layer perceptrons nearly always performed better at modelling a synthetic time-series than multi layer perceptrons (MLPs). These authors used conventional cross-validation rather than L1OB to arrive at this conclusion. Geman et al. (1992) presents an in-depth tutorial of the generalisation properties of *tabula rasa* (ie *model-free*) inference methods such as ANNs, GAs, CART etc. These authors concluded that;

> Inferring ... complexity from examples, that is learning it, although theoretically achievable, is, for all practical matters, not feasible: too many examples would be needed. Important properties must be built-in or "hard-wired," perhaps to be tuned later by experience, but not learned in any statistically meaningful way.

Further more, they argued that examples where ANNs had been successful at learning from data, or more specifically, where they had been able to achieve an inference task not possible with a conventional constrained approach, tended to be characterised by either unlimited data, or were essentially tasks of *interpolation* rather than *extrapolation*. It may be reasoned that the present results and those of Hwarng and Ang (2001), can be explained by the hypothesis that perceptrons are "hard-wired" to generalisation on a global time scale by being constrained to linear decision boundaries. Conversely, ANNs perform well on local time scales because this allows them to be interpolating rather than extrapolating in a non-trivial way. Clearly, more research is needed to determine the validity of these hypotheses.

6.7.3.5 The Effect of Data Availability

The results clearly show that ANNs trained on datasets characterised by high availability of training records (ie short sample period and/or large number of records) generally perform significantly better than ANNs trained for datasets characterised by low availability (ie long sample period and/or low number of records). The time-series plots show very high correlation between predicted and observed values for most Darling River and Myponga Reservoir models which, with generally short sample periods (ie ≈ 7 days for the output variable) and a large number of training records (ie 200 – 400), classify as relatively "data rich" datasets.

A question raised by these observations is, which aspect of data availability is more important in determining model performance – sampling period or number of records? The performance of the model predicting chlorophyll *a* in Lake Soyang provides a clue to this question, because the sampling frequency for the output variable and many input variables abruptly changes from monthly to weekly half way through the time-series. The plot of prediction and observations in figure M.6 shows that the model performs relatively poorly in the sparsely sampled region prior to 1993, but dramatically improves as the sampling regime is intensified. Since, obviously enough, the same quantity of training data is available to the ANN when making predictions throughout the time-series, this evidence suggests that it is sampling density that is more important to modelling outcomes.

Possibly, reduction of the sample interval strengthens time-series interactions between neighbouring samples. Thus the value of a variable at sample time t is more dependent on the value at sample $t \pm 1$ as the time unit t is shortened. Also, as suggested previously, it is possible that time-series of limnological and water quality observations are generally characterised by single or multi-dimensional non-stationarities, since shorter sample intervals and more data means that the ANN is more likely to be able to generalise relationships learned over short time periods.

6.7.4 Reservations about Models and Methods

6.7.4.1 Data Strip Mining

The results showed that, in general, at least one of the output specific models, that is, either the starting, single pass or final pass model, was likely to be better performing than the generic model. A criticism of the present study may be that even better performance could have been achieved if the best model was selected from the *entire* sequence of models generated during the strip-mining process. This is no doubt the case, but the problem with such an approach is that the selection would have to be made on the basis of generalisation error,

meaning that the validation data is used in the modelling process and is no longer independent. This effectively turns what was an unsupervised model selection method into a supervised one. Instead, it was intended to identify a methodology that, if applied the same way for every output and dataset, would have a high chance of approaching optimal performance.

6.7.4.2 Forward Selection

As stated in section 6.2.2, the forward selection technique used was coarse in that input selection is performed on functional groups of inputs rather than individual inputs. It must therefore be conceded that a more comprehensive forward selection experiment may better modelling outcomes. Thus further work is required to determine if this is indeed the case.

6.7.4.3 Alternative Model Selection Methods

Olden and Jackson (2000) performed Monte Carlo simulations to compare the performance of eight model selection techniques at identifying the correct independent variables for a synthetic multiple regression task. They found that all approaches erroneously included or excluded predictor variables, but that the relative performance of each method depended on the sample size. Obviously, the present study does not provide an exhaustive search of selection methods. However, it does served to show that a supervised method does perform better than a more recently introduced unsupervised approach. Clearly, more research is needed to investigate further selection methods in the context of the types of inference approaches and datasets used in the present study. However, it should be noted that, as discussed previously, ANNs are generally more robust with respect to redundant variables than the regression approaches investigated by Olden and Jackson (2000) meaning that the differences between competing methods may be relatively minor.

6.7.4.4 Consideration of Spatial Information

The review of ANN applications to modelling phytoplankton in the literature in chapter 2 concluded that a feature of existing applications is a lack of consideration of spatial variability of conditions in lake ecosystems. This is despite clear evidence in the literature regarding the effect of factors such as thermal stratification and wind on both water quality monitoring and ecosystem processes. This issue was not addressed in the present study. However, it needs to be pointed out that the input window technique employed in the present study could easily be extended to summarise spatial as well as temporal dimensions as a means of simplifying input data representation. This would enable experimentation with a

range of summary statistics – for example it may be possible to consider variance and gradients over an area or volume in addition to average values.

6.8 Conclusions and Recommendations

With respect to the hypotheses posed in section 6.1, it can be concluded that both feature selection methods investigated identified input layers that performed better than the generic 5-variable and fully parameterised models. Given this outcome, it can be concluded that elimination of redundant variables and retention of relevant variables to achieve the most predictive and yet parsimonious input layer is beneficial to modelling outcomes in the context of the inference methods and data used in the present study. This finding is consistent with the experience of the literature on the subject. However, it needs to be further noted that the results also support the suggestion of Scardi (2001) that ANNs are robust with respect to the presence of redundant inputs, since reasonable performance was achieved for several models that had more inputs than training records!

The results have also raised a number of issues that deserve further attention;

- While the supervised forward selection method performed better than the unsupervised strip mining method in terms of resultant model performance, it is reasonable to hypothesise that an unsupervised forward selection approach may yield comparable results since "starting" model inference will not be hampered by an excessive number of redundant input variables.
- It is clear that the validation method interacts with perceived model performance in a non-trivial way, with methods that place a high percentage of validation records close in time to training set data leading to better results. Further research is needed to investigate the presence of trends and non-stationarities in data and the degree to which different validation methods cause cross-contamination of training and independent subsamples.
- It was found that perceptrons tended to indicate better prediction accuracy when validated by CV, but that ANNs appear better when validated by L1OB. It was hypothesised that CV tests "global" generalisation abilities, while L1OB places a greater emphasis on "local" generalisation. Further research is needed to identify the properties of model inference techniques affecting local and global generalisation respectively.
- The fact that ANNs only performed better than perceptrons under limited conditions puts a new perspective on the value of *tabula rasa* model inference in the context of the present application. This finding emphasises the need for constrained models to be considered as a control for all ANN modelling applications.

- Sampling period and/or dataset size emerges as the most important factor in determining the performance of ANN models. This prompts a clear recommendation to water management authorities that real time forecasting applications will benefit from frequent and consistent water quality monitoring.
- The results showed that the use of longer lags was universally beneficial. This suggests that longer term time-series interactions than previously supposed exist in the data. It can be concluded, as per Maier and Dandy (2000), that more work to resolve useful lags may bring further improvements to model performance.
- The input window method provides scope for alternative data representations that describe the properties of distributions of independent or dependent variables in space or time. More research is needed to determine how alternative data representations will affect modelling outcomes.

Chapter 7

Conclusion

This thesis has addressed the ongoing need for computer models that can aid decision making in the implementation of eutrophication control options. Working on the premise that, as far as possible, "machine learning" techniques should operate autonomously, the goal has been to identify generic ANN model representations and methodologies that reduce the need for user intervention on a case by case basis. The approach taken has been to identify the "bottlenecks" that require significant decisions to be made by users at each step of the model development process. It is concluded that the methodologies discussed in this thesis make a significant contribution towards a more generic ANN methodology for modelling phytoplankton dynamics in lakes.

The most profound innovation presented is the representation of model inputs as summary statistics of sliding time windows. The added flexibility of this approach makes reconciliation of time-series ANN model structures with available datasets easier on two counts. Firstly, it allows variables that differ in sampling frequency to be used in the same model. This means that a greater selection of variables may be available to be included in a model for a given dataset than previously. Secondly, it eliminates the need for interpolation of data to a constant sampling interval in order to "fit" the length of time delay connections in the ANN structure. This significantly reduces the total information processing task on the part of users, since data preprocessing is simplified and overall training times are reduced. No decisions need to be made by users regarding the mode of interpolation. Also, it reduces the risk that performance expectations will be biased by "data contamination" between past and future states[1].

It was demonstrated (chapter 4) that bagging significantly decreased ANN model sensitivity to overfitting. It was concluded that the bagged model is relatively unaffected by the increase in the variance component of model error during overfitting because uncorrelated predictions between individual member models of the

[1]Lee et al. (2003) demonstrated how interpolation can blur information between time periods leading to unrealistic model performance expectations.

bagging ensemble tend to cancel each other out. When used in combination with efficient model approximation techniques[2], bagging mostly eliminates the need for user intervention to optimise the approximation and generalisation characteristics of the model. This removes a major technical headache so far as users are concerned, since optimising arcane parameters such as learning rate, momentum, hidden layer configuration, training time, weight decay, jitter *et cetera* is a difficult and error prone task[3].

Rotation estimators, such as leave-k-out cross-validation, have been used for validating all ANN models. This approach makes all records in a given dataset available for validation without significantly degrading training set representation. No user decisions are required regarding division of data into training and validation sets. This eliminates a possible bias from the methodology, since practitioners are prevented from "coaching" performance outcomes by selecting validation data to emphasise or hide certain model characteristics. It can be proposed that rotation estimators, when used in combination with bagging, lead to repeatable modelling outcomes. This is because it cannot be claimed that performance has been influenced by random or intentional variations in either training and validation data.

This thesis has also highlighted steps in the model development process-model that are, as yet, difficult to apply generic methodologies to. It was demonstrated (chapter 6) that input layer selection is best done on a case by case basis because;

- Datasets from different lakes are unique in terms of the feature set available for modelling.
- Dataset/output specific models performed significantly better than models restricted to commonly available variables.
- Including all available variables as inputs does not result in the best modelling outcomes – redundant inputs degrade performance.

Two approaches to automated selection of case-specific input layers were applied and demonstrated to be moderately successful. It is clear that including a mix of short and longer term lags for crucial input variables improves performance. However, there was still considerable variation in outcomes between individual models. Thus it can be concluded that the generic ANN modelling approach presented in this thesis would benefit from future research on generalised input selection methods or heuristics.

[2]It was shown (chapter 4) that the Scaled Conjugate Gradient (SCG) (Møller, 1993) and incremental mode backpropagation (BP) are effective training algorithms. However, SCG observed to be significantly faster and does not require tuning of learning rate or momentum coefficients.

[3]Breiman (1996b) states that, even when properly regularised, unstable inference methods such as ANNs may still not perform optimally.

Models were developed to forecast 21 output variables in 6 datasets. Several observations were made in the course of this modelling work that may have profound implications. It was observed that MLPs did not always perform better at predicting validation data than perceptrons, despite being demonstrated to be consistently better at learning training sets. Furthermore, it was found (chapter 6) that there is a significant interaction between the ANN architecture and the validation method; MLPs consistently outperformed perceptrons when validated using the leave-one-out bootstrap, whereas perceptrons performed better when validated by blocked 20-fold cross-validation. These two validation methods differ from each other in one important respect – the leave-one-out bootstrap permits every record in the validation set to be bounded by training set records in the time-series. Thus it can be concluded that MLPs generalise well over short-term "local" time scales, but not over longer term "global" time scales. Clearly research is warranted to further characterise the effect of time on the generalisation characteristics of ANN models.

In terms of performance, it is demonstrated that feedforward ANNs can be credibly used to make one to two week forecasts of the abundance of chlorophyll *a* or phytoplankton species/functional groups using datasets typical in structure – that is, with irregular sampling intervals, different sample dates for input and output variables and long periods of missing data. It can be concluded that the best validation set performance is returned when the model is configured with output/dataset specific input layers, 20 hidden layer units and validated using the leave-one-out bootstrap. The time-series plots showed moderate to good agreement between the observed and predicted abundance. Furthermore, U2 error rates ≤ 1 showed this configuration performed better than the "no-change" model ($y_t = y_{t-1}$) and positive average κ values indicated better classification performance than a random classifier.

7.1 Summary of Findings and Recommendations

Selection of Model Inputs:

- The generic input layer, comprising oxidised nitrogen, orthophosphate, water temperature and secchi disk depth, is compatible with typical monitoring datasets and is a useful starting point for identification of ANN models.
- For optimum performance, output specific input layers need to be selected. Selection approaches based on forward selection appear to perform better than backwards elimination.
- Elimination of redundant variables identified by sensitivity analysis yields moderate performance gains.

- Input layers should consider a mixture of short and long term lags, although more research is needed to identify a selection of lag intervals for inputs likely to generalise for many outputs and datasets.

Representation of Time Series Interactions:

- Model generalisation was possible using the input-window approach, even when sampling intervals were variable or there were long periods of missing data.
- As a means of ensuring compatibility with existing datasets, the input-window model representation is arguably superior to interpolation. It strictly defines the "temporal definition" of the model meaning that issues relating to data contamination (ie the model having access to information it shouldn't) identified by Lee et al. (2003) can be minimised. It also significantly reduces the information processing task, both during model inference and data preprocessing.
- The input-window model representation provides scope for further research determining the effects of different approaches to summarising or preprocessing input (and/or output) window conditions.

Training Algorithms:

- SCG is a significantly more efficient approximator than BP because it requires no tuning of learning rate or momentum coefficients and it is faster to train large networks.
- There were no differences between BP and SCG in terms of generalising ability. While not an exhaustive test of all approaches to supervised training, the results suggest that algorithms should be selected on the basis of their approximation properties rather than generalisation capabilities.

Model Complexity:

- Distinct underfitting, optimum and overfitting phases were observed for all models with increasing model complexity.
- Bagging greatly reduced increased prediction error due to overfitting. However, contrary to results of Cannon and Whitfield (2002), there was still a slight increase in validation error in the overfitting phase meaning that tuning fitting power by cross-validation is still necessary for optimum model performance.
- Where tuning of model complexity is deemed necessary, training error was shown to be a far more convenient parameter than the number of hidden

layer units. Thus, specification of hidden layer size should only be guided by the approximation requirements of the dataset.

- The results show that indices derived from model variance may be useful as a goal functions for tuning ANN complexity.

Model Validation:

- Rotation estimators provide a more robust model performance evaluation in data limited applications than conventional split-plot cross-validation since they allow better validation set representation without reduction of training set representation.
- L1OB returns more optimistic model performance estimates than blocked CV. It was hypothesised that L1OB permits training and validation set records to be closer in time, thus increasing the probability of "local" generalisation over short time scales.
- Further research is needed to clarify the properties of different rotation estimators in terms of;
  - resisting temporal contamination of training sets,
  - accounting for the effects of local versus global generalisation in the context of time-series data and
  - data efficiency.

Error Measures:

- Visual assessment of time-series plots and classification statistics are a useful way to characterise different aspects of model performance such as resistance to false positive or negative predictions.
- Objective error measures are useful for developing approaches to model selection or algorithm tuning.
- Normalised error measures are useful for comparison of model performance between different outputs/datasets.
- Theil's inequality type 2 (U2), which indicates comparative RMSE performance of the ANN model and a naive no-change model, is a useful "reality check" where other performance measures indicate very high performance.
- Further consultation with domain experts, water resource managers and process engineers is needed to develop error measures that are useful for describing the meaning of the model performance to interested parties.

Knowledge Discovery:

- The correlation between the input perturbation and the output response can be used to quantify the overall complexity and polarity of the of the correlation between a given input variable and the output.
- Sensitivity analysis identifies redundant inputs that can be dropped from the model structure to improve performance.
- Further work is required to analyse the sensitivity analysis data gained in this work in order to characterise complex interactions between the effects of input variables on the output that may have been learned.
- Further research is needed to investigate the potential of using the sensitivity analysis through time feature as a means of deriving management decisions for water resource process engineers.

7.2 The Future

There are many ways in which the generic ANN methodology developed in the present study could be extended or improved. The review of existing ANN applications (chapter 2) showed that no models have taken account of the effect of spatial variability of conditions. The input-window approach could easily be extended, where data is available, to define summary windows in space as well as time. Also, the present study only considered the average of the input-window conditions. Other summary statistics such as the range, minimum, maximum, trend over time, variance etc may be useful. Similarly, outputs as well as inputs could be represented as the summary statistic of a time window.

While a "sensitivity analysis through time" procedure was described and implemented, the results presented (chapter 5) only described the relative importance and complexity of interactions between the inputs and the output variables. There is still a need to explore ways of expressing interactions between the effects of input variables. It is proposed that the timing of sensitive periods for a particular input variable may yield information that can be exploited when designing tactical responses to predicted algal blooms. Similarly, validation data could be mined to identify the ranges of certain variables when the model is likely to perform well and when the model is likely to perform poorly. This would provide a measure of confidence in model predictions.

The reviews of Recknagel (2003) and Jørgensen (1999) suggest that the future of research in the ecological informatics domain is likely to be focused on two issues;

- Increasing the transparency inductive modelling techniques by development of adaptive knowledge representations that make sense to ecologists (eg

Todorovski et al. (1998), Bobbin and Recknagel (2001), Maier et al. (2001), Recknagel et al. (2002)).

- Incorporation of structural dynamic properties in models to account for the adaptive nature of ecosystem processes and parameters through time (see proposals of Recknagel (2002)).

Thus the "cutting edge" of research into computational modelling of ecosystem variables appears to be shifting attention from ANNs to novel machine learning techniques that are able to learn model representations consistent with the above goals. It should be pointed out the methods implemented in the present study (ie use of input-windows, bootstrap aggregation and k-fold cross-validation) are appropriate for any type of regression estimation, regardless of the estimation method or model structure. Indeed it can be argued that bagging is particularly appropriate to novel computation modelling techniques that are unstable[4]. However, it must be commented that the approaches advocated by this thesis are computationally intensive because ensembles of models are required to make predictions and perform validation. This may particularly be the case for algorithms that are relatively slow to converge, such as those based on genetic algorithms.

[4]Geman et al. (1992) argues that any *model-free* approach to regression estimation is characterised by high variance.

Appendix A

Tactical Responses to Algal Blooms

Table A.1: Examples of tactical controls that may benefit from short term forecasts of algal abundance.

Response Type	Response	Reference
a–priori	Algicide application (eg $CuSO_4$)	Whitaker et al. (1978); Burch (1990)
	Intermittent destratification to prevent domination by *r*– or *k*– selected algal species	Reynolds et al. (1984); Burns (1994)
	Barley straw application to promote herbivorous zooplankton and produce phytogenic compounds	Everall and Lees (1996)
	Nutrient precipitation	Hall et al. (1995)
	Flushing with low nutrient, low biomass water	eg Senate Standing Committee on Environment Recreation and the Arts (Aust.) (1993)
a–posteriori	Opportunistic water usage. Bloom affected reservoirs taken "off–line".	Whitehead and Hornberger (1984)
	Alert levels and an appropriate response framework	Burch (1993)
	Cyanotoxin removal during water treatment using adsorption and oxidation techniques	Burch and Nicholson (2000)
	Deployment of booms around offtakes. Alteration of offtake depth.	

Appendix B

Box and Whisker Plots

Box and whisker plots (Tukey, 1977; McGill et al., 1978) are a convenient method for displaying and comparing distributions in graphical form. Examples of notched box–and–whisker plots are used in section 3.3 to compare water quality parameters. They convey the following information;

- *Range* is indicated by the whiskers extending from each side of the box.
- *Median prediction* is indicated by the central line in the box.
- *Upper and lower quartiles* are indicated by the upper and lower sides of the box. These are the median values of the upper and lower halves of the distribution respectively.
- *95% confidence intervals* around the median is indicated by the notch.
- *Outliers* are indicated by circles beyond the reaches of the whiskers.

The calculation of 95% confidence interval is shown in equation B.1 where IQR corresponds to the interquartile range. This approximation is only strictly valid if the distribution of data is normal. Velleman and Hoaglin (1981) gives further information regarding this confidence interval calculation including derivation of the 1.58 constant.

$$95\%\text{CI} = \text{median} \pm \frac{1.58 * \text{IQR}}{\sqrt{\text{no. obs.}}} \tag{B.1}$$

Boxplots may be used to provide basic model diagnostics when used to display distributions of predictions from a bagging ensemble (for example, see figure 4.7). The width of the boxplots is indicative of variance. Very wide boxplots will tend to indicate that overfitting is occurring or that the input data for that record is not well represented in training data. Also, the size of the notch and the position of the median line in the box provides information about the normality of the distribution of model predictions. If the notch is not central it indicates that the distribution is

probably skewed, and if the notch is very wide, it is likely that the distribution of predictions is strongly multi–modal indicating several distinct "classes" of model being generated.

Appendix C

Effect of Model Aggregation on RMSE

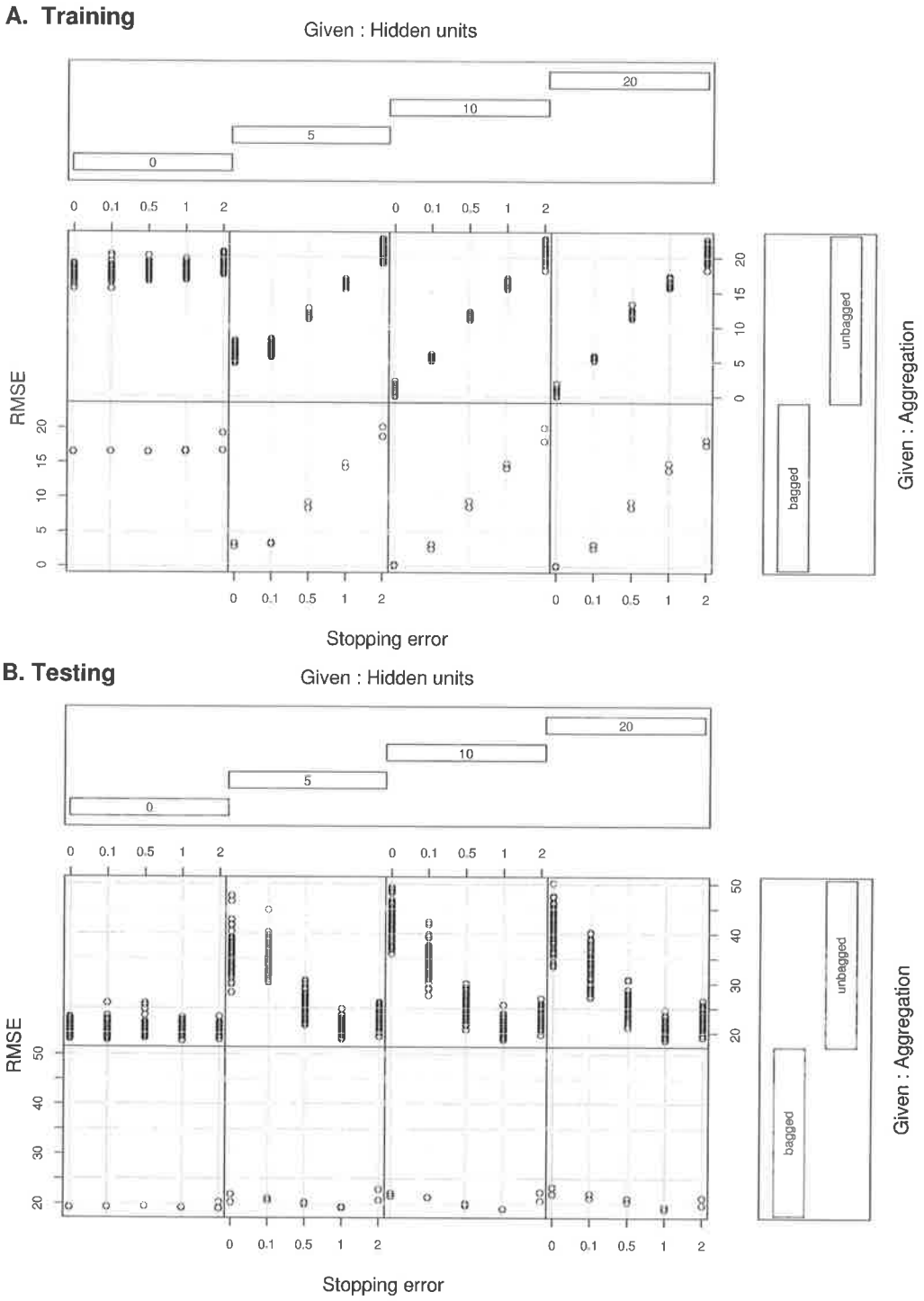

Figure C.1: Model No. 1. **A** Train RMSE v Stop error given No. Hidden units and Aggregation. **B** Test RMSE v Stop error given No. Hidden units and Aggregation.

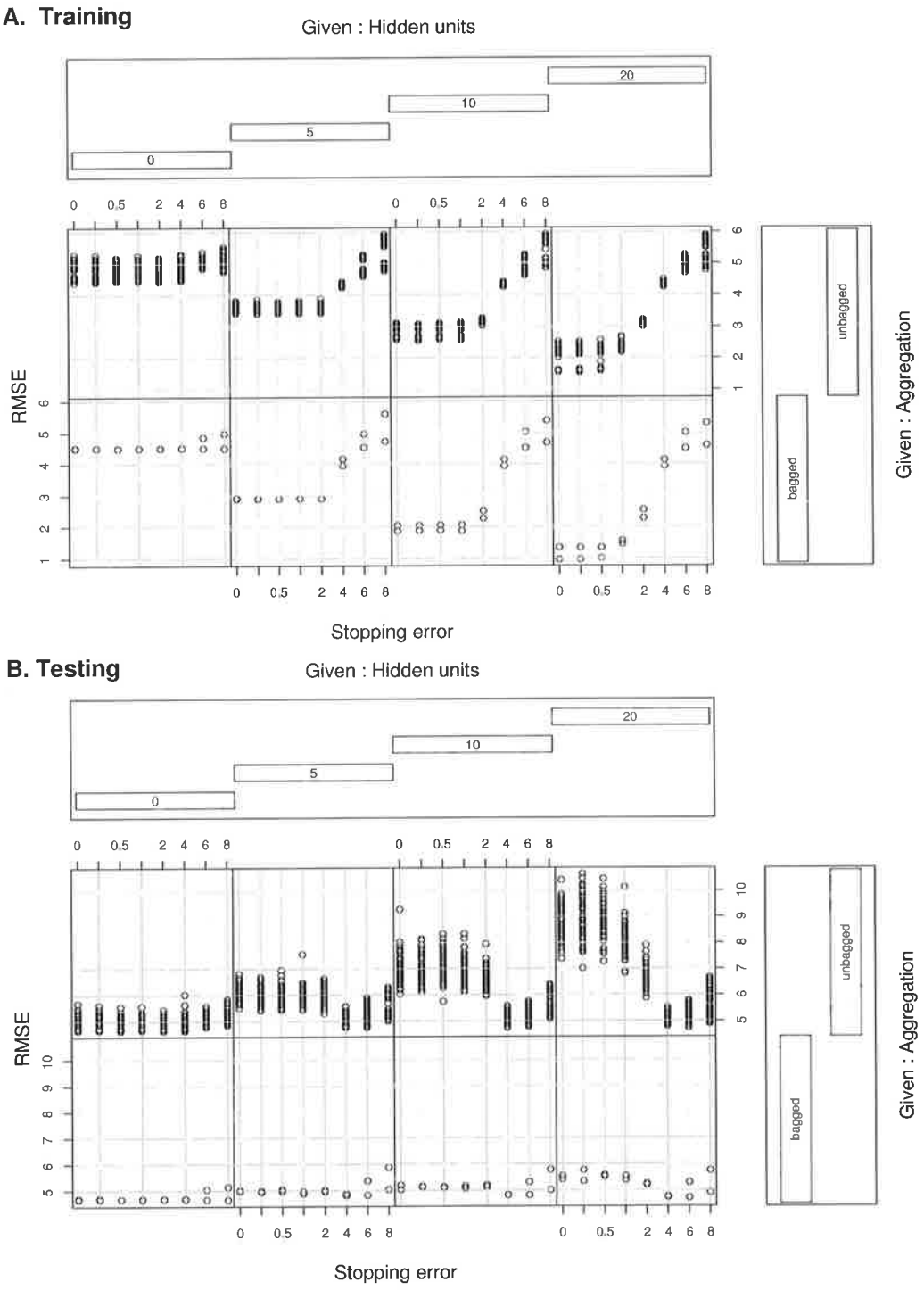

Figure C.2: Model No. 2. **A** Train RMSE v Stop error given No. Hidden units and Aggregation. **B** Test RMSE v Stop error given No. Hidden units and Aggregation.

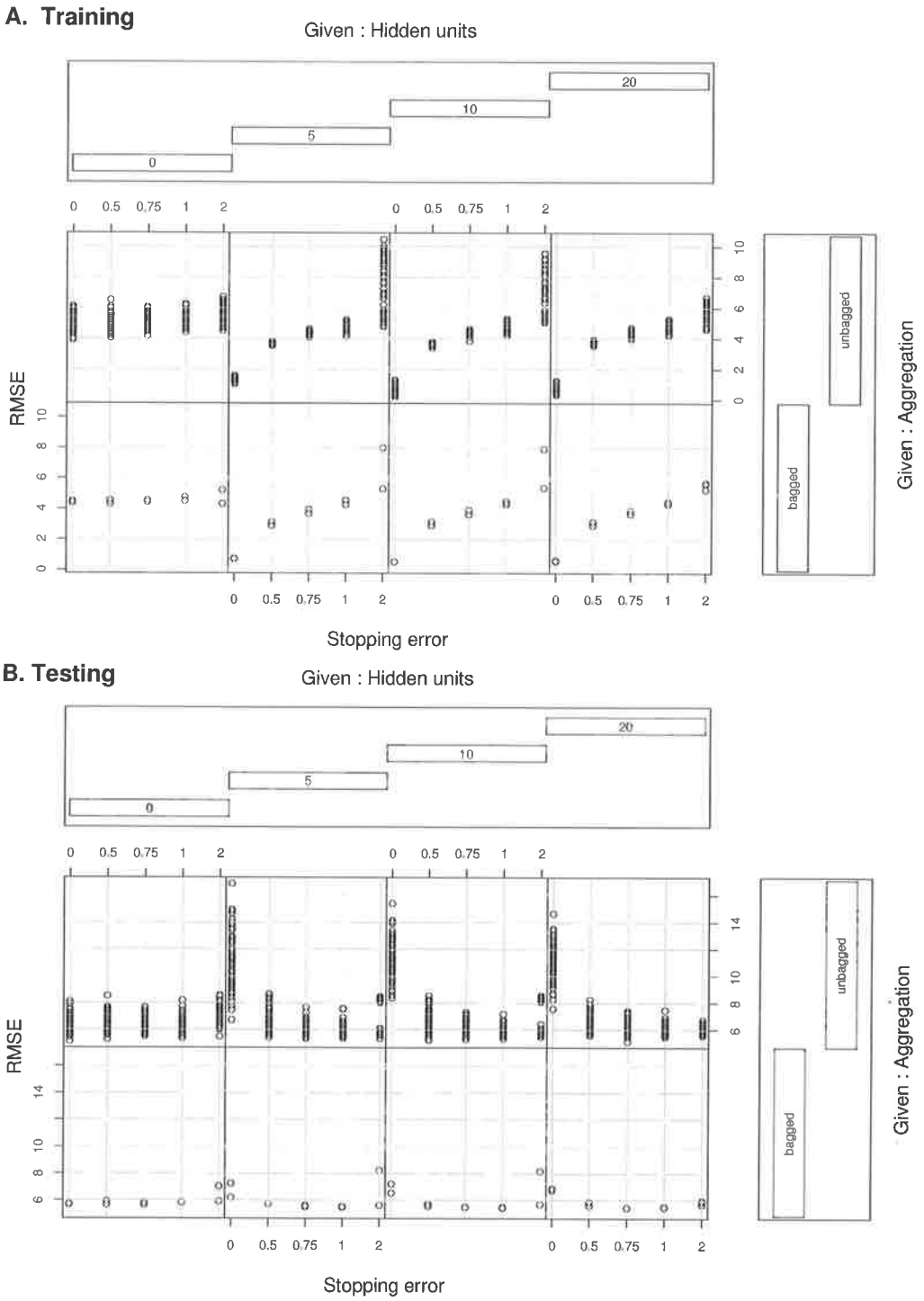

Figure C.3: Model No. 3. **A** Train RMSE v Stop error given No. Hidden units and Aggregation. **B** Test RMSE v Stop error given No. Hidden units and Aggregation.

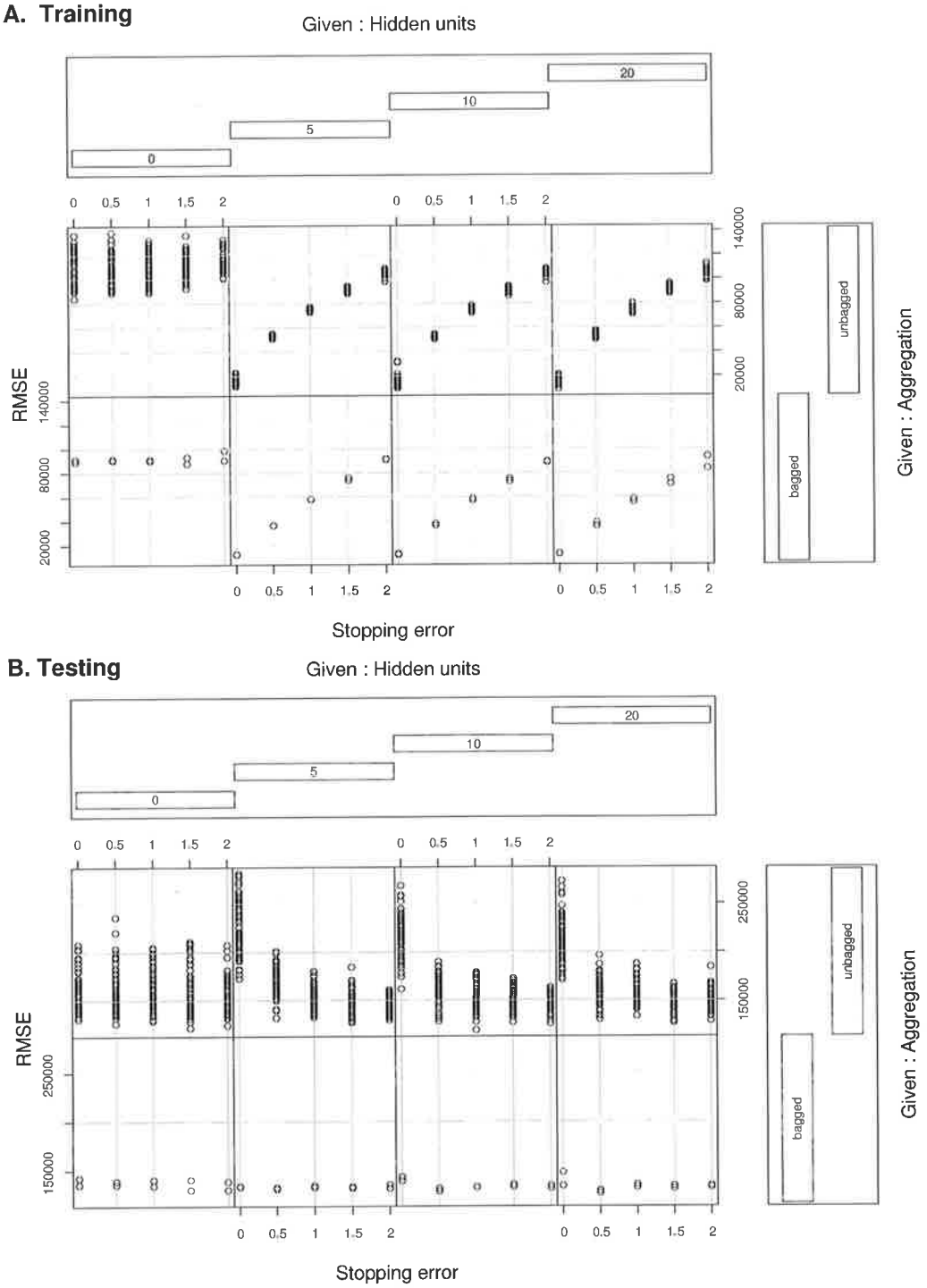

Figure C.4: Model No. 4. **A** Train RMSE v Stop error given No. Hidden units and Aggregation. **B** Test RMSE v Stop error given No. Hidden units and Aggregation.

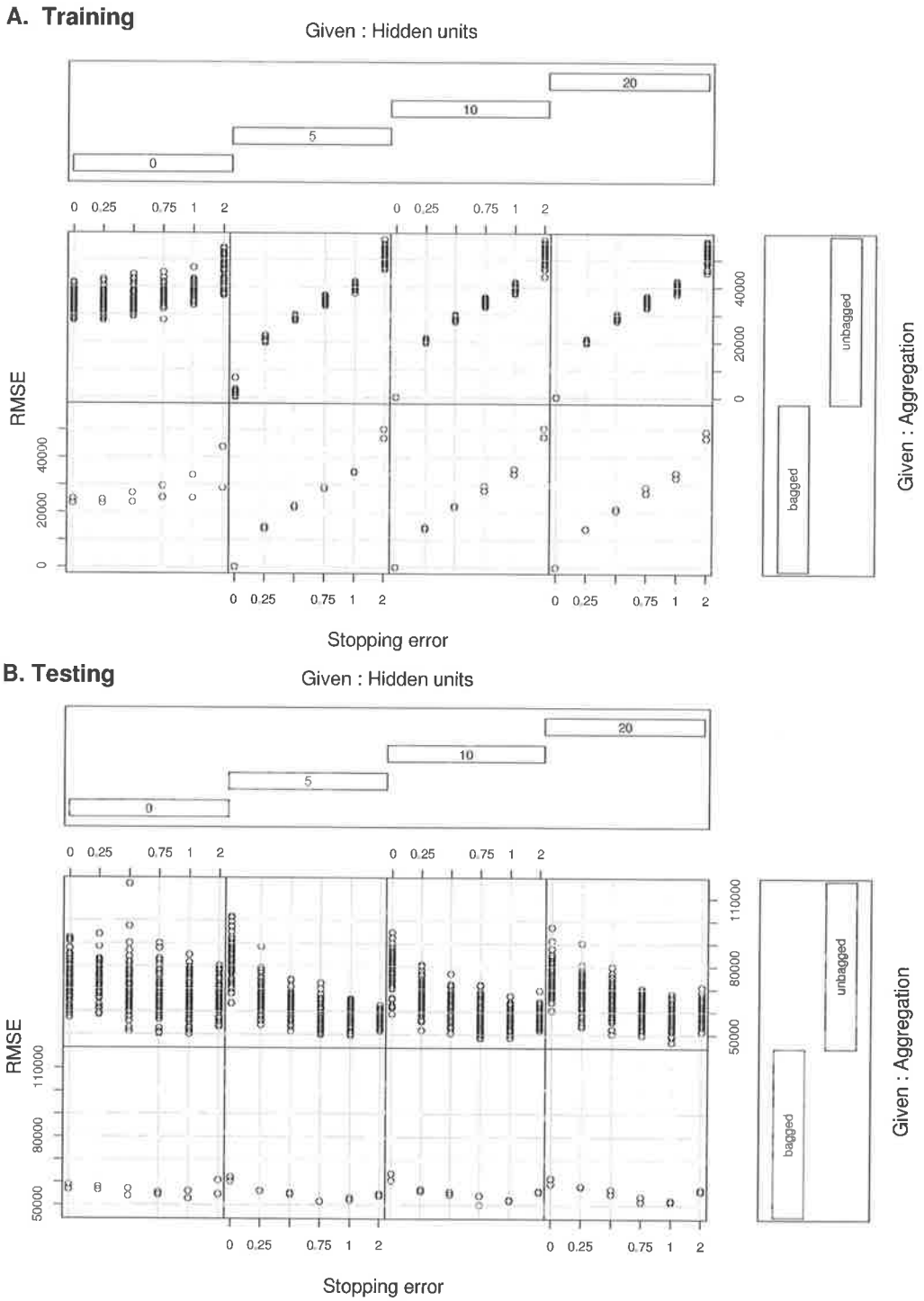

Figure C.5: Model No. 5. **A** Train RMSE v Stop error given No. Hidden units and Aggregation. **B** Test RMSE v Stop error given No. Hidden units and Aggregation.

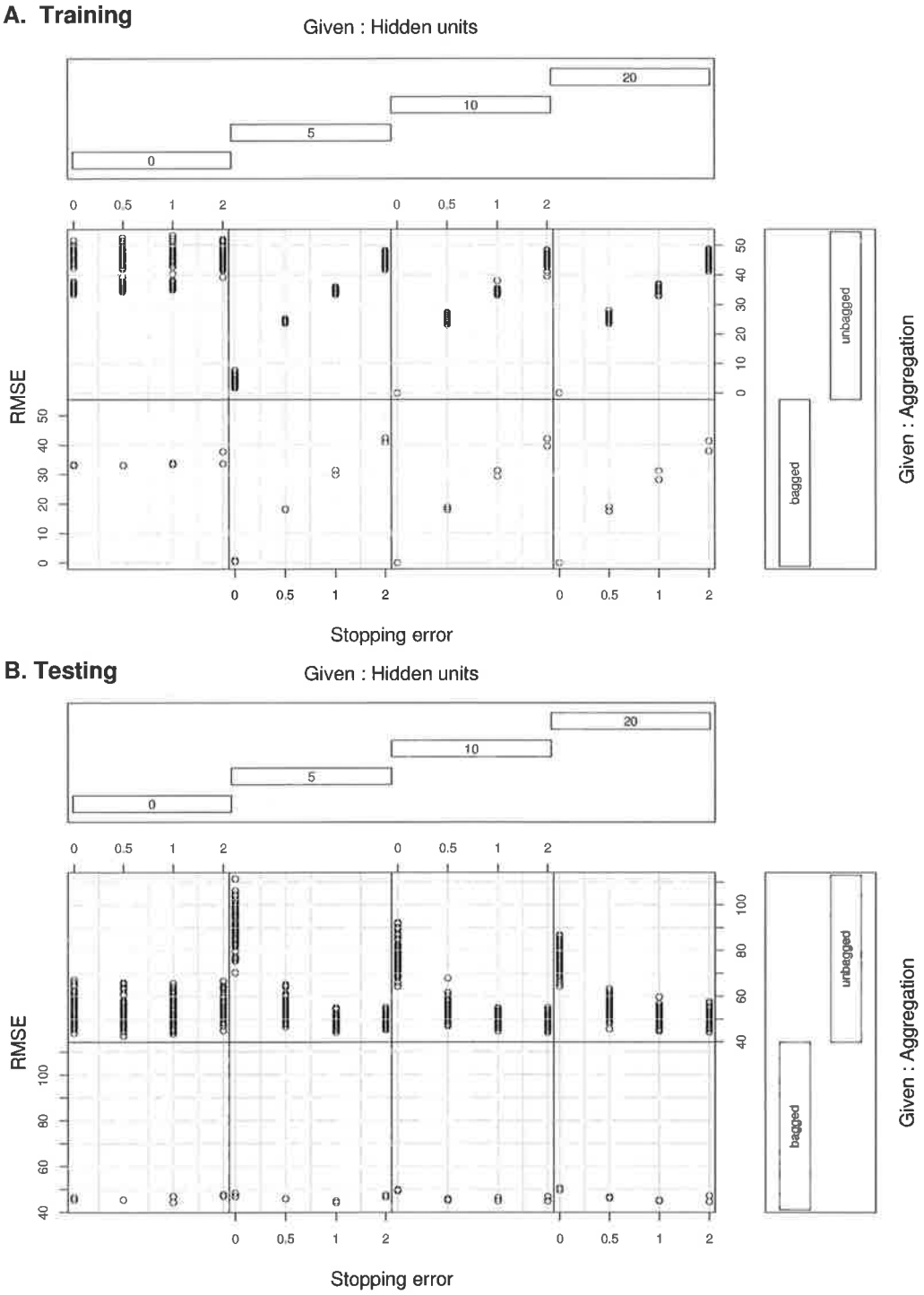

Figure C.6: Model No. 6. **A** Train RMSE v Stop error given No. Hidden units and Aggregation. **B** Test RMSE v Stop error given No. Hidden units and Aggregation.

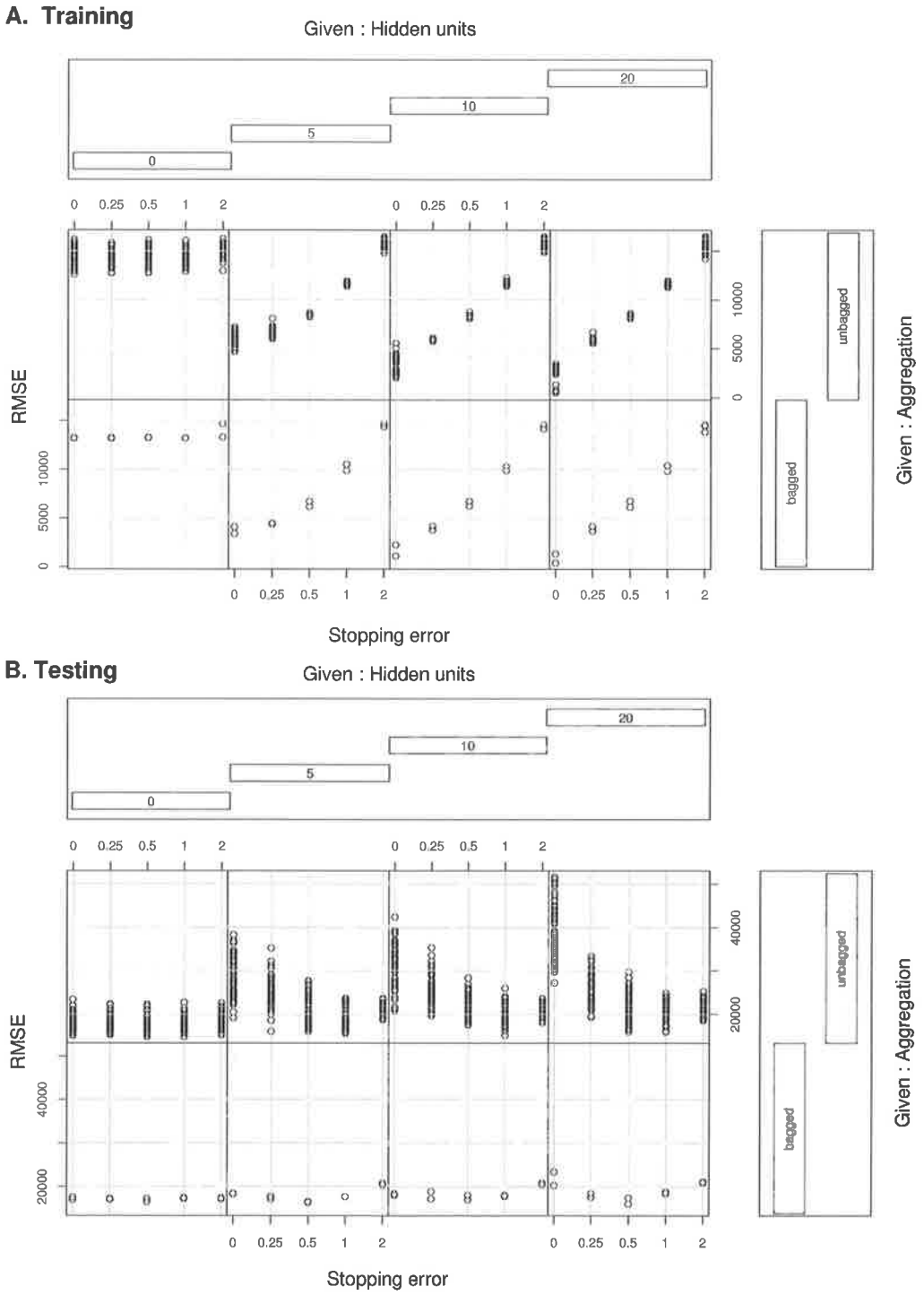

Figure C.7: Model No. 7. **A** Train RMSE v Stop error given No. Hidden units and Aggregation. **B** Test RMSE v Stop error given No. Hidden units and Aggregation.

Appendix D

Effect of Model Complexity on Model Predictions

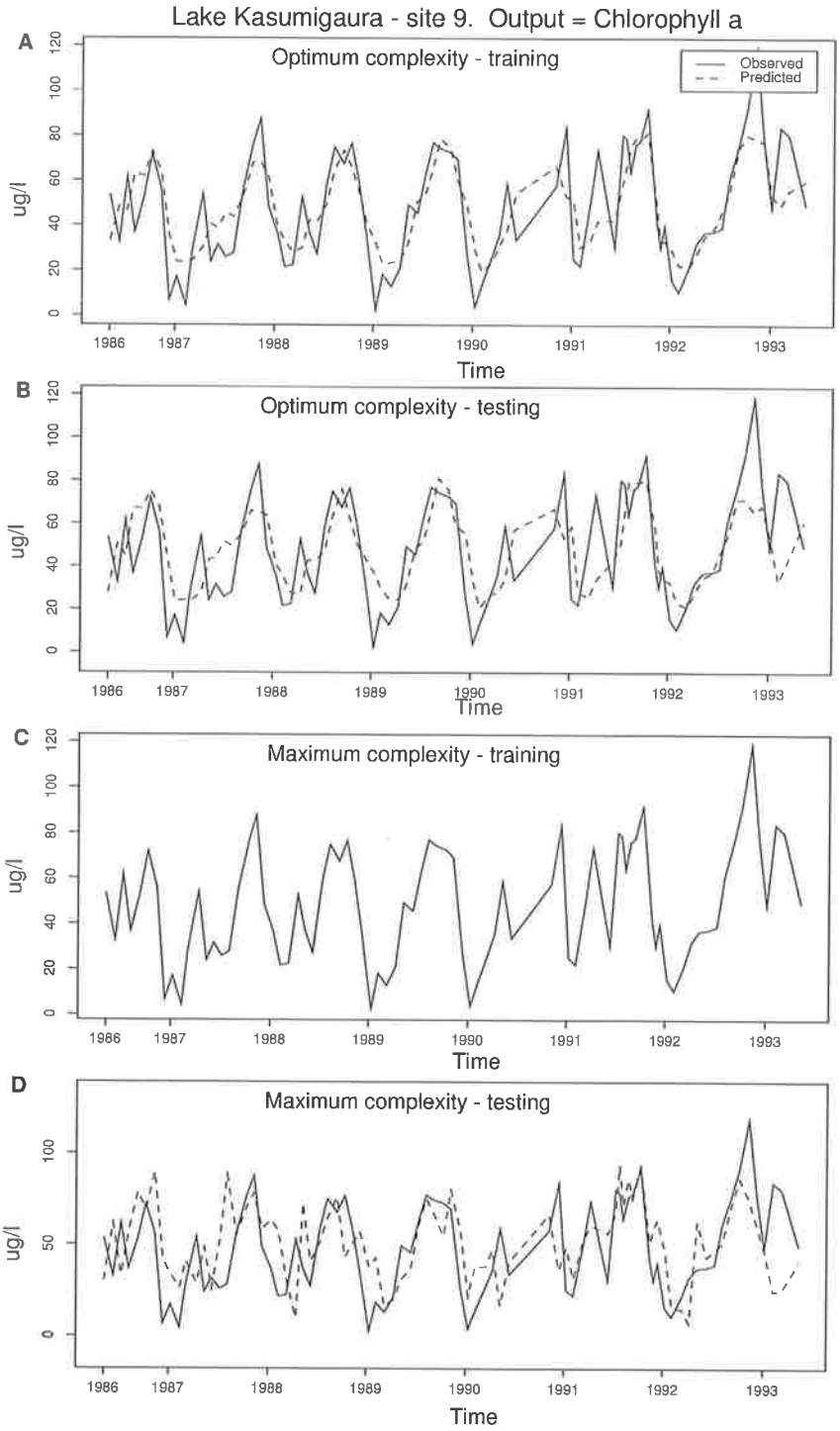

Figure D.1: Model no. 1 – Time series plots of observed and predicted algal abundance. **A** Optimum complexity – training. **B** Optimum complexity – validation. **C** Maximum complexity – training. **D** Maximum complexity – validation.

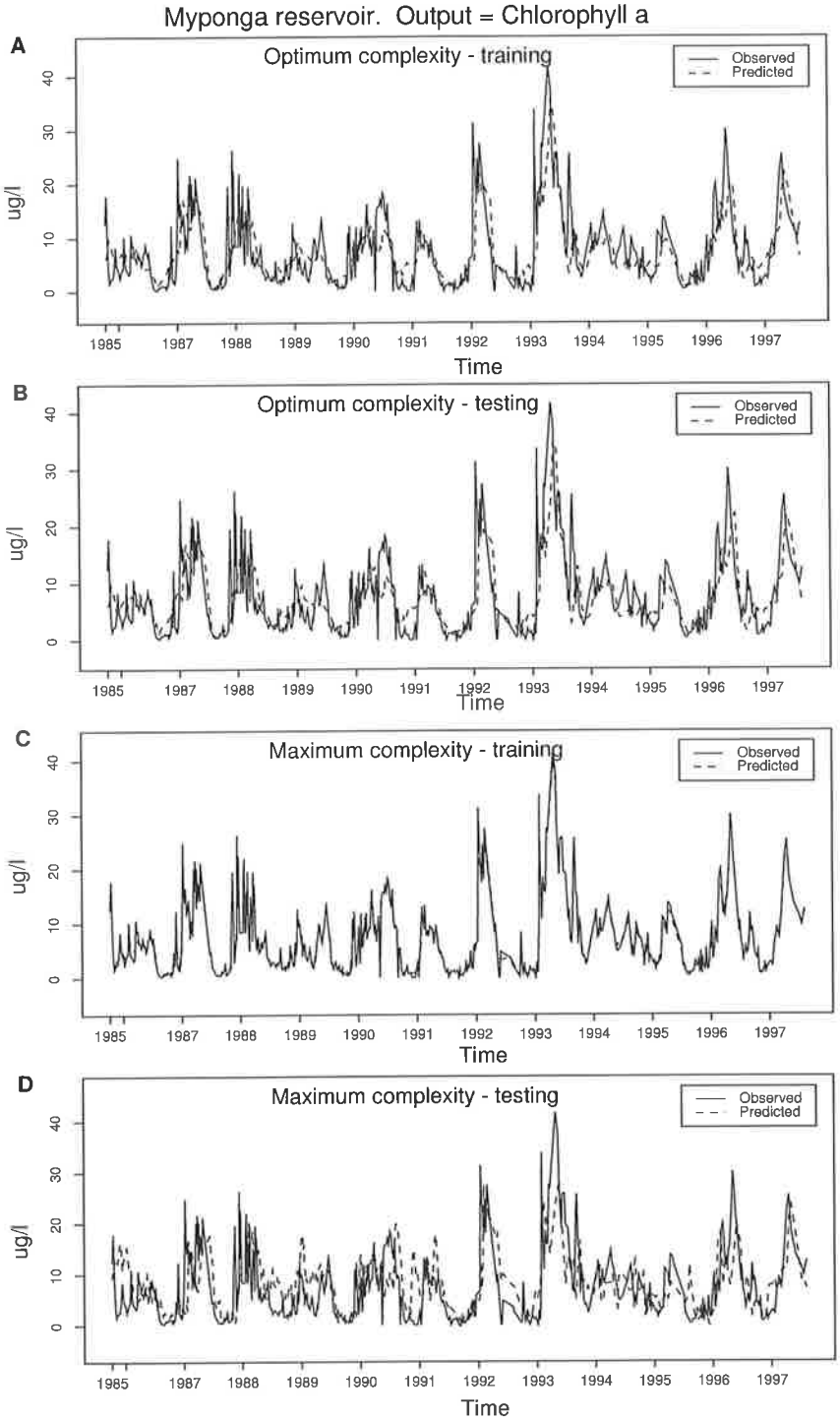

Figure D.2: Model no. 2 – Time series plots of observed and predicted algal abundance. **A** Optimum complexity – training. **B** Optimum complexity – validation. **C** Maximum complexity – training. **D** Maximum complexity – validation.

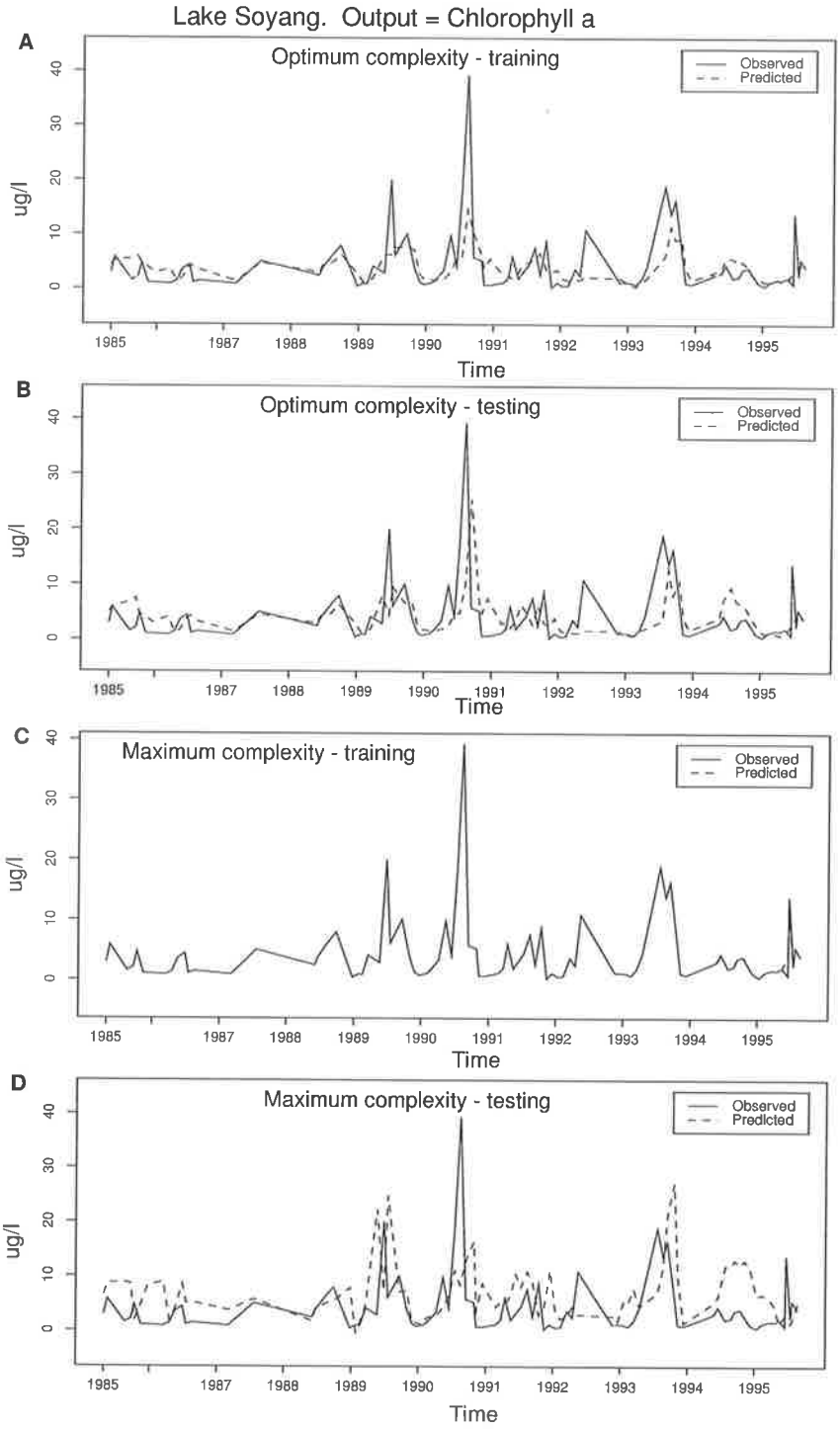

Figure D.3: Model no. 3 – Time series plots of observed and predicted algal abundance. **A** Optimum complexity – training. **B** Optimum complexity – validation. **C** Maximum complexity – training. **D** Maximum complexity – validation.

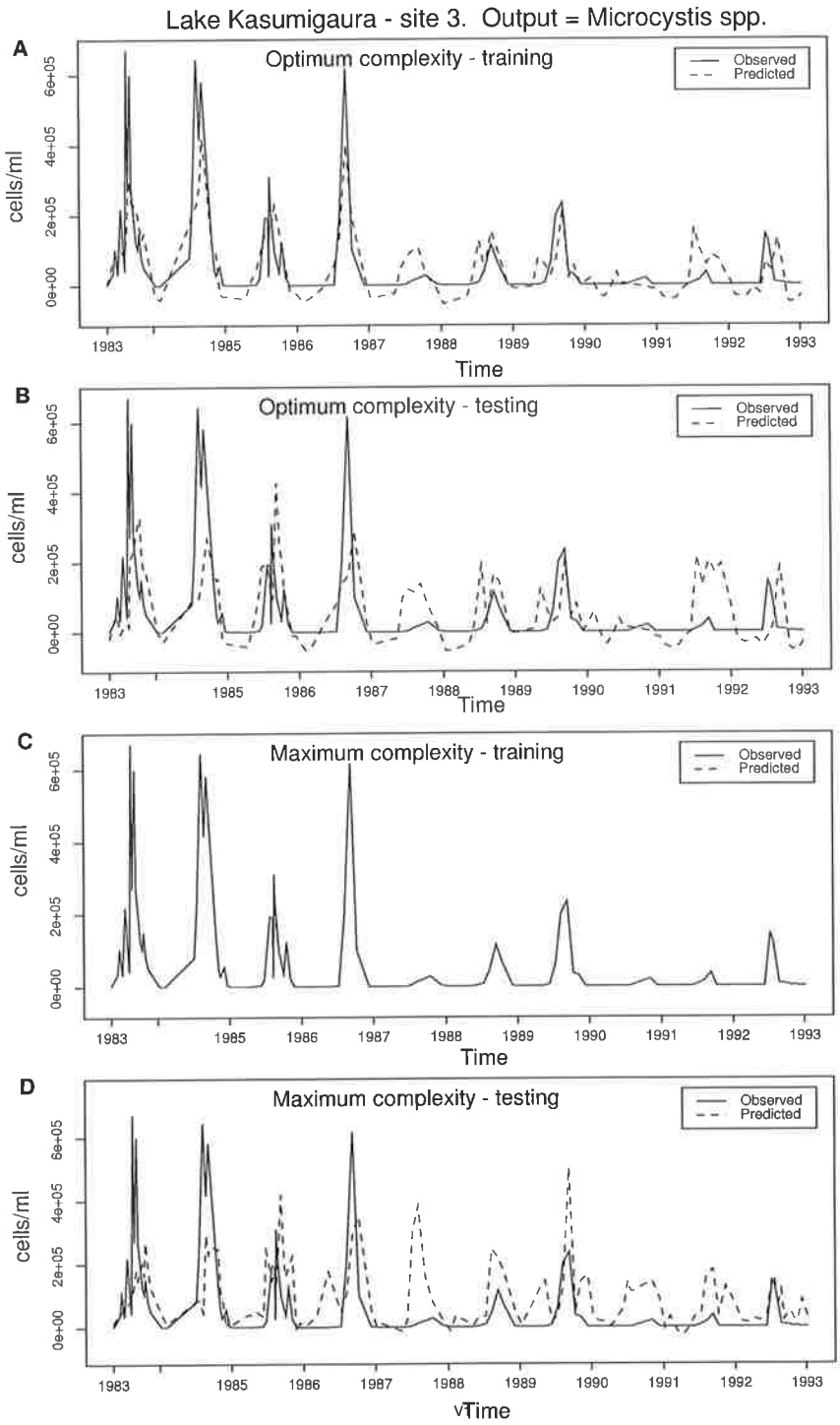

Figure D.4: Model no. 4 – Time series plots of observed and predicted algal abundance. **A** Optimum complexity – training. **B** Optimum complexity – validation. **C** Maximum complexity – training. **D** Maximum complexity – validation.

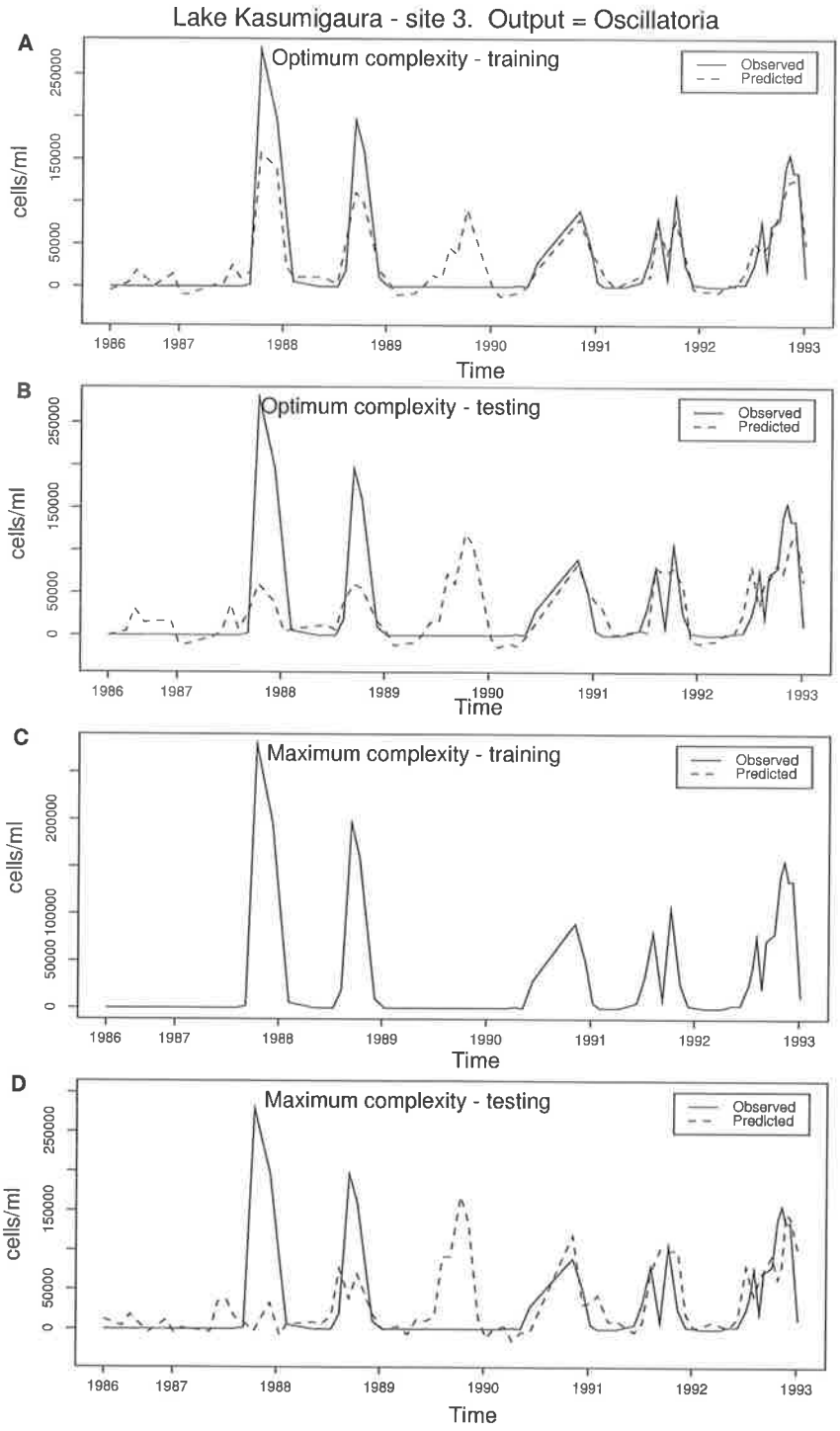

Figure D.5: Model no. 5 – Time series plots of observed and predicted algal abundance. **A** Optimum complexity – training. **B** Optimum complexity – validation. **C** Maximum complexity – training. **D** Maximum complexity – validation.

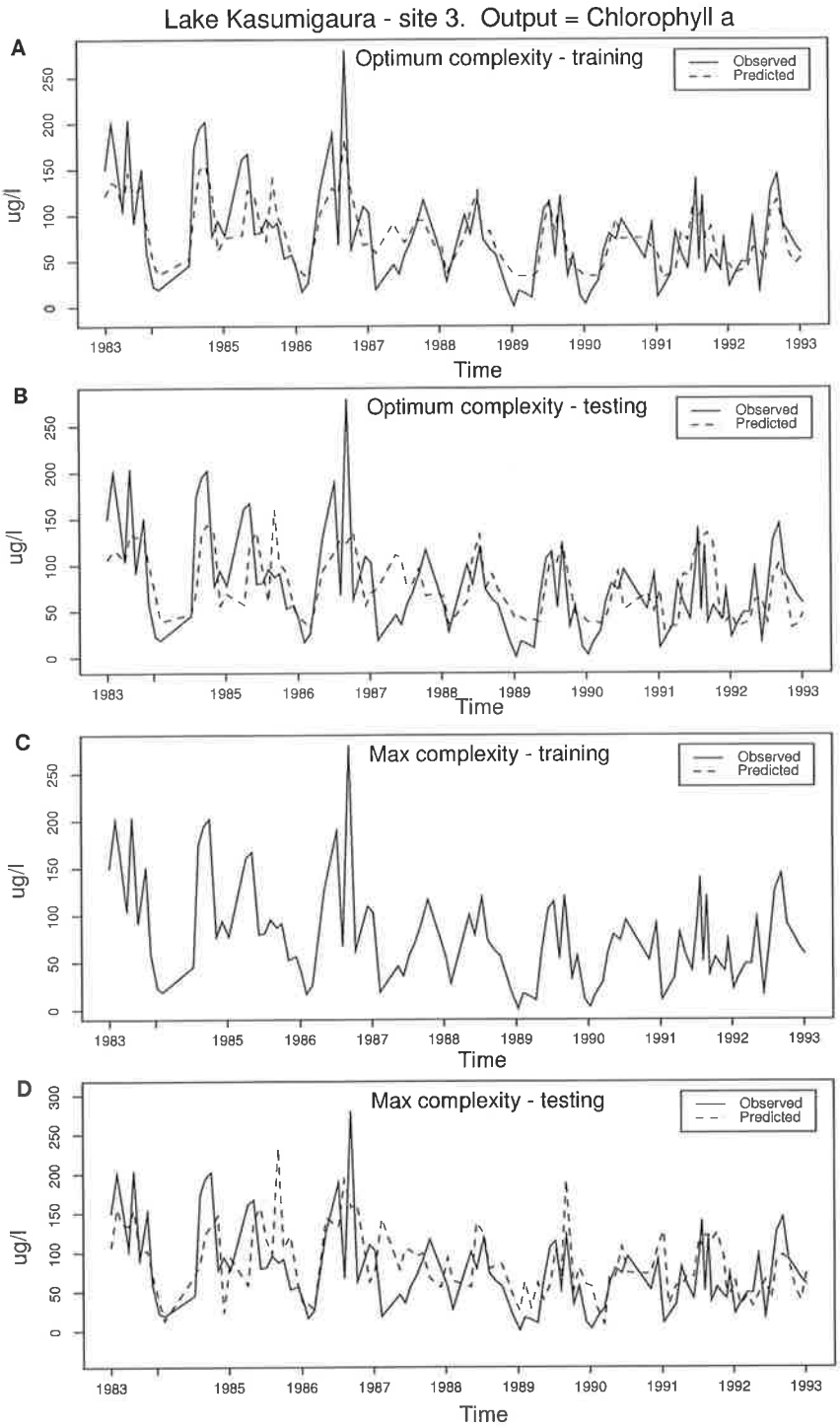

Figure D.6: Model no. 6 – Time series plots of observed and predicted algal abundance. **A** Optimum complexity – training. **B** Optimum complexity – validation. **C** Maximum complexity – training. **D** Maximum complexity – validation.

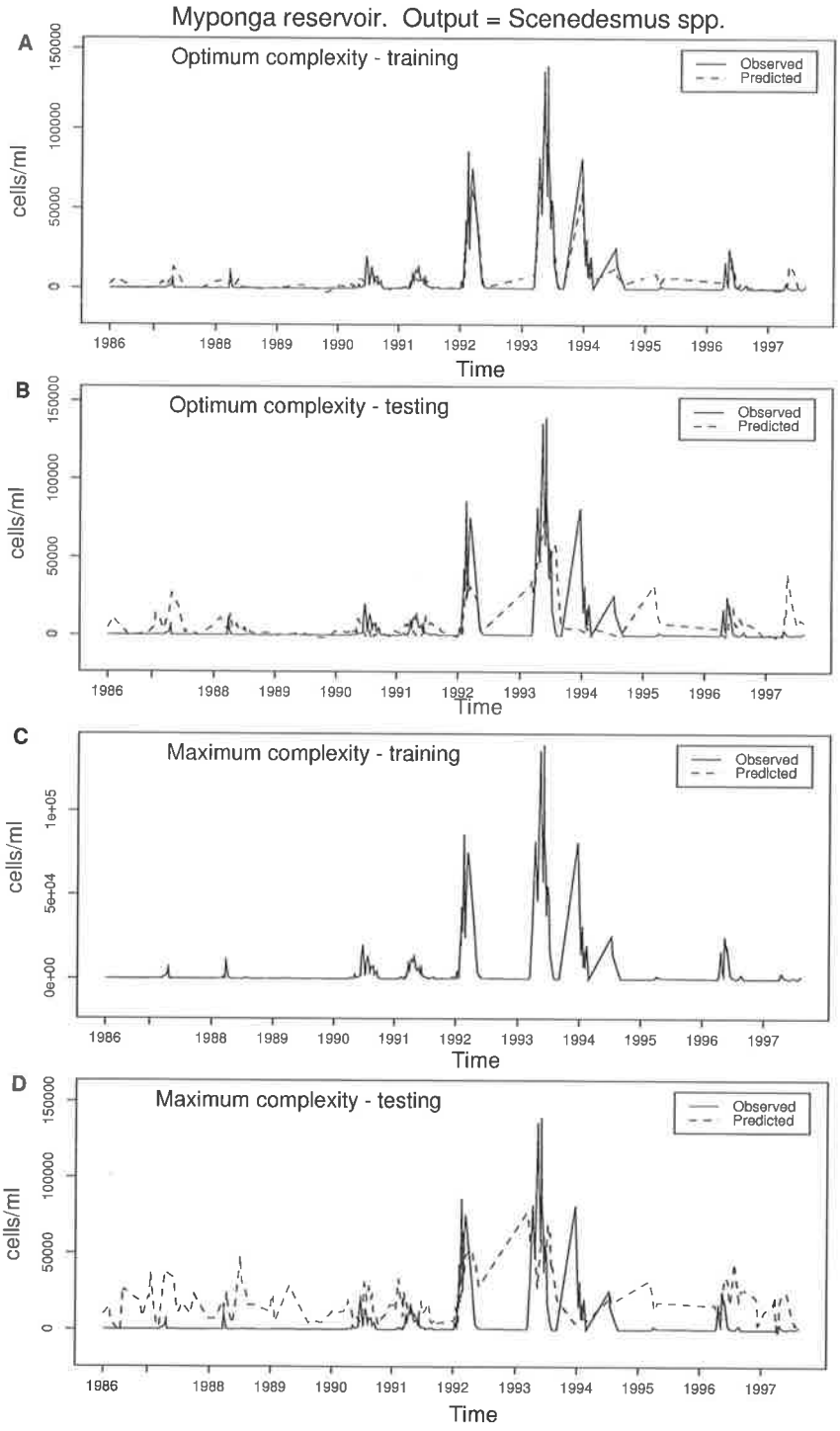

Figure D.7: Model no. 7 – Time series plots of observed and predicted algal abundance. **A** Optimum complexity – training. **B** Optimum complexity – validation. **C** Maximum complexity – training. **D** Maximum complexity – validation.

Appendix E

Effect of Training Algorithm on RMSE

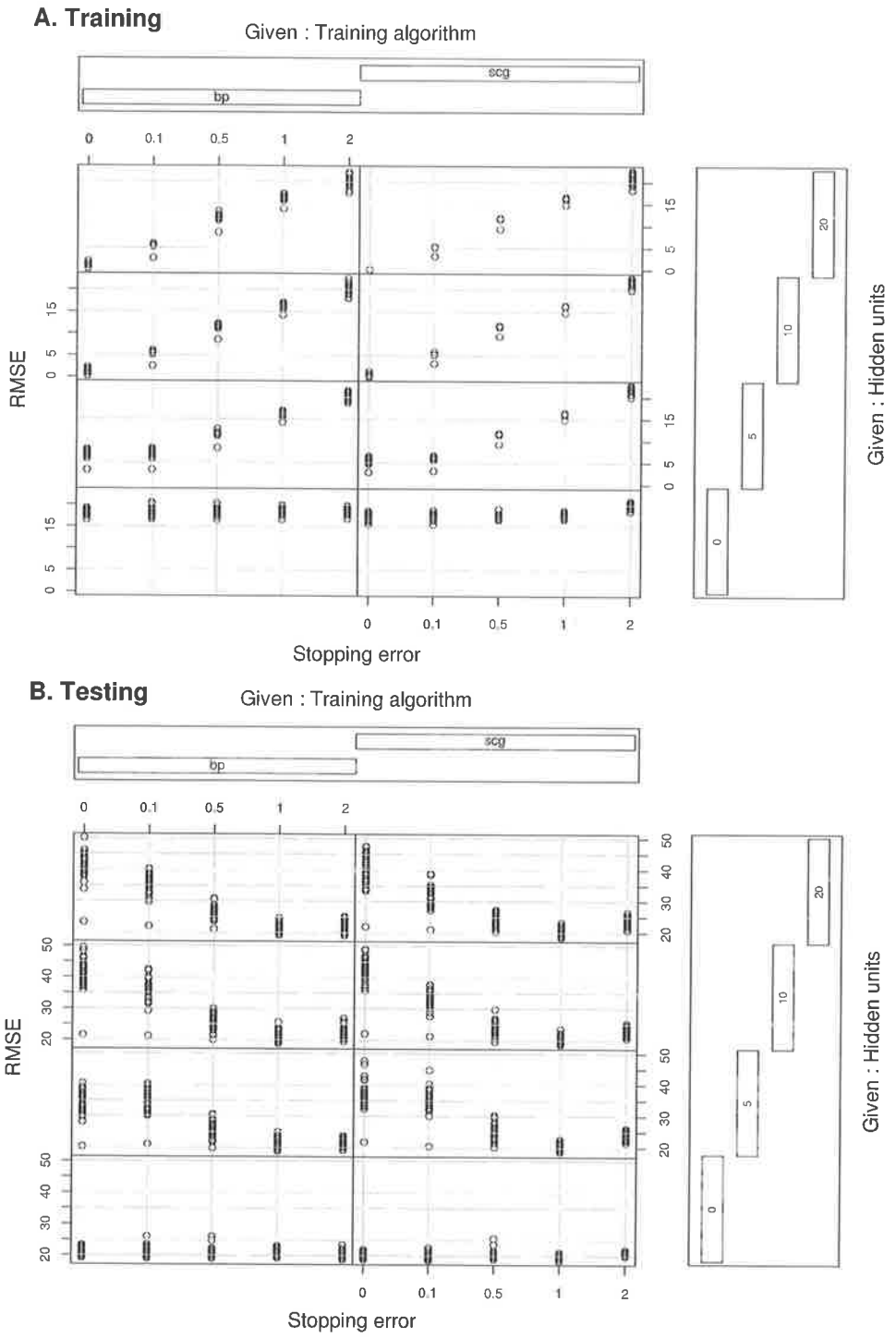

Figure E.1: Model No. 1. **A** Train RMSE v Stop error given No. Hidden units and Training Algorithm. **B** Test RMSE v Stop error given No. Hidden units and Training Algorithm.

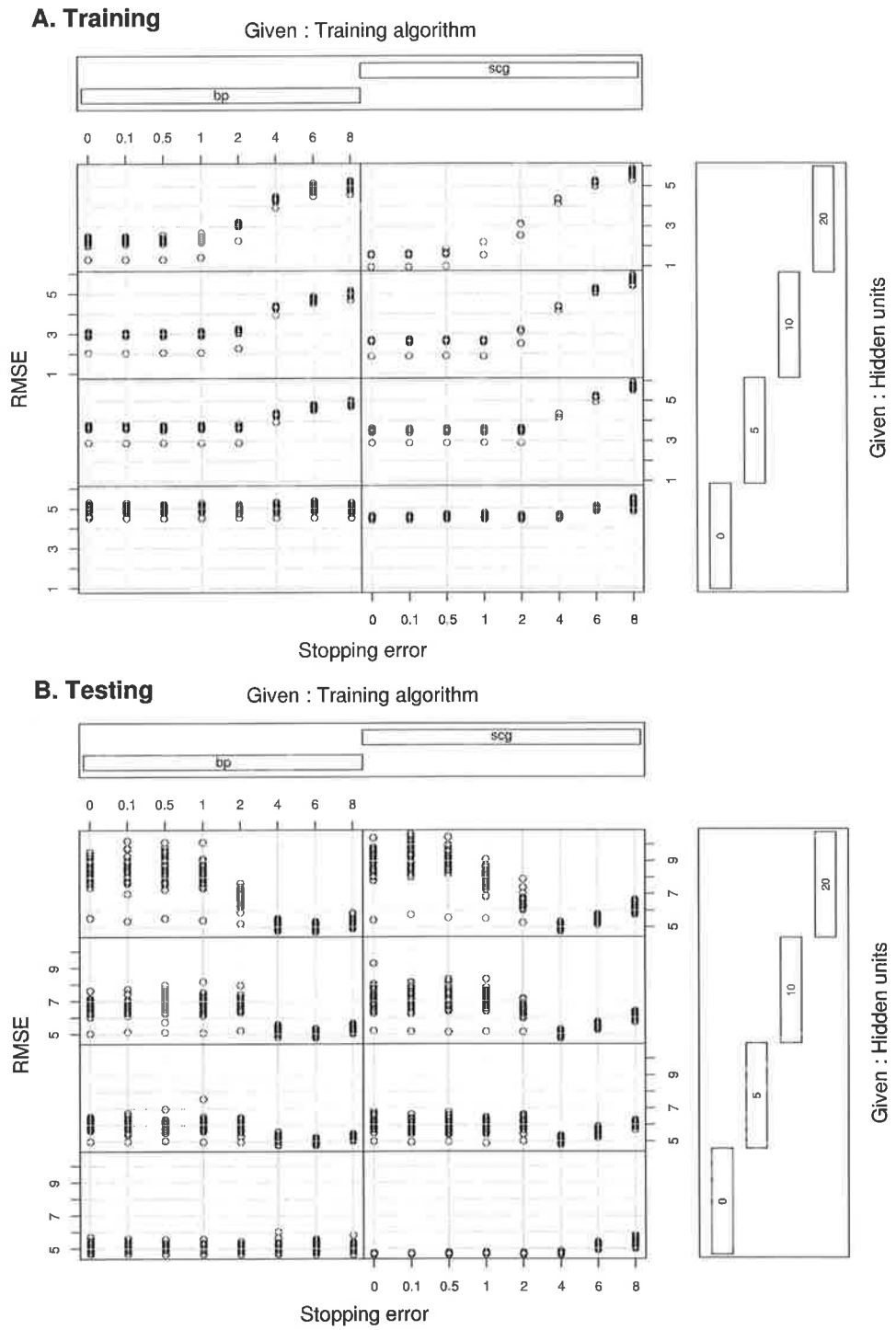

Figure E.2: Model No. 2. **A** Train RMSE v Stop error given No. Hidden units and Training Algorithm. **B** Test RMSE v Stop error given No. Hidden units and Training Algorithm.

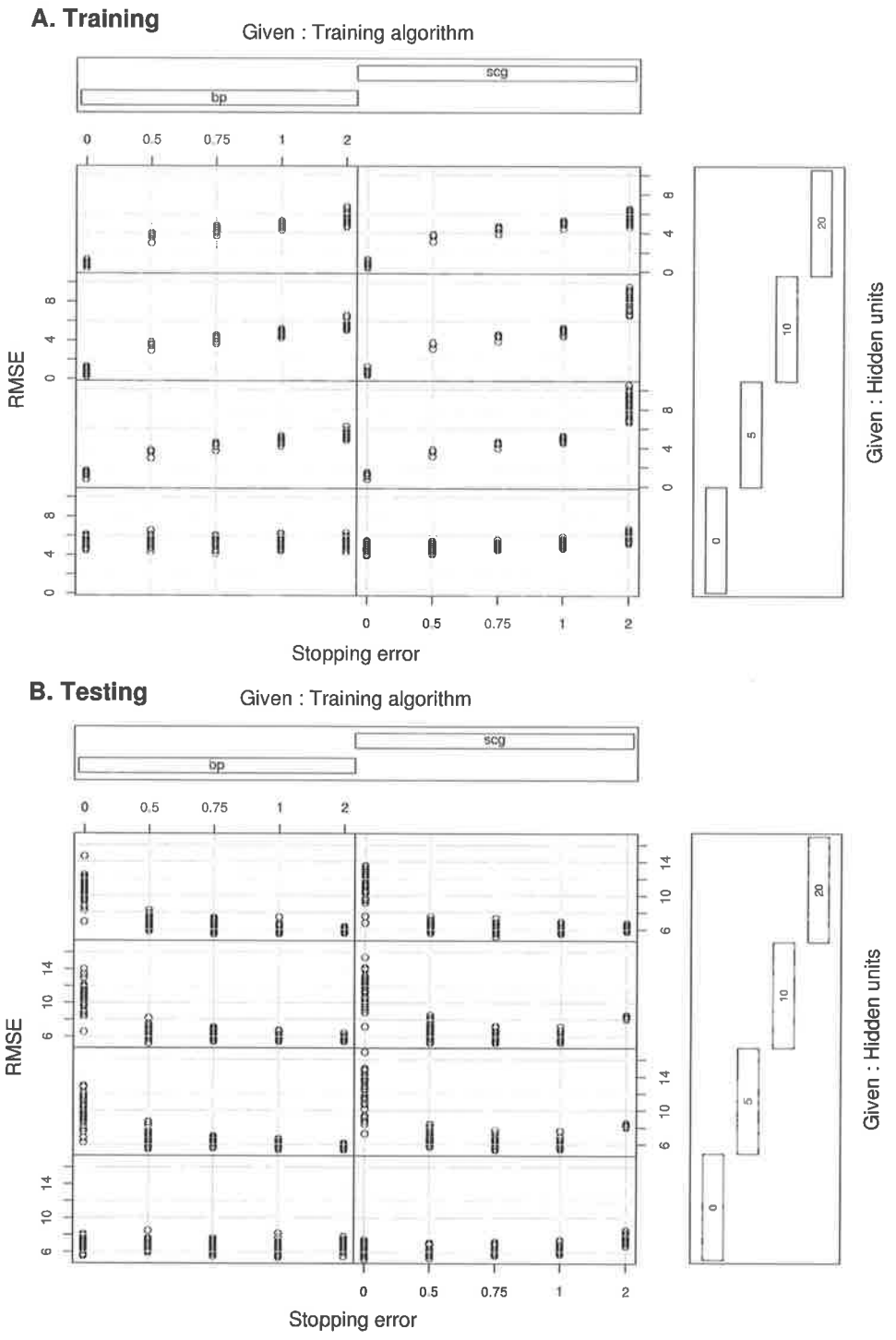

Figure E.3: Model No. 3. **A** Train RMSE v Stop error given No. Hidden units and Training Algorithm. **B** Test RMSE v Stop error given No. Hidden units and Training Algorithm.

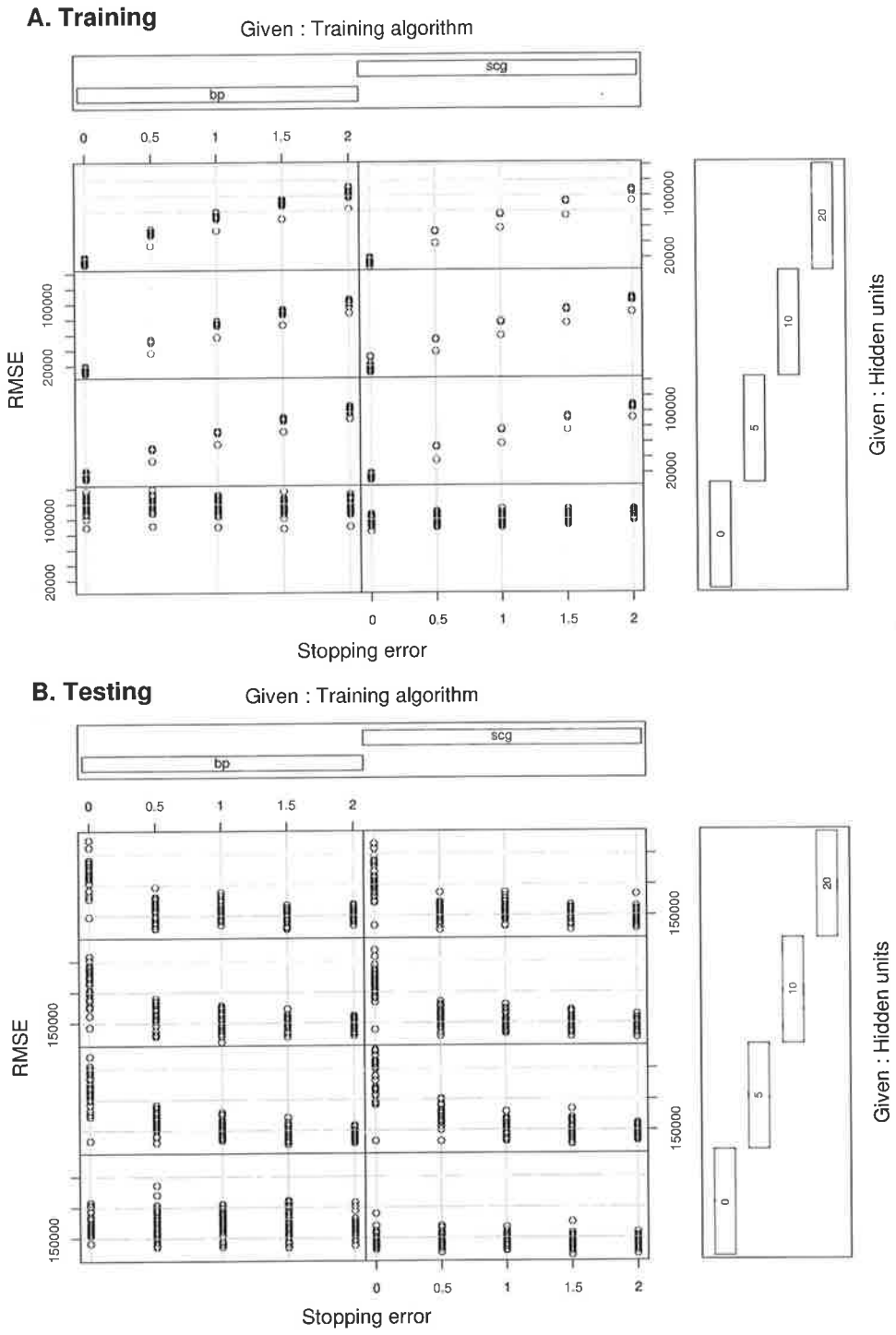

Figure E.4: Model No. 4. **A** Train RMSE v Stop error given No. Hidden units and Training Algorithm. **B** Test RMSE v Stop error given No. Hidden units and Training Algorithm.

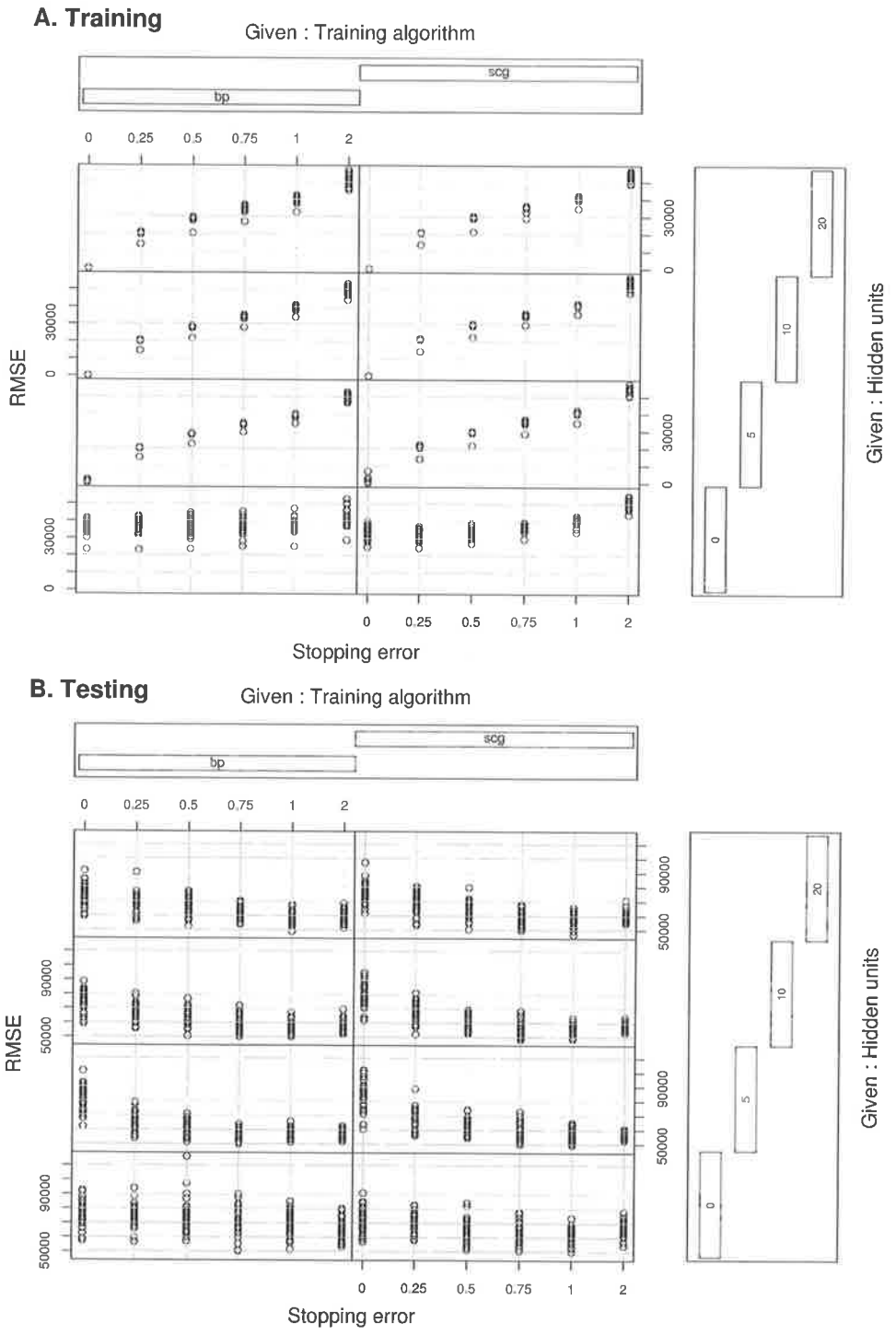

Figure E.5: Model No. 5. **A** Train RMSE v Stop error given No. Hidden units and Training Algorithm. **B** Test RMSE v Stop error given No. Hidden units and Training Algorithm.

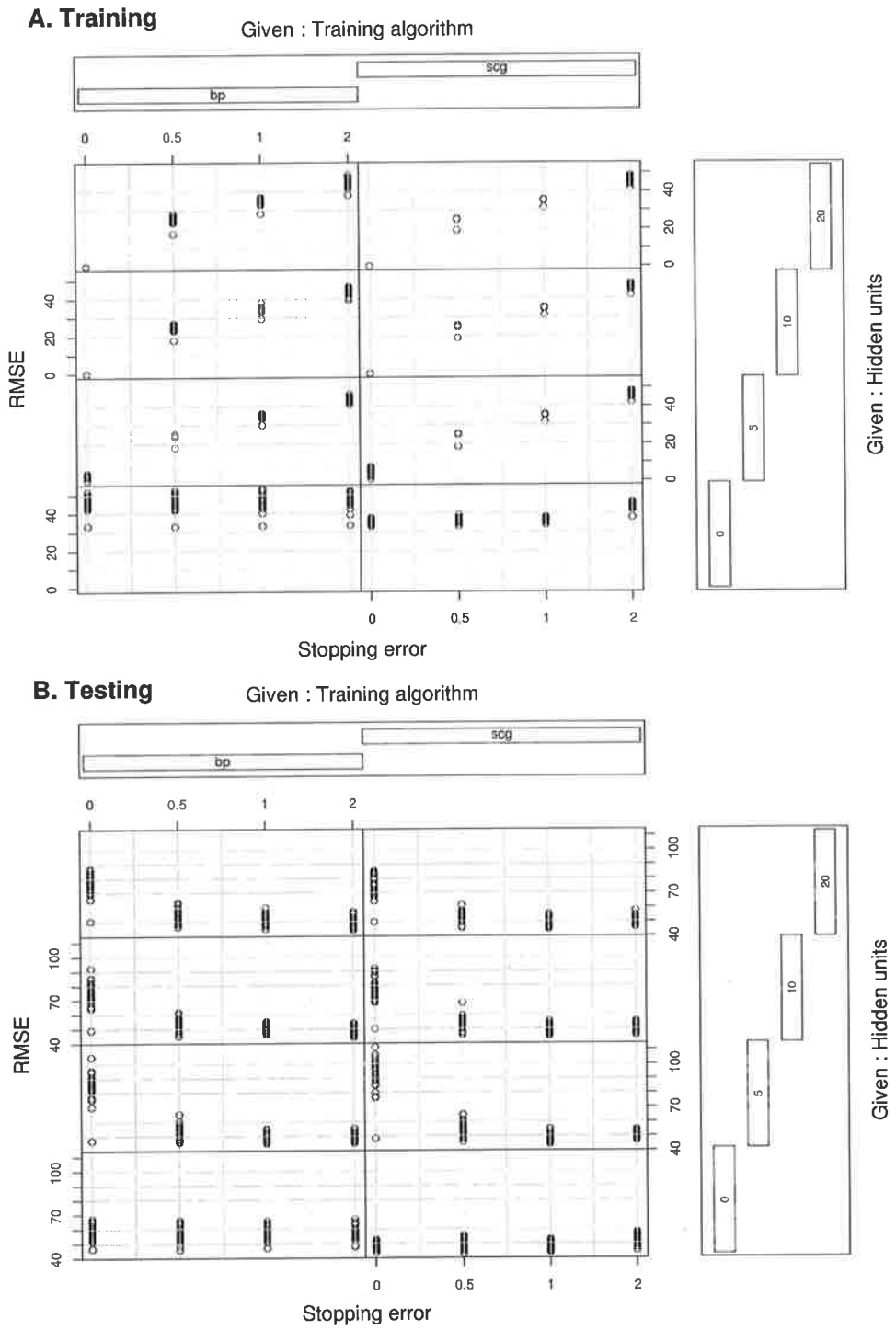

Figure E.6: Model No. 6. **A** Train RMSE v Stop error given No. Hidden units and Training Algorithm. **B** Test RMSE v Stop error given No. Hidden units and Training Algorithm.

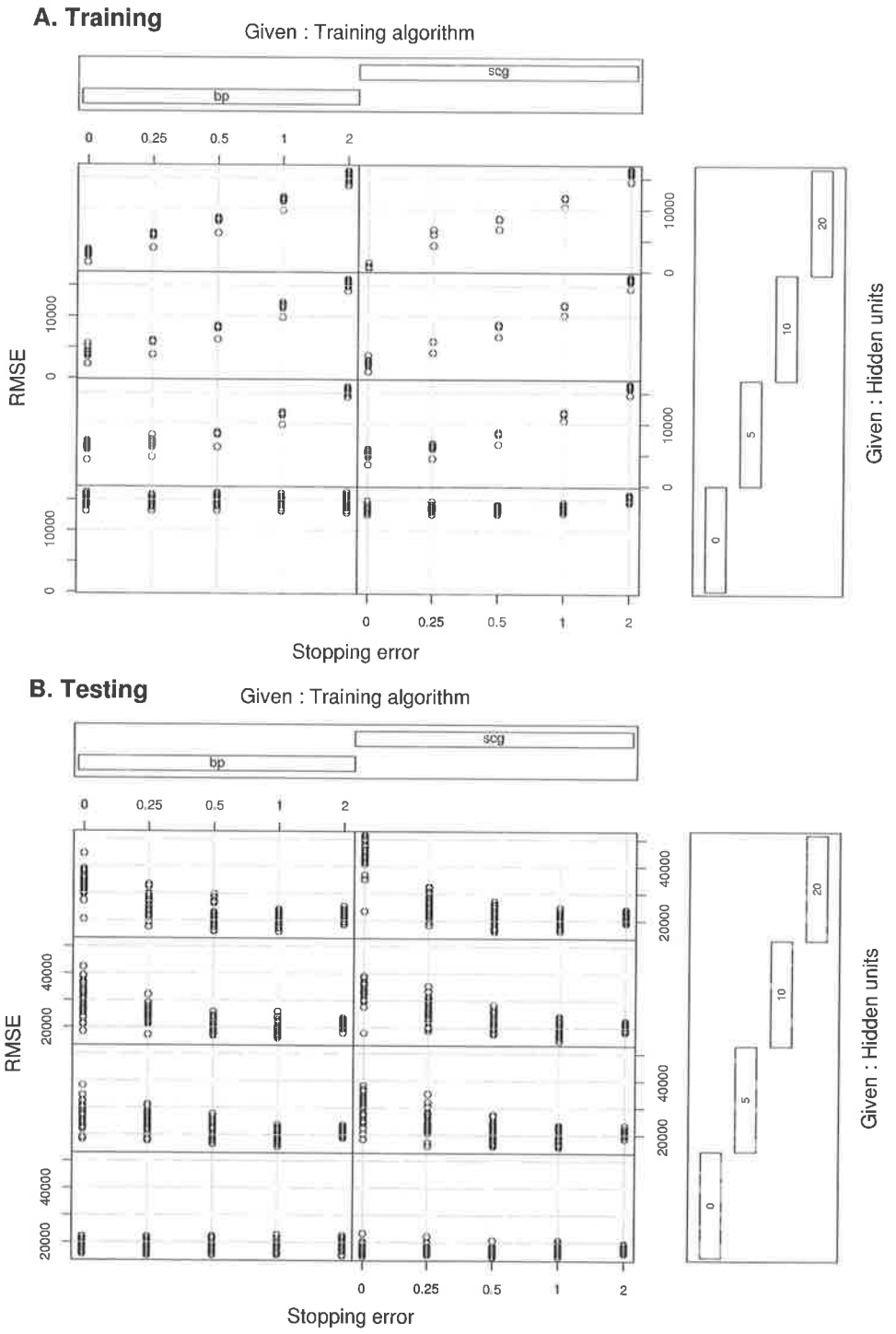

Figure E.7: Model No. 7. **A** Train RMSE v Stop error given No. Hidden units and Training Algorithm. **B** Test RMSE v Stop error given No. Hidden units and Training Algorithm.

Appendix F

Generic Model Predictions

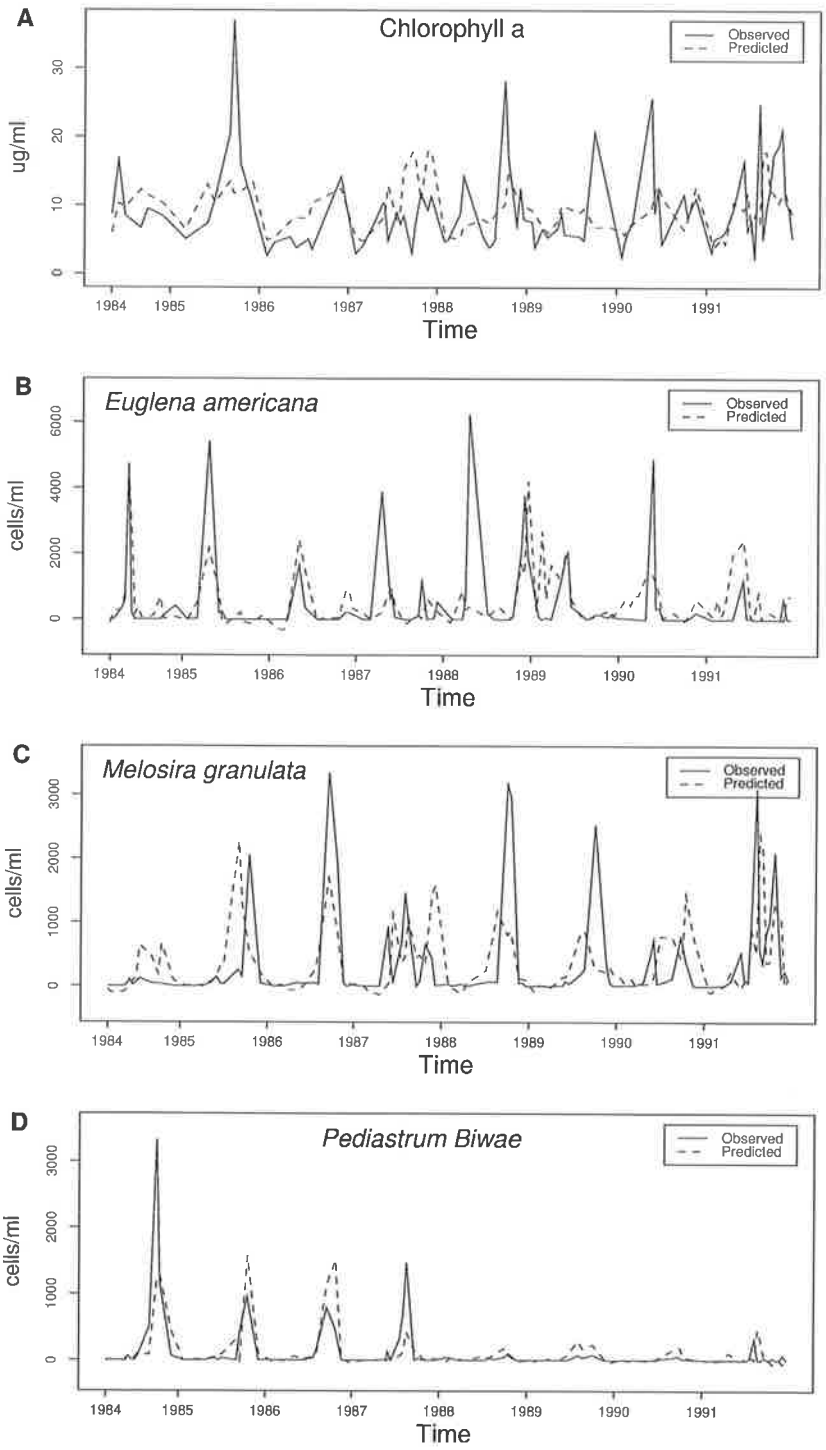

Figure F.1: Lake Biwa. Generic input layer. **A** Chlorophyll *a*. **B** *Euglena americana*. **C** *Melosira granulata*. **D** *Pediastrum biwa*.

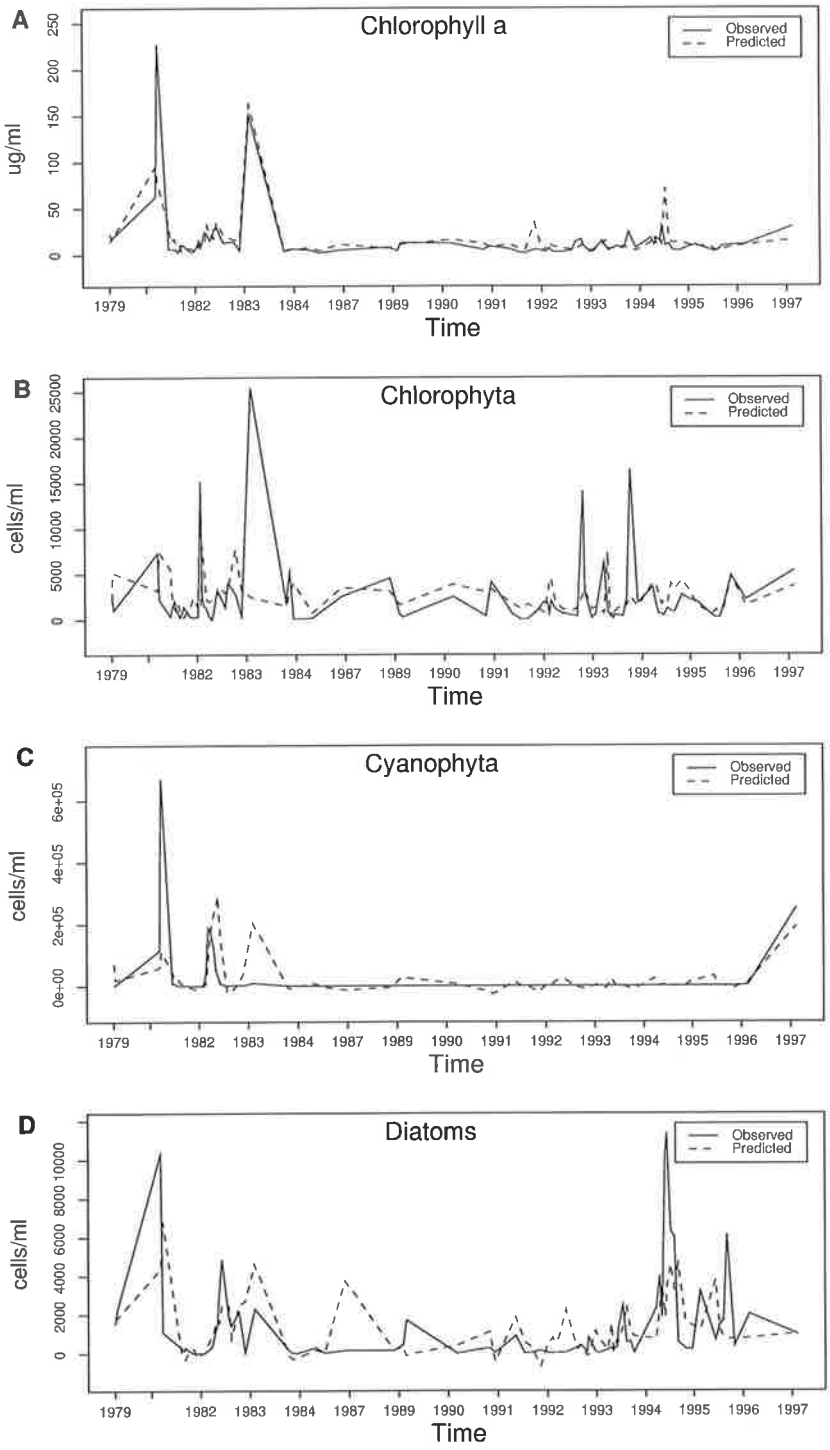

Figure F.2: Burrinjuck Dam. Generic input layer. **A** Chlorophyll *a*. **B** Chlorophyta. **C** Cyanophyta. **D** Diatoms.

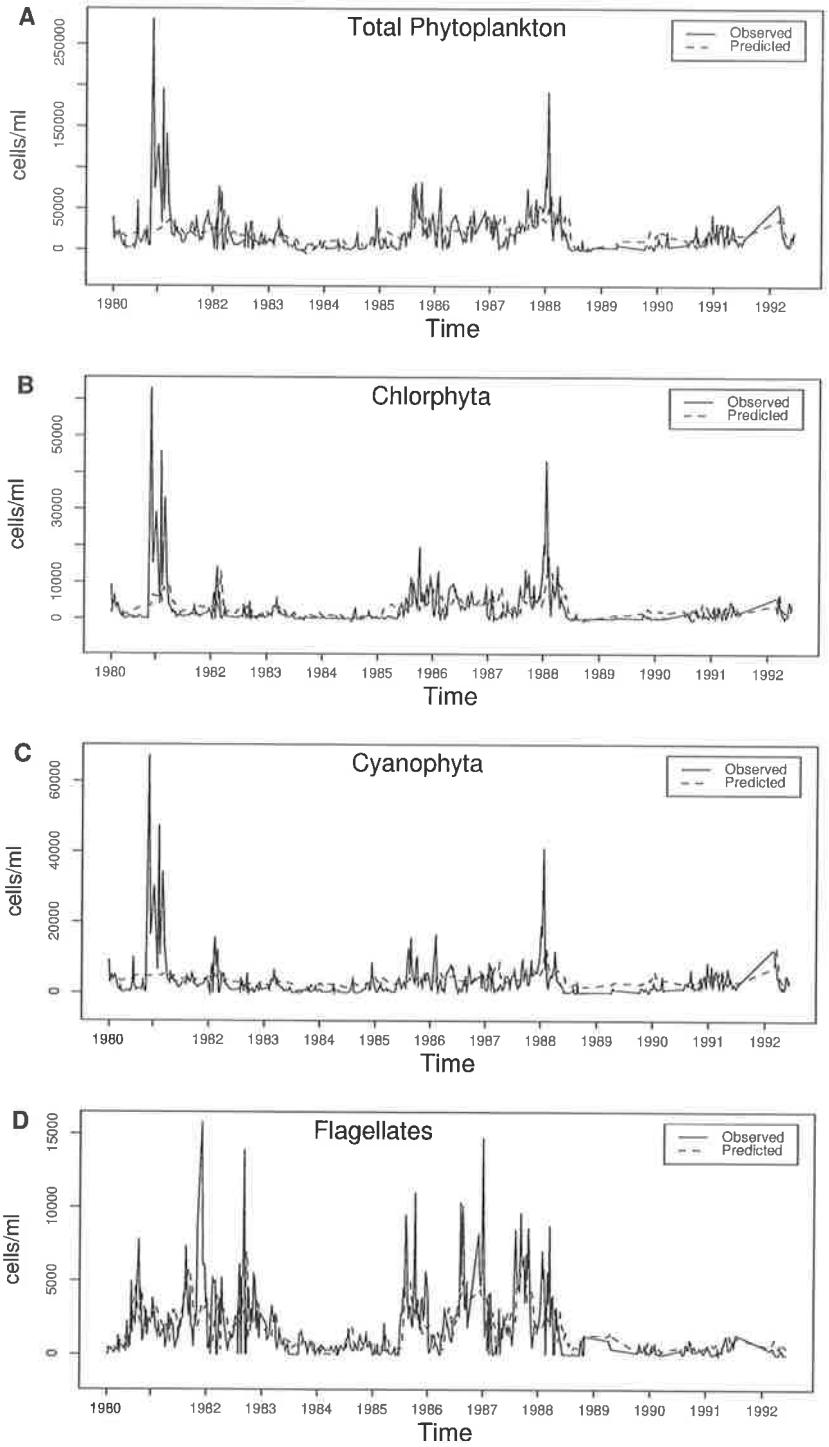

Figure F.3: Darling River. Generic input layer. **A** Total phytoplankton. **B** Chlorophyta. **C** Cyanophyta. **D** Flagellates.

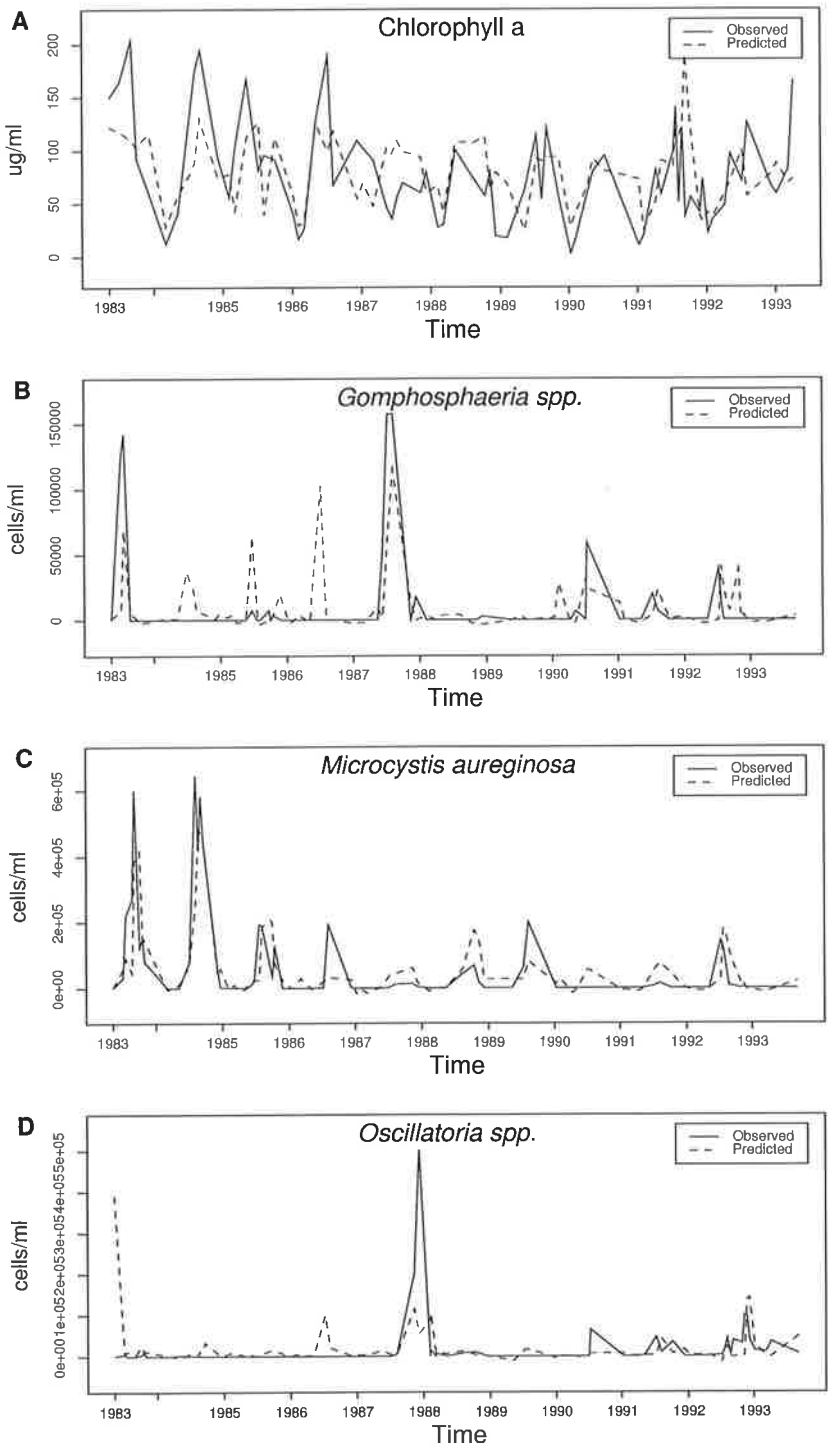

Figure F.4: Lake Kasumigaura. Generic input layer. **A** Chlorophyll *a*. **B** *Gomphosphaeria spp.* **C** *Microcystis aeruginosa*. **D** *Oscillatoria spp.*

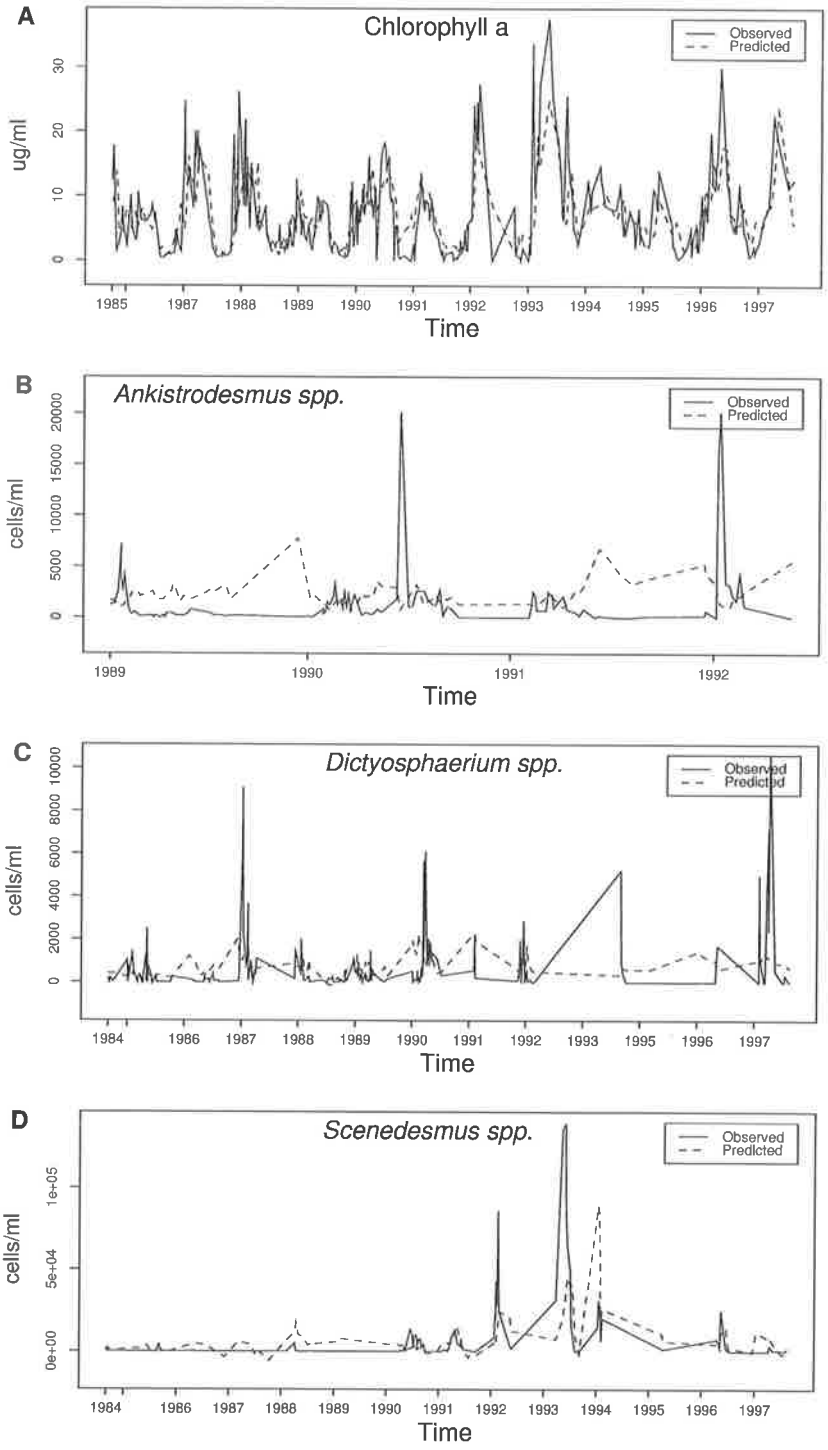

Figure F.5: Myponga Reservoir. Generic input layer. **A** Chlorphyll *a*. **B** *Ankistrodesmus spp.* **C** *Dictyosphaerium spp.* **D** *Scenedesmus spp.*

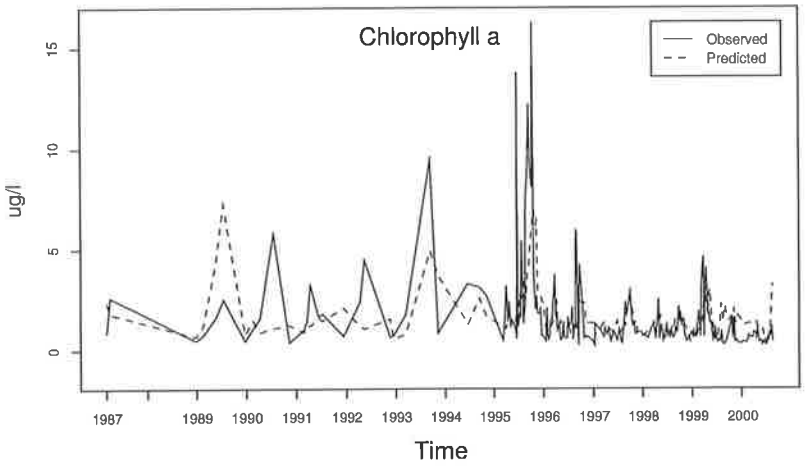

Figure F.6: Lake Soyang. Generic input layer. Chlorophyll *a*

Appendix G

Generic Model Classification Statistics - 14 Day Forecasts

Table G.1: Classification error rates. Lake Biwa – Chlorophyll *a*.

Threshold	Prev.	Sens.	Spec.	PPP	Kappa
4.5	0.86	0.96	0.00	0.86	-0.06
5.4	0.69	0.93	0.26	0.74	0.23
7.15	0.57	0.84	0.45	0.67	0.30
8.6	0.42	0.76	0.55	0.55	0.29
10.8	0.27	0.54	0.77	0.46	0.29

Table G.2: Classification error rates. Lake Biwa – *Euglena americana*.

Threshold	Prev.	Sens.	Spec.	PPP	Kappa
2	0.56	0.95	0.35	0.65	0.32
30	0.46	0.92	0.33	0.53	0.24
110	0.35	0.79	0.45	0.44	0.21
300	0.24	0.74	0.64	0.39	0.29
1000	0.12	0.79	0.92	0.58	0.61

Table G.3: Classification error rates. Lake Biwa – *Melosira granulata*.

Threshold	Prev.	Sens.	Spec.	PPP	Kappa
0.2	0.87	0.84	0.47	0.91	0.25
14	0.69	0.88	0.54	0.81	0.45
46	0.52	0.88	0.57	0.69	0.46
110	0.33	0.86	0.59	0.51	0.38
700	0.15	0.47	0.84	0.35	0.27

Table G.4: Classification error rates. Lake Biwa – *Pediastrum biwae*.

Threshold	Prev.	Sens.	Spec.	PPP	Kappa
0.25	0.46	0.81	0.47	0.57	0.27
5	0.38	0.86	0.57	0.55	0.38
30	0.29	0.73	0.73	0.53	0.42
50	0.20	0.77	0.78	0.46	0.44
150	0.11	0.50	0.90	0.38	0.35

Table G.5: Classification error rates. Burrinjuck Dam – Chlorophyll *a*.

Threshold	Prev.	Sens.	Spec.	PPP	Kappa
3.7	0.85	0.96	0.00	0.84	-0.06
5.3	0.68	0.88	0.33	0.74	0.24
7	0.52	0.84	0.49	0.64	0.33
10.55	0.34	0.76	0.68	0.55	0.40
16	0.18	0.53	0.86	0.44	0.36

Table G.6: Classification error rates. Burrinjuck Dam – Chlorophyta.

Threshold	Prev.	Sens.	Spec.	PPP	Kappa
310	0.82	0.99	0.00	0.81	-0.02
525	0.66	0.98	0.00	0.65	-0.02
950	0.49	0.93	0.18	0.53	0.11
1850	0.34	0.73	0.60	0.49	0.29
2900	0.17	0.67	0.72	0.33	0.28

Table G.7: Classification error rates. Burrinjuck Dam – Cyanophyta.

Threshold	Prev.	Sens.	Spec.	PPP	Kappa
10	0.64	0.67	0.50	0.71	0.16
65	0.51	0.63	0.44	0.54	0.07
300	0.39	0.70	0.47	0.46	0.15
1300	0.25	0.90	0.54	0.40	0.31
10000	0.11	0.89	0.73	0.29	0.32

Table G.8: Classification error rates. Burrinjuck Dam – Diatoms.

Threshold	Prev.	Sens.	Spec.	PPP	Kappa
30	0.82	0.85	0.47	0.88	0.30
135	0.67	0.89	0.44	0.77	0.37
300	0.49	0.90	0.50	0.63	0.40
700	0.35	0.90	0.64	0.58	0.48
2000	0.20	0.50	0.86	0.47	0.36

Table G.9: Classification error rates. Darling River – Total Phytoplankton.

Threshold	Prev.	Sens.	Spec.	PPP	Kappa
3000	0.84	0.97	0.22	0.87	0.26
6500	0.67	0.96	0.27	0.73	0.27
13000	0.49	0.92	0.47	0.63	0.39
21000	0.33	0.68	0.76	0.59	0.42
35000	0.16	0.35	0.95	0.55	0.35

Table G.10: Classification error rates. Darling River – Chlorophyta.

Threshold	Prev.	Sens.	Spec.	PPP	Kappa
165	0.84	0.97	0.29	0.88	0.34
455	0.66	0.97	0.26	0.72	0.28
1000	0.50	0.92	0.37	0.59	0.29
2300	0.34	0.83	0.64	0.54	0.42
6000	0.15	0.41	0.93	0.51	0.37

Table G.11: Classification error rates. Darling River – Cyanophyta.

Threshold	Prev.	Sens.	Spec.	PPP	Kappa
400	0.84	1.00	0.00	0.84	0.00
1000	0.67	1.00	0.07	0.68	0.09
1800	0.50	1.00	0.13	0.53	0.13
3150	0.33	0.83	0.50	0.45	0.27
5500	0.15	0.24	0.93	0.36	0.20

Table G.12: Classification error rates. Darling River – Flagellates.

Threshold	Prev.	Sens.	Spec.	PPP	Kappa
100	0.83	0.97	0.06	0.83	0.05
375	0.66	0.91	0.22	0.70	0.15
800	0.50	0.87	0.69	0.74	0.56
1900	0.32	0.74	0.81	0.65	0.53
3500	0.16	0.44	0.93	0.55	0.41

Table G.13: Classification error rates. Lake Kasumigaura – Chlorophyll *a*.

Threshold	Prev.	Sens.	Spec.	PPP	Kappa
30	0.83	0.98	0.33	0.88	0.41
55	0.69	0.85	0.45	0.77	0.33
70	0.49	0.68	0.47	0.55	0.15
92	0.31	0.50	0.65	0.39	0.14
120	0.17	0.25	0.97	0.60	0.28

Table G.14: Classification error rates. Lake Kasumigaura – *Gomphosphaeria spp.*

Threshold	Prev.	Sens.	Spec.	PPP	Kappa
100	0.20	0.71	0.37	0.21	0.04
5000	0.16	0.57	0.78	0.33	0.27
10000	0.11	0.50	0.83	0.28	0.25
35000	0.08	0.43	0.94	0.38	0.34

Table G.15: Classification error rates. Lake Kasumigaura – *Microcystis aeruginosa*.

Threshold	Prev.	Sens.	Spec.	PPP	Kappa
100	0.55	0.94	0.59	0.74	0.54
2000	0.44	1.00	0.63	0.68	0.60
13000	0.34	1.00	0.60	0.57	0.50
65000	0.23	0.80	0.91	0.73	0.69
160000	0.13	0.55	0.92	0.50	0.45

Table G.16: Classification error rates. Lake Kasumigaura – *Oscillatoria spp.*

Threshold	Prev.	Sens.	Spec.	PPP	Kappa
500	0.37	0.78	0.16	0.35	-0.04
3500	0.31	0.70	0.50	0.39	0.17
6000	0.23	0.65	0.61	0.33	0.20
15000	0.15	0.46	0.84	0.33	0.26
40000	0.08	0.43	0.93	0.33	0.31

Table G.17: Classification error rates. Myponga Reservoir – Chlorophyll *a*.

Threshold	Prev.	Sens.	Spec.	PPP	Kappa
1.4	0.84	0.98	0.22	0.87	0.27
2.8	0.68	0.97	0.42	0.78	0.45
5.5	0.50	0.86	0.65	0.71	0.51
8.4	0.34	0.66	0.84	0.67	0.50
12	0.17	0.47	0.93	0.57	0.42

Table G.18: Classification error rates. Myponga Reservoir – *Ankistrodesmus spp.*

Threshold	Prev.	Sens.	Spec.	PPP	Kappa
150	0.83	1.00	0.00	0.83	0.00
400	0.67	1.00	0.00	0.67	0.00
800	0.51	1.00	0.00	0.51	0.00
1300	0.35	1.00	0.00	0.35	0.00
2500	0.17	0.78	0.02	0.14	-0.07

Table G.19: Classification error rates. Myponga Reservoir – *Dictyosphaerium spp.*

Threshold	Prev.	Sens.	Spec.	PPP	Kappa
3	0.81	0.98	0.05	0.82	0.04
100	0.64	0.97	0.03	0.64	-0.00
220	0.50	0.97	0.15	0.53	0.12
460	0.35	0.84	0.38	0.43	0.18
1100	0.16	0.40	0.79	0.27	0.16

Table G.20: Classification error rates. Myponga Reservoir – *Scenedesmus spp.*

Threshold	Prev.	Sens.	Spec.	PPP	Kappa
20	0.84	0.86	0.50	0.90	0.33
100	0.68	0.87	0.35	0.74	0.24
750	0.51	0.90	0.40	0.61	0.30
2500	0.35	0.77	0.49	0.45	0.22
10000	0.18	0.64	0.86	0.52	0.46

Table G.21: Classification error rates. Lake Soyang – Chlorophyll *a*.

Threshold	Prev.	Sens.	Spec.	PPP	Kappa
0.47	0.84	0.99	0.03	0.84	0.03
0.66	0.67	0.95	0.12	0.68	0.09
0.93	0.51	0.88	0.31	0.57	0.19
1.45	0.32	0.61	0.79	0.59	0.40
2.3	0.16	0.42	0.88	0.41	0.30

Appendix H

Generic Model Classification Statistics - 7 Fay Forecasts

Table H.1: Classification error rates. Darling – 7 day forecast – Total phytoplankton.

Threshold	Prev.	Sens.	Spec.	PPP	Kappa
3000	0.84	0.97	0.27	0.87	0.32
6500	0.68	0.95	0.27	0.73	0.26
13000	0.49	0.87	0.50	0.63	0.37
21000	0.34	0.66	0.77	0.59	0.42
35000	0.16	0.56	0.92	0.56	0.47

Table H.2: Classification error rates. Darling – 7 day forecast – Chlorophyta.

Threshold	Prev.	Sens.	Spec.	PPP	Kappa
165	0.84	0.95	0.38	0.89	0.39
455	0.67	0.94	0.34	0.74	0.33
1000	0.50	0.91	0.41	0.61	0.33
2300	0.34	0.79	0.76	0.63	0.52
6000	0.15	0.54	0.91	0.51	0.44

Table H.3: Classification error rates. Darling – 7 day forecast – Cyanophyta.

Threshold	Prev.	Sens.	Spec.	PPP	Kappa
400	0.84	0.98	0.26	0.88	0.33
1000	0.67	0.95	0.28	0.73	0.27
1800	0.50	0.90	0.42	0.61	0.32
3150	0.34	0.58	0.77	0.56	0.34
5500	0.15	0.49	0.92	0.53	0.43

Table H.4: Classification error rates. Darling – 7 day forecast – Flagellates.

Threshold	Prev.	Sens.	Spec.	PPP	Kappa
100	0.83	0.98	0.00	0.83	-0.03
375	0.66	0.92	0.20	0.69	0.13
800	0.50	0.88	0.71	0.75	0.59
1900	0.32	0.76	0.83	0.68	0.58
3500	0.15	0.52	0.94	0.60	0.48

Table H.5: Classification error rates. Myponga – 7 day forecast – Chlorophyll *a*.

Threshold	Prev.	Sens.	Spec.	PPP	Kappa
1.4	0.85	0.95	0.52	0.92	0.50
2.8	0.68	0.95	0.57	0.82	0.58
5.5	0.52	0.89	0.70	0.76	0.59
8.4	0.34	0.76	0.80	0.66	0.54
12	0.17	0.44	0.93	0.55	0.40

Table H.6: Classification error rates. Myponga – 7 day forecast – *Ankistrodesmus spp.*

Threshold	Prev.	Sens.	Spec.	PPP	Kappa
150	0.83	0.94	0.00	0.83	-0.08
400	0.68	0.89	0.12	0.68	0.01
800	0.51	0.91	0.50	0.66	0.41
1300	0.37	0.79	0.62	0.55	0.37
2500	0.17	0.11	0.82	0.12	-0.07

Table H.7: Classification error rates. Myponga – 7 day forecast – *Dictyosphaerium spp.*

Threshold	Prev.	Sens.	Spec.	PPP	Kappa
3	0.80	0.98	0.00	0.80	-0.03
100	0.64	0.97	0.10	0.66	0.09
220	0.49	0.92	0.23	0.53	0.14
460	0.36	0.86	0.49	0.48	0.30
1100	0.16	0.53	0.81	0.35	0.27

Table H.8: Classification error rates. Myponga – 7 day forecast – *Scenedesmus spp.*

Threshold	Prev.	Sens.	Spec.	PPP	Kappa
20	0.84	0.92	0.26	0.87	0.20
100	0.68	0.92	0.19	0.71	0.13
750	0.51	0.96	0.31	0.59	0.27
2500	0.35	0.90	0.59	0.54	0.43
10000	0.19	0.54	0.91	0.58	0.45

Table H.9: Classification error rates. Soyang – 7 day forecast – Chlorophyll *a*.

Threshold	Prev.	Sens.	Spec.	PPP	Kappa
0.47	0.82	0.98	0.11	0.84	0.13
0.66	0.65	0.91	0.24	0.69	0.17
0.93	0.49	0.79	0.52	0.61	0.30
1.45	0.29	0.67	0.82	0.61	0.48
2.3	0.14	0.41	0.92	0.48	0.36

Appendix I

Generic Model Sensitivity Analyses

Table I.1: Sensitivity Analysis. Lake Biwa – chlorophyll *a*.

Input	Sens.	R
chlorophyll a	0.24	0.48
nitrate	0.22	-0.45
orthophosphate	0.20	-0.12
water temperature	0.19	0.28
secchi depth	0.15	0.03

Table I.2: Sensitivity Analysis. Lake Biwa – *Euglena americana*.

Input	Sens.	R
Euglena americana	0.29	0.47
water temperature	0.23	0.20
nitrate	0.20	0.48
orthophosphate	0.15	-0.32
secchi depth	0.12	-0.10

Table I.3: Sensitivity Analysis. Lake Biwa – *Melosira granulata*.

Input	Sens.	R
Melosira granulata	0.26	0.31
water temperature	0.25	0.49
secchi depth	0.18	-0.24
orthophosphate	0.15	-0.10
nitrate	0.15	-0.11

Table I.4: Sensitivity Analysis. Lake Biwa – *Pediastrum biwae*.

Input	Sens.	R
Pediastrum biwae	0.63	0.74
orthophosphate	0.11	0.17
secchi depth	0.09	0.30
water temperature	0.09	0.46
nitrate	0.07	-0.05

Table I.5: Sensitivity Analysis. Burrinjuck Dam – chlorophyll *a*.

Input	Sens.	R
chlorophyll a	0.36	0.46
water temperature	0.19	0.38
total oxidised nitrogen	0.18	-0.05
dissolved inorganic phosphorous	0.15	0.06
secchi depth	0.12	0.09

Table I.6: Sensitivity Analysis. Burrinjuck Dam – chlorophyta.

Input	Sens.	R
chlorophyta	0.32	0.40
secchi depth	0.22	0.06
water temperature	0.18	0.03
total oxidised nitrogen	0.15	-0.47
dissolved inorganic phosphorous	0.14	-0.22

Table I.7: Sensitivity Analysis. Burrinjuck Dam – cyanophyta.

Input	Sens.	R
cyanophyta	0.56	0.65
water temperature	0.15	0.49
total oxidised nitrogen	0.12	0.42
secchi depth	0.10	-0.29
dissolved inorganic phosphorous	0.07	-0.23

Table I.8: Sensitivity Analysis. Burrinjuck Dam – diatoms.

Input	Sens.	R
diatoms	0.30	0.65
dissolved inorganic phosphorous	0.23	-0.68
water temperature	0.18	0.53
secchi depth	0.17	-0.57
total oxidised nitrogen	0.12	0.22

Table I.9: Sensitivity Analysis. Darling River – chlorophyta.

Input	Sens.	R
chlorophyta	0.33	0.58
water temperature	0.17	0.29
turbidity	0.14	-0.30
total oxidised nitrogen	0.13	-0.48
soluble reactive phosphorous	0.12	-0.34
flow	0.11	-0.43

Table I.10: Sensitivity Analysis. Darling River – cyanophyta.

Input	Sens.	R
cyanophyta	0.35	0.10
water temperature	0.18	0.32
soluble reactive phosphorous	0.15	-0.28
turbidity	0.12	-0.36
total oxidised nitrogen	0.10	-0.39
flow	0.09	-0.31

Table I.11: Sensitivity Analysis. Darling River – flagellates.

Input	Sens.	R
flagellates	0.22	0.44
total oxidised nitrogen	0.18	0.17
water temperature	0.17	-0.18
turbidity	0.16	-0.14
soluble reactive phosphorous	0.14	-0.20
flow	0.12	-0.06

Table I.12: Sensitivity Analysis. Darling River – total phytoplankton.

Input	Sens.	R
total phytoplankton	0.31	0.42
water temperature	0.17	0.30
soluble reactive phosphorous	0.14	-0.45
turbidity	0.14	-0.47
flow	0.12	-0.46
total oxidised nitrogen	0.12	-0.45

Table I.13: Sensitivity Analysis. Lake Kasumigaura – chlorophyll *a*.

Input	Sens.	R
orthophosphate	0.24	0.46
secchi depth	0.20	0.24
chlorophyll a	0.20	0.51
nitrate	0.19	-0.36
water temperature	0.17	0.18

Table I.14: Sensitivity Analysis. Lake Kasumigaura – *Gomphosphaeria spp.*

Input	Sens.	R
Gomphosphaeria spp.	0.47	0.75
orthophosphate	0.15	0.35
nitrate	0.13	-0.35
secchi depth	0.13	0.03
water temperature	0.12	-0.31

Table I.15: Sensitivity Analysis. Lake Kasumigaura – *Microcystis aeruginosa*.

Input	Sens.	R
Microcystis aeruginosa	0.48	0.78
orthophosphate	0.16	-0.04
water temperature	0.15	0.38
secchi depth	0.12	0.31
nitrate	0.08	-0.15

Table I.16: Sensitivity Analysis. Lake Kasumigaura – *Oscillatoria spp.*

Input	Sens.	R
Oscillatoria spp.	0.38	0.29
secchi depth	0.22	-0.29
orthophosphate	0.16	-0.34
water temperature	0.15	0.11
nitrate	0.09	-0.13

Table I.17: Sensitivity Analysis. Myponga Reservoir – *Ankistrodesmus spp.*

Input	Sens.	R
nitrate	0.26	-0.03
water temperature	0.23	-0.07
turbidity	0.20	0.07
filter reactive phosphorous	0.19	0.09
Ankistrodesmus spp.	0.14	0.13

Table I.18: Sensitivity Analysis. Myponga Reservoir – chlorophyll *a*.

Input	Sens.	R
water temperature	0.22	0.25
nitrate	0.21	0.10
turbidity	0.20	-0.06
filter reactive phosphorous	0.19	-0.16
chlorophyll a	0.18	0.45

Table I.19: Sensitivity Analysis. Myponga Reservoir – *Dictyosphaerium spp.*

Input	Sens.	R
turbidity	0.26	-0.52
nitrate	0.22	0.26
filter reactive phosphorous	0.21	-0.18
Dictyosphaerium spp.	0.17	0.03
water temperature	0.13	0.23

Table I.20: Sensitivity Analysis. Myponga Reservoir – *Scenedesmus spp.*

Input	Sens.	R
Scenedesmus spp.	0.40	0.75
water temperature	0.19	0.30
turbidity	0.16	0.30
filter reactive phosphorous	0.13	-0.36
nitrate	0.12	-0.07

Table I.21: Sensitivity Analysis. Lake Soyang – chlorophyll *a*.

Input	Sens.	R
chlorophyll a	0.27	0.71
water temperature	0.24	0.31
dissolved inorganic phosphorous	0.20	-0.37
secchi depth	0.15	-0.35
nitrate	0.13	0.20

Appendix J

Starting Models for Strip Mining

Table J.1: Staring model – Lake Biwa

Input variable	In. grp.	Window end	Window start
Water quality & physical conditions			
Chlorophyll *a*	+gen	-7	-37
Dissolved oxygen		-7	-37
Nitrate	+gen	-7	-37
Orthophosphate	+gen	-7	-37
pH	+pH	-7	-37
Secchi depth	+gen	-7	-37
Si	+Si	-7	-37
Water temperature	+gen	-7	-37
Weather (fine, cloudy, rain)	+wea	-7	-37
Wind speed	+wea	-7	-37
Phytoplankton			
Euglena americana	+spe	-7	-37
Melosira granulata	+spe	-7	-37
Cyclotella glomerata		-7	-37
Asterionella formosa		-7	-37
Rhodomonas spp.		-7	-37
Micractinium pusillum		-7	-37
Dictyosphaerium sp		-7	-37
Ankistrodesmus fal v mirabile		-7	-37
Pediastrum biwae	+spe	-7	-37
Coelastrum cambricum		-7	-37

All inputs repeated for -7 – -67 window.

Table J.2: Staring model – Burrinjuck Dam

Input variable	In. grp.	Window end	Window start
Water quality & physical conditions			
Area		-7	-14
Chlorophyll *a*	+gen	-7	-37
Dissolved P	+gen	-7	-37
Dissolved oxygen		-7	-37
Evaporation	+wea	-7	-14
Precipitation	+wea	-7	-14
Relative humidity 900	+wea	-7	-14
Relative humidity 1500	+wea	-7	-14
Secchi depth	+gen	-7	-37
Wind speed 900	+wea	-7	-14
Wind speed 1500	+wea	-7	-14
Stratification	+str	-7	-37
Sunshine hrs	+wea	-7	-14
Water temperature	+gen	-7	-37
Air temp, max	+wea	-7	-14
Air temp, min	+wea	-7	-14
Total P		-7	-37
Total oxidised N	+gen	-7	-37
Volume		-7	-14
Water level	+dep	-7	-14
Inflow			
Ginnind & Charnwood	+inf	-7	-14
Goodradigbee	+inf	-7	-14
Molonglo Coppins	+inf	-7	-14
Mountain Creek	+inf	-7	-14
Murrum MtMcD	+inf	-7	-14
S410008	+inf	-7	-14
S410700	+inf	-7	-14
S410731	+inf	-7	-14
S410745	+inf	-7	-14
S410761	+inf	-7	-14
Yass	+inf	-7	-14
Phytoplankton			
Chlorophyta	+spe	-7	-37
Cyanophyta	+spe	-7	-37
Diatoms	+spe	-7	-37

All inputs repeated for -7 – -67 window.

Table J.3: Staring model – Darling river

Input variable	In. grp.	Window end	Window start
Water quality & physical conditions			
Bicarbonate		-7	-37
Calcium		-7	-21
Chloride		-7	-37
Colour		-7	-37
E.C. - Field		-7	-21
E.C. - Lab		-7	-21
Flow	+gen	-7	-14
Magnesium		-7	-21
NO_x	+gen	-7	-21
pH - Field	+pH	-7	-21
pH - Lab		-7	-21
Potassium		-7	-21
Silica	+Si	-7	-21
Sodium		-7	-21
Sol React Phosphorus	+gen	-7	-21
Sulphate		-7	-37
Temperature	+gen	-7	-21
Tot Phosphorus		-7	-21
Turbidity	+gen	-7	-21
Phytoplankton			
Centric diatoms		-7	-21
Chlorococcales		-7	-21
Chlorophyta	+spe	-7	-21
Cyanophyta	+spe	-7	-21
Diatoms unicellular		-7	-21
Ditomophyta		-7	-21
Flagellates	+spe	-7	-21
Planctonema		-7	-21
Scenedesmus		-7	-21
Total phytoplankton	+gen	-7	-21
Ulothricales		-7	-21

All inputs repeated for -7 – -67 window.

Table J.4: Staring model – Lake Kasumigaura

Input variable	In. grp.	Window end	Window start
Water quality & physical conditions			
Chlorophyll *a*	+gen	-7	-37
Dissolved inorganic N		-7	-37
Dissolved oxygen		-7	-37
Dissolved total P	+gen	-7	-37
Light	+wea	-7	-37
NH_4		-7	-37
NO_2		-7	-37
NO_3	+gen	-7	-37
pH	+pH	-7	-37
PO_4	+gen	-7	-37
Radiation Time (Kashima)	+wea	-7	-14
Radiation Time (Tsuchiura)	+wea	-7	-14
Rain (Kashima)	+wea	-7	-14
Rain (Tsuchiura)	+wea	-7	-14
Si	+Si	-7	-37
Total N		-7	-37
Total P		-7	-37
Secchi depth	+gen	-7	-37
Water temperature	+gen	-7	-37
Phytoplankton			
Merismopedia spp.		-7	-37
Oscillatoria spp.	+spe	-7	-37
Phormidium spp.		-7	-37
Cyclotella sp. 1		-7	-37
Synedra rumpens		-7	-37
Anabaena flos–aquae		-7	-37
Ochromonas spp.		-7	-37
Microcystis aeruginosa	+spe	-7	-37
Microcystis wesen		-7	-37
Gomphosphaeria spp.	+spe	-7	-37
Zooplankton			
Bosmina fatalis	+zoo	-7	-37
Cladocera	+zoo	-7	-37
Copepoda	+zoo	-7	-37
Diaphanosoma brachyurum	+zoo	-7	-37
Rotifera	+zoo	-7	-37
Total zooplankton	+zoo	-7	-37

All inputs repeated for -7 – -67 window.

Table J.5: Staring model – Myponga reservoir

Input variable	In. grp.	Window end	Window start
Water quality & physical conditions			
Aluminium – soluble	+hea	-7	-37
Aluminium – total	+hea	-7	-37
NH_4		-7	-37
Chlorophyll *a*	+gen	-7	-21
Chlorophyll *b*		-7	-21
Copper – soluble	+hea	-7	-37
Copper – dissolved	+hea	-7	-37
Dissolved organic carbon		-7	-37
Iron – soluble	+hea	-7	-37
Iron – total	+hea	-7	-37
Manganese – soluble	+hea	-7	-37
Manganese – total	+hea	-7	-37
NO_2		-7	-37
NO_3	+gen	-7	-37
Odour – cold		-7	-21
Odour – hot		-7	-21
Total phosophorous	+gen	-7	-37
Water temperature	+gen	-7	-21
Turbidity	+gen	-7	-21

All inputs repeated for -7 – -67 window.

Table J.6: Staring model – Lake Soyang

Input variable	In. grp.	Window end	Window start
Water quality & physical conditions			
Inflow	+inf	-7	-14
Rainfall	+wea	-7	-14
Chlorophyll *a*	+gen	-7	-37
Conductivity		-7	-37
Dissolved inorganic P	+gen	-7	-37
Dissolved oxygen		-7	-37
NO_3	+gen	-7	-37
pH	+pH	-7	-37
Productivity		-7	-37
Secchi depth	+gen	-7	-37
Water temperature	+gen	-7	-37
Total N		-7	-37
Total P		-7	-37
Turbidity		-7	-37

All inputs repeated for -7 – -67 window.

Appendix K

Strip Mining – Error Rate Comparison

Table K.1: Effect of "data strip–mining" on model error rates. Lake Biwa.

Output	Inputs	RMSE	U1	U2	R^2	Av κ
chlorophyll *a*						
all inputs	40	**6.36**	**0.282**	**0.874**	**0.189**	0.250
first strip	18	6.48	0.289	0.891	0.142	0.176
last strip	8	7.02	0.296	0.965	0.131	**0.266**
(generic)	5	6.96	0.303	0.957	0.087	0.120
Euglena americana						
all inputs	40	1830	0.603	0.853	0.060	0.202
first strip	33	**1690**	**0.547**	**0.784**	**0.160**	**0.222**
last strip	3	2700	0.658	1.261	0.000	0.098
(generic)	6	1950	0.560	0.910	0.095	0.350
Melosira granulata						
all inputs	40	607	0.457	0.951	0.204	0.326
first strip	29	**592**	**0.444**	**0.929**	**0.231**	**0.378**
last strip	14	625	0.465	0.980	0.178	0.326
(generic)	6	609	0.461	0.950	0.195	0.400
Pediastrum biwae						
all inputs	40	555	0.566	0.950	0.126	0.252
first strip	26	515	0.533	0.887	0.197	0.242
last strip	5	**471**	**0.483**	**0.817**	**0.304**	**0.298**
(generic)	6	554	0.593	0.948	0.106	0.438

Table K.2: Effect of "data strip–mining" on model error rates. Burrinjuck Dam.

Output	Inputs	RMSE	U1	U2	R^2	Av κ
chlorophyll *a*						
all inputs	68	24.2	0.490	0.815	0.141	**0.302**
first strip	45	**18.5**	**0.373**	**0.650**	**0.391**	0.294
last strip	16	23.0	0.490	0.799	0.144	0.228
(generic)	5	19.4	0.399	0.626	0.394	0.346
chlorophyta						
all inputs	68	4040	0.525	**0.840**	0.039	**0.170**
first strip	54	4310	0.556	0.908	0.010	0.156
last strip	2	**3420**	**0.453**	0.864	**0.062**	0.084
(generic)	6	3670	0.489	0.810	0.0735	0.130
cyanophyta						
all inputs	68	75800	0.637	0.977	0.043	0.268
first strip	45	54000	**0.615**	**0.802**	**0.083**	**0.320**
last strip	1	**49900**	0.699	0.862	0.043	0.166
(generic)	6	68300	0.605	0.750	0.139	0.296
diatoms						
all inputs	68	1450	0.319	1.019	0.556	0.354
first strip	49	**1420**	**0.306**	**1.01**	**0.591**	0.374
last strip	32	1740	0.356	1.250	0.479	**0.384**
(generic)	6	1870	0.409	1.05	0.318	0.372

Table K.3: Effect of "data strip–mining" on model error rates. Darling River.

Output	Inputs	RMSE	U1	U2	R^2	Av κ
total phytoplankton						
all inputs	60	19800	0.317	0.989	0.477	0.542
first strip	16	**19100**	**0.310**	**0.979**	**0.480**	**0.582**
last strip	0	n/a	n/a	n/a	n/a	n/a
(generic)	6	22100	0.345	1.04	0.351	0.500
chlorophyta						
all inputs	60	4300	**0.343**	**1.00**	**0.500**	0.484
first strip	31	**4140**	0.352	1.03	0.475	**0.584**
last strip	8	4690	0.374	1.16	0.389	0.484
(generic)	7	4430	0.367	0.950	0.435	0.512
cyanophyta						
all inputs	60	4440	**0.361**	**1.00**	**0.442**	0.436
first strip	38	**4380**	0.381	1.04	0.389	**0.526**
last strip	4	4720	0.376	1.008	0.379	0.396
(generic)	7	4940	0.390	1.02	0.331	0.448
flagellates						
all inputs	60	**1850**	**0.318**	0.910	**0.454**	**0.520**
first strip	8	1990	0.391	**0.881**	0.303	0.410
last strip	3	2150	0.423	0.954	0.206	0.284
(generic)	7	1780	0.330	0.930	0.400	0.452

Table K.4: Effect of "data strip–mining" on model error rates. Lake Kasumigaura.

Output	Inputs	RMSE	U1	U2	R^2	Av κ
chlorophyll *a*						
all inputs	70	57.3	0.304	0.952	**0.113**	0.178
first strip	68	**56.7**	**0.303**	**0.946**	0.104	**0.188**
last strip	2	67.4	0.352	1.216	0.000	0.062
(generic)	5	54.0	0.297	0.990	0.109	0.274
Gomphosphaeria spp.						
all inputs	70	**26200**	0.532	1.037	0.148	0.165
first strip	27	27500	**0.497**	**1.00**	**0.179**	**0.240**
last strip	2	28400	0.626	1.249	0.006	0.073
(generic)	6	26100	0.588	1.08	0.0884	0.205
Microcystis aeruginosa						
all inputs	70	110000	**0.341**	**0.800**	**0.485**	0.518
first strip	53	113000	0.354	0.840	0.449	**0.544**
last strip	18	**106000**	0.368	0.878	0.433	0.398
(generic)	6	99700	0.376	0.820	0.460	0.636
Oscillatoria spp.						
all inputs	70	58100	0.623	1.103	0.034	0.206
first strip	16	48300	**0.500**	**0.966**	**0.191**	**0.428**
last strip	1	**40800**	0.758	1.034	0.000	0.003
(generic)	6	45800	0.476	0.950	0.243	0.268

Table K.5: Effect of "data strip–mining" on model error rates. Myponga Reservoir.

Output	Inputs	RMSE	U1	U2	R^2	Av κ
chlorophyll *a*						
all inputs	38	**3.51**	**0.184**	**0.868**	**0.711**	**0.728**
first strip	16	3.70	0.189	0.915	0.706	0.666
last strip	11	3.98	0.197	0.995	0.682	0.660
(generic)	5	3.92	0.212	1.00	0.614	0.632
Ankistrodesmus spp.						
all inputs	40	**2711**	0.554	0.902	0.016	0.332
first strip	37	2760	0.546	0.919	0.017	0.356
last strip	5	2886	**0.509**	**0.874**	**0.102**	**0.402**
(generic)	6	2240	0.528	0.830	0.0352	0.250
Dictyosphaerium spp.						
all inputs	40	1716	0.498	1.046	0.118	0.222
first strip	37	**1640**	**0.482**	**1.00**	**0.148**	**0.232**
last strip	35	1645	0.496	1.002	0.123	0.194
(generic)	6	1860	0.487	1.07	0.123	0.144
Scenedesmus spp.						
all inputs	40	11707	0.253	0.755	0.738	0.446
first strip	38	**11600**	**0.240**	**0.746**	**0.748**	**0.454**
last strip	9	11724	0.279	0.918	0.673	0.358
(generic)	6	10800	0.255	0.780	0.726	0.448

Table K.6: Effect of "data strip–mining" on model error rates. Lake Soyang.

Output	Inputs	RMSE	U1	U2	R^2	Av κ
chlorophyll *a*						
all inputs	28	2.35	0.375	0.923	0.252	**0.242**
first strip	19	**2.26**	**0.354**	**0.903**	**0.297**	0.188
last strip	6	2.63	0.421	1.126	0.102	0.170
(generic)	5	2.14	0.375	0.976	0.242	0.260

Appendix L

Forward Selection

Table L.1: Performance comparison – extended generic models. Lake Biwa.

Output	Inputs	RMSE	U1	U2	R^2	Av κ
Chlorophyll *a*						
generic	5	6.96	0.303	0.957	0.087	0.120
+species	8	6.37	0.282	0.877	0.170	0.210
+lag	10	6.29	0.279	0.864	0.184	0.270
+species +lag	16	**5.93**	**0.262**	**0.815**	**0.247**	**0.342**
+pH	6	6.83	0.295	0.940	0.112	0.208
+weather	7	6.63	0.287	0.912	0.146	0.240
+Si	6	7.05	0.308	0.969	0.078	0.112
+spe +lag + pH +wea	16	6.07	0.273	0.834	0.200	0.324
Euglena americana						
generic	6	1950	0.560	0.910	0.095	**0.350**
+species	9	2010	0.555	0.930	0.0976	0.302
+lag	12	1790	0.547	0.834	0.127	0.224
+species +lag	18	**1690**	0.542	**0.785**	**0.168**	0.290
+pH	7	1860	**0.540**	0.867	0.129	0.294
+weather	8	1890	0.562	0.879	0.093	0.248
+Si	7	1830	0.563	0.850	0.104	0.284
+lag +pH + wea + Si	16	1780	0.558	0.827	0.123	0.286
Melosira granulata						
generic	6	609	0.461	0.950	0.195	0.400
+species	9	655	0.492	1.03	0.132	0.416
+lag	12	601	0.463	0.942	0.198	**0.460**
+species +lag	18	**562**	**0.422**	**0.882**	**0.290**	0.426
+pH	7	617	0.452	0.967	0.211	0.384
+weather	8	634	0.495	0.994	0.128	0.392
+Si	7	615	0.477	0.964	0.161	0.390
Pediastrum biwae						
generic	6	554	0.593	0.948	0.106	**0.438**
+species	9	579	0.613	0.990	0.0686	0.346
+lag	12	**528**	0.572	**0.904**	**0.155**	0.362
+species +lag	18	568	**0.565**	0.972	0.122	0.308
+pH	7	566	0.586	0.968	0.097	0.406
+weather	8	549	0.618	0.939	0.094	0.346
+Si	7	586	0.587	1.00	0.084	0.390
+lag +wea	14	557	0.604	0.953	0.094	0.324

Table L.2: Performance comparison – extended generic models. Burrinjuck Dam.

Output	Inputs	RMSE	U1	U2	R^2	Av κ
Chlorophyll *a*						
generic	5	19.4	0.399	0.626	0.394	0.346
+species	8	21.1	0.441	0.670	0.302	0.356
+lag	10	**18.0**	**0.370**	**0.581**	**0.478**	0.364
+species +lag	16	21.9	0.452	0.695	0.262	**0.394**
+weather	14	21.2	0.419	0.649	0.353	0.322
+inflow	16	18.9	0.387	0.610	0.424	0.304
+stratification + depth	7	18.2	0.376	0.586	0.467	0.370
+lag +inf +str +dep	23	18.9	0.383	0.608	0.431	0.338
Chlorophyta						
generic	6	3670	0.489	0.810	0.0735	0.130
+species	9	**3390**	**0.463**	**0.740**	**0.145**	0.178
+lag	12	3430	0.470	0.759	0.126	**0.328**
+species +lag	18	3600	0.489	0.787	0.091	0.278
+weather	15	3720	0.516	0.794	0.047	0.140
+inflow	17	4040	0.530	0.893	0.021	0.158
+stratification + depth	8	**3390**	0.495	0.751	0.098	0.224
+lag +spe +str +dep	16	3540	0.487	0.774	0.098	0.192
Cyanophyta						
generic	6	68300	0.605	**0.750**	0.139	0.296
+species	9	72600	0.554	0.800	0.151	0.254
+lag	12	**67400**	**0.534**	**0.750**	**0.200**	0.298
+species +lag	18	68600	0.588	0.753	0.149	0.268
+weather	15	80600	0.588	0.870	0.079	0.294
+inflow	17	74500	0.613	0.847	0.058	0.206
+stratification + depth	8	67600	0.545	0.770	0.166	**0.370**
+lag +str +dep	14	68400	0.588	0.778	0.122	0.334
Diatoms						
generic	6	1870	0.409	1.05	0.318	0.372
+species	9	**1610**	**0.341**	**0.900**	**0.494**	0.384
+lag	12	1870	0.419	1.05	0.308	0.400
+species +lag	18	1650	0.362	0.921	0.448	0.356
+weather	15	1800	0.388	1.02	0.385	0.374
+inflow	17	2100	0.434	1.18	0.249	0.406
+stratification + depth	8	1820	0.390	1.02	0.366	0.394
+spe +wea +str +dep	20	1660	0.350	0.927	0.480	**0.434**

Table L.3: Performance comparison – extended generic models. Darling River.

Output	Inputs	RMSE	U1	U2	R^2	Av κ
Total phytoplankton						
generic	6	22100	0.345	1.04	0.351	0.500
+species	9	19900	0.321	0.930	0.444	0.512
+lag	12	21200	0.347	0.990	0.372	0.518
+species +lag	18	21300	0.344	0.999	0.391	0.540
+pH	7	21100	0.336	0.992	0.386	0.502
+Si	7	21300	0.330	0.995	0.404	0.530
+pH +spe +Si	12	**19300**	**0.312**	**0.901**	**0.480**	**0.568**
Chlorophyta						
generic	7	4430	**0.367**	0.950	**0.435**	0.512
+species	10	4530	0.374	0.970	0.415	0.492
+lag	14	4720	0.394	1.01	0.369	0.506
+species +lag	20	4660	0.400	0.997	0.372	**0.516**
+pH	8	**4380**	**0.367**	**0.937**	**0.435**	0.508
+Si	8	4530	0.371	0.962	0.421	0.484
Cyanophyta						
generic	7	4940	0.390	1.02	0.331	0.448
+species	10	**4570**	0.368	**0.948**	**0.396**	0.470
+lag	14	4850	0.403	1.01	0.319	0.460
+species +lag	20	4730	0.391	0.980	0.352	0.494
+pH	8	4650	**0.364**	0.965	**0.396**	**0.506**
+Si	8	4800	0.376	0.987	0.368	0.486
+spe +lag +pH +Si	19	4850	0.402	0.998	0.326	0.468
Flagellates						
generic	7	1780	0.330	0.930	0.400	0.452
+species	10	**1680**	**0.308**	**0.880**	**0.465**	0.448
+lag	14	1690	0.312	0.887	0.454	**0.530**
+species +lag	20	1720	0.319	0.904	0.438	0.528
+pH	8	1900	0.343	0.998	0.357	0.458
+Si	8	1870	0.342	0.977	0.360	0.436

Table L.4: Performance comparison – extended generic models. Lake Kasumigaura.

Output	Inputs	RMSE	U1	U2	R^2	Av κ
Chlorophyll *a*						
generic	5	54.0	0.297	0.990	0.109	0.274
+species	8	54.3	0.289	0.984	0.129	0.210
+lag	10	48.9	0.272	0.896	0.199	0.292
+species +lag	16	49.6	0.272	0.896	0.166	0.304
+pH	6	53.2	0.289	0.953	0.129	0.284
+zoo	11	47.8	0.264	0.863	0.257	**0.380**
+weather	10	49.3	0.277	0.904	0.175	0.294
+Si	6	50.7	0.280	0.912	0.171	0.332
+lag +zoo +Si +wea	22	**46.9**	**0.255**	**0.824**	**0.272**	0.306
Gomphosphaeria spp.						
generic	6	26100	0.588	1.08	0.0884	0.205
+species	9	25100	0.518	1.04	0.179	0.198
+lag	12	26800	0.559	1.11	0.103	0.260
+species +lag	18	26500	0.544	1.10	0.126	0.220
+pH	7	27900	0.571	1.16	0.079	0.135
+zoo	12	25800	0.534	1.04	0.160	0.205
+weather	11	28200	0.558	1.14	0.094	0.150
+Si	7	**22400**	**0.518**	**0.883**	**0.230**	0.240
+zoo +Si +spe	15	23700	0.523	0.919	0.194	**0.268**
Microcystis aeruginosa						
generic	6	99700	0.376	0.820	0.460	**0.636**
+species	9	99200	0.352	0.810	0.485	0.632
+lag	12	**83700**	**0.302**	**0.690**	**0.622**	0.572
+species +lag	18	88700	0.319	0.727	0.579	0.562
+pH	7	99600	0.351	0.789	0.493	0.596
+zoo	12	105000	0.360	0.839	0.456	0.528
+weather	11	92700	0.331	0.746	0.549	0.554
+Si	7	101000	0.354	0.802	0.484	0.542
+lag +wea	17	87500	0.309	0.704	0.599	0.536
Oscillatoria spp.						
generic	6	45800	**0.476**	0.950	**0.243**	0.268
+species	9	**45200**	0.502	**0.940**	0.222	**0.348**
+lag	12	55000	0.527	1.14	0.137	0.282
+species +lag	18	52100	0.542	1.08	0.117	0.254
+pH	7	48500	0.509	0.982	0.197	0.300
+zoo	12	49100	0.512	0.988	0.188	0.298
+weather	11	49500	0.494	1.01	0.189	0.304
+Si	7	49800	0.515	1.01	0.185	0.336

Table L.5: Performance comparison – extended generic models. Myponga Reservoir.

Output	Inputs	RMSE	U1	U2	R^2	Av κ
Chlorophyll *a*						
generic	5	3.92	0.212	1.00	0.614	0.632
+lag	10	3.53	0.190	0.903	0.685	0.646
+heavy metals	13	3.76	0.197	0.961	0.661	0.632
+lag +hea	18	**3.52**	**0.187**	**0.901**	**0.695**	**0.654**
Ankistrodesmus spp.						
generic	6	**2240**	0.528	0.830	0.0352	0.250
+lag	12	2340	0.543	0.865	0.023	**0.342**
+heavy metals	14	2250	**0.509**	**0.829**	**0.050**	0.306
Dictyosphaerium spp.						
generic	6	1860	0.487	1.07	0.123	0.144
+lag	12	1720	0.472	0.992	0.158	0.210
+heavy metals	14	1820	0.469	1.04	0.158	0.164
+lag +hea	20	**1650**	**0.448**	**0.944**	**0.217**	**0.226**
Scenedesmus spp.						
generic	6	10800	0.255	0.780	0.726	0.448
+lag	12	10440	0.249	0.749	0.745	**0.542**
+heavy metals	14	10700	0.247	0.767	0.737	0.446
+lag +hea	20	**10300**	**0.238**	**0.734**	**0.756**	0.476

Table L.6: Comparison of starting and final model error rates. Lake Soyang.

Output	Inputs	RMSE	U1	U2	R^2	Av κ
Chlorophyll *a*						
generic	5	2.14	0.375	0.976	0.242	0.260
+lag	10	**2.09**	0.371	0.954	0.261	0.290
+pH	6	**2.09**	0.367	**0.946**	**0.274**	0.278
+weather	6	2.24	**0.365**	0.955	0.273	0.172
+inflow	6	2.30	0.376	0.986	0.238	0.188
+lag +pH	11	2.12	0.379	0.963	0.236	**0.294**

Appendix M

Specific Model Predictions

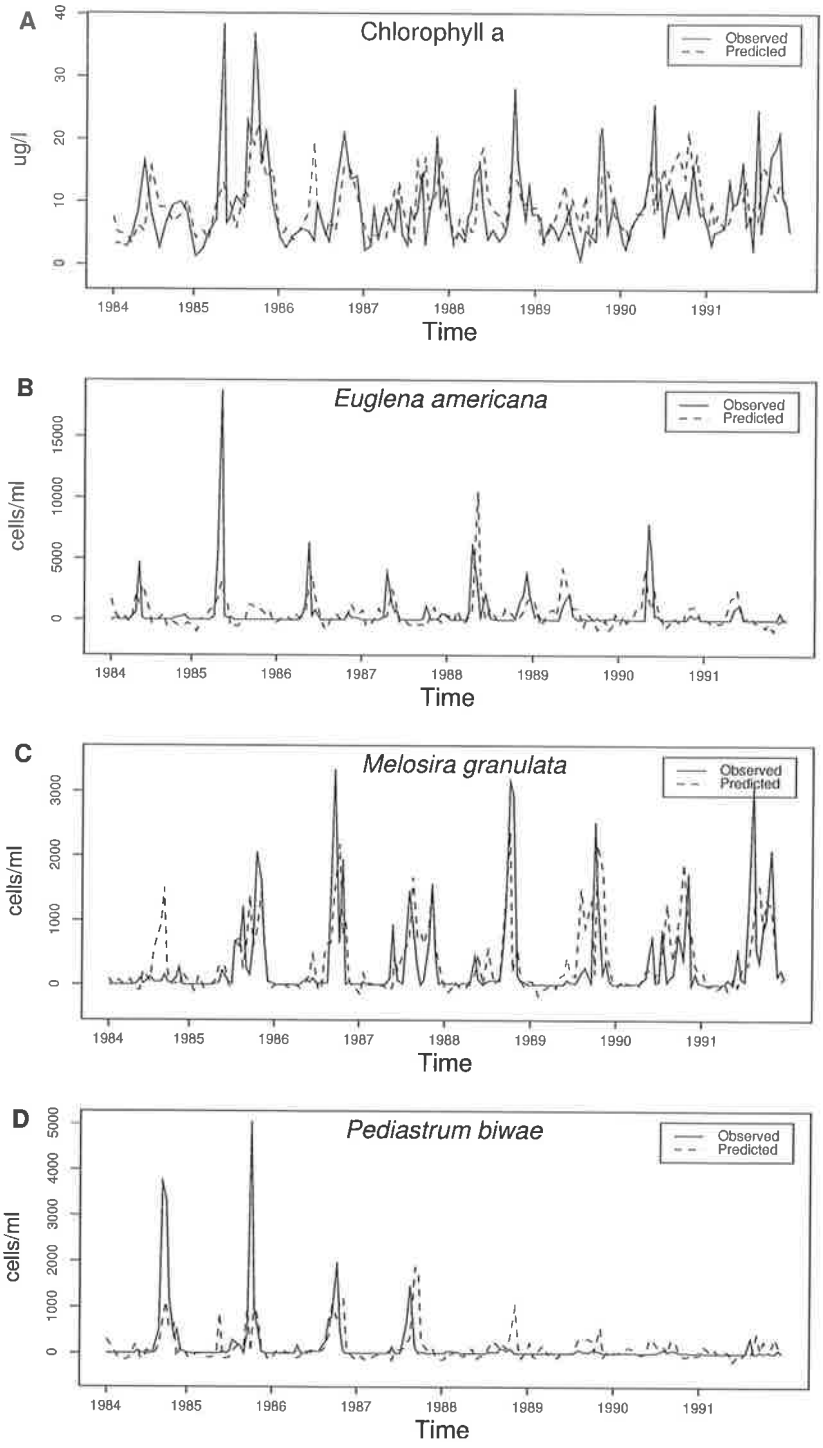

Figure M.1: Lake Biwa. Specific input layer. **A** Chlorophyll *a*. **B** *Euglena americana*. **C** *Melosira granulata*. **D** *Pediastrum biwa*.

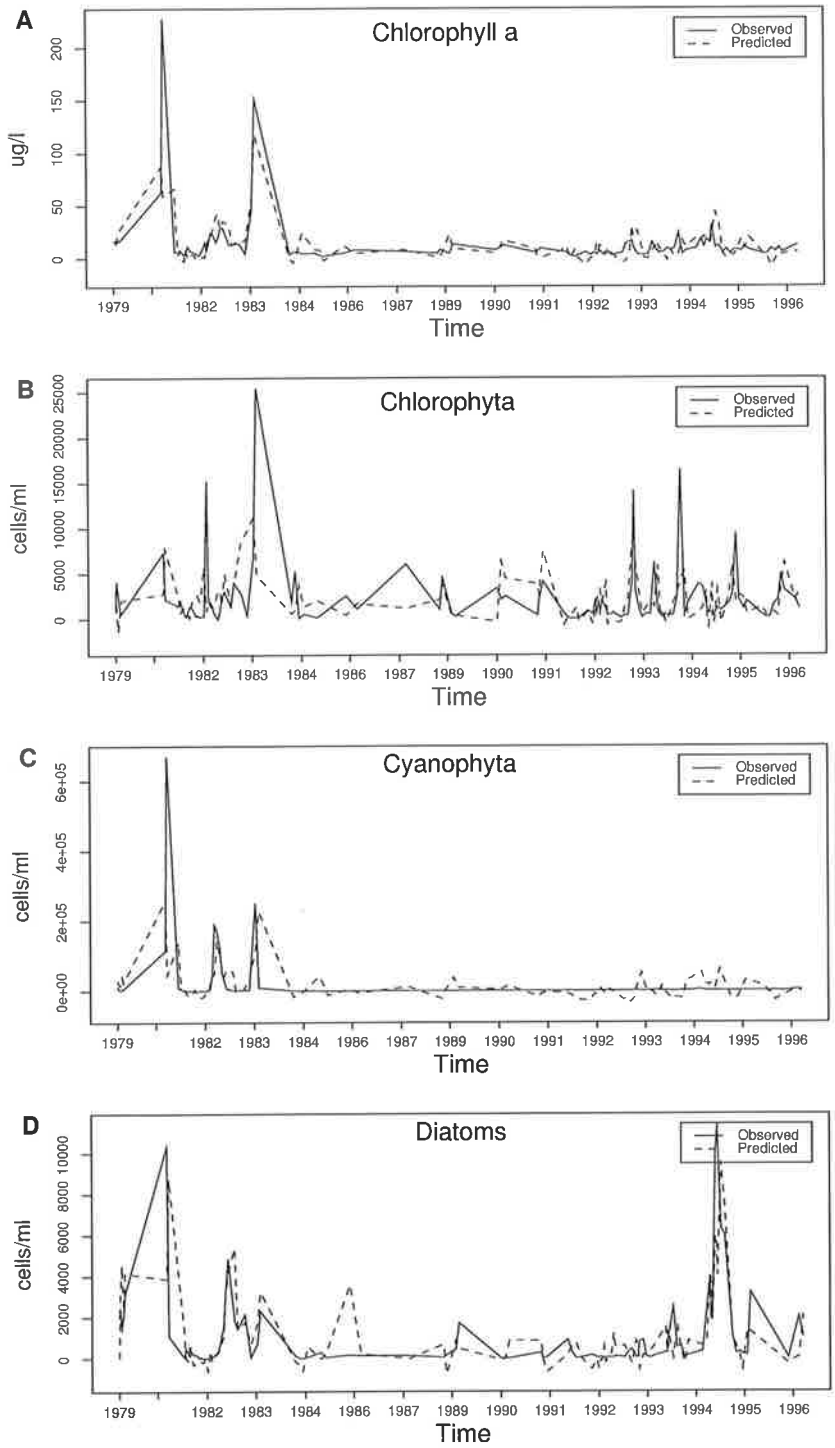

Figure M.2: Burrinjuck Dam. Specific input layer. **A** Chlorophyll *a*. **B** Chlorophyta. **C** Cyanophyta. **D** Diatoms.

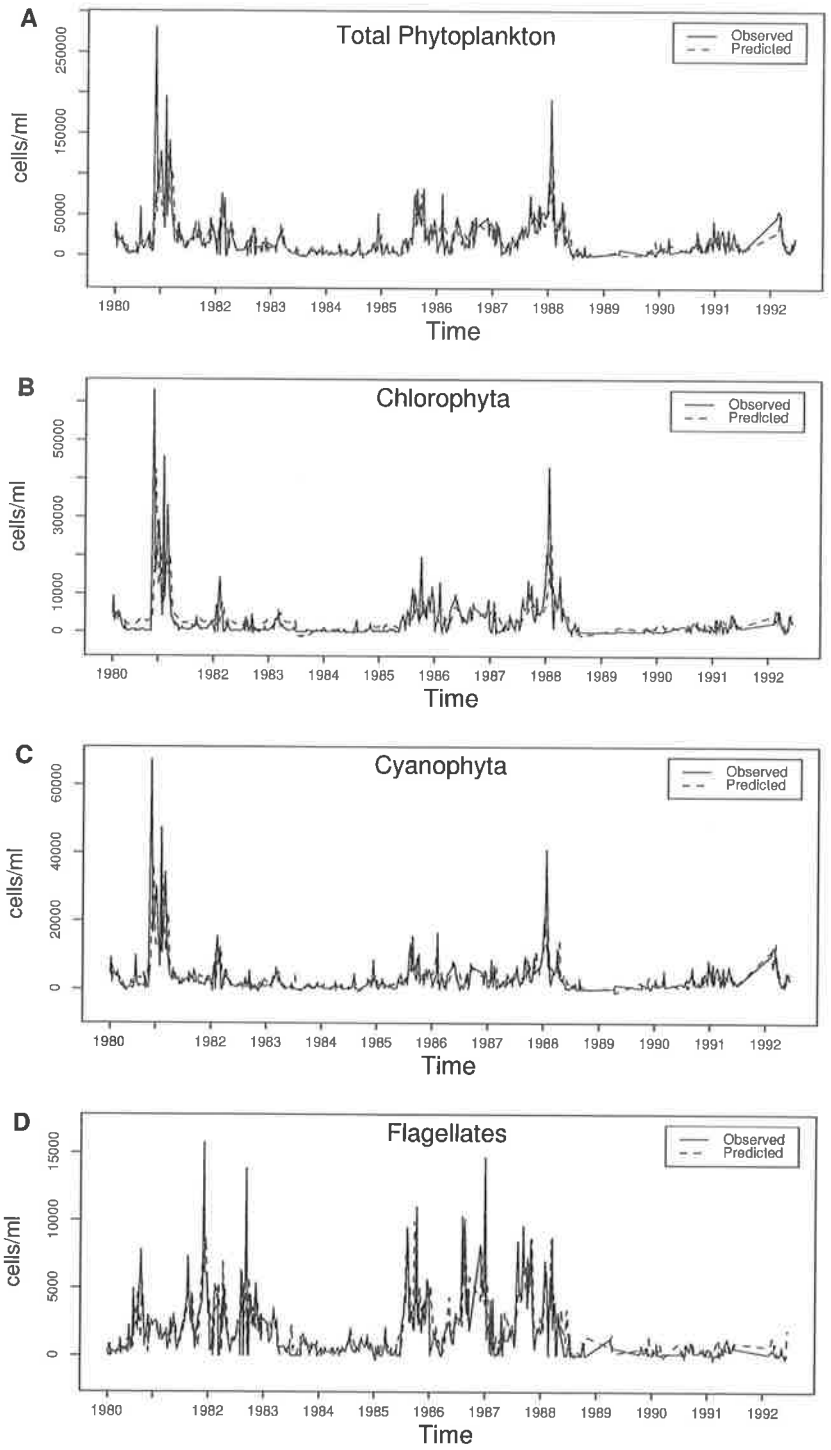

Figure M.3: Darling River. Specific input layer. **A** Total phytoplankton. **B** Chlorophyta. **C** Cyanophyta. **D** Flagellates.

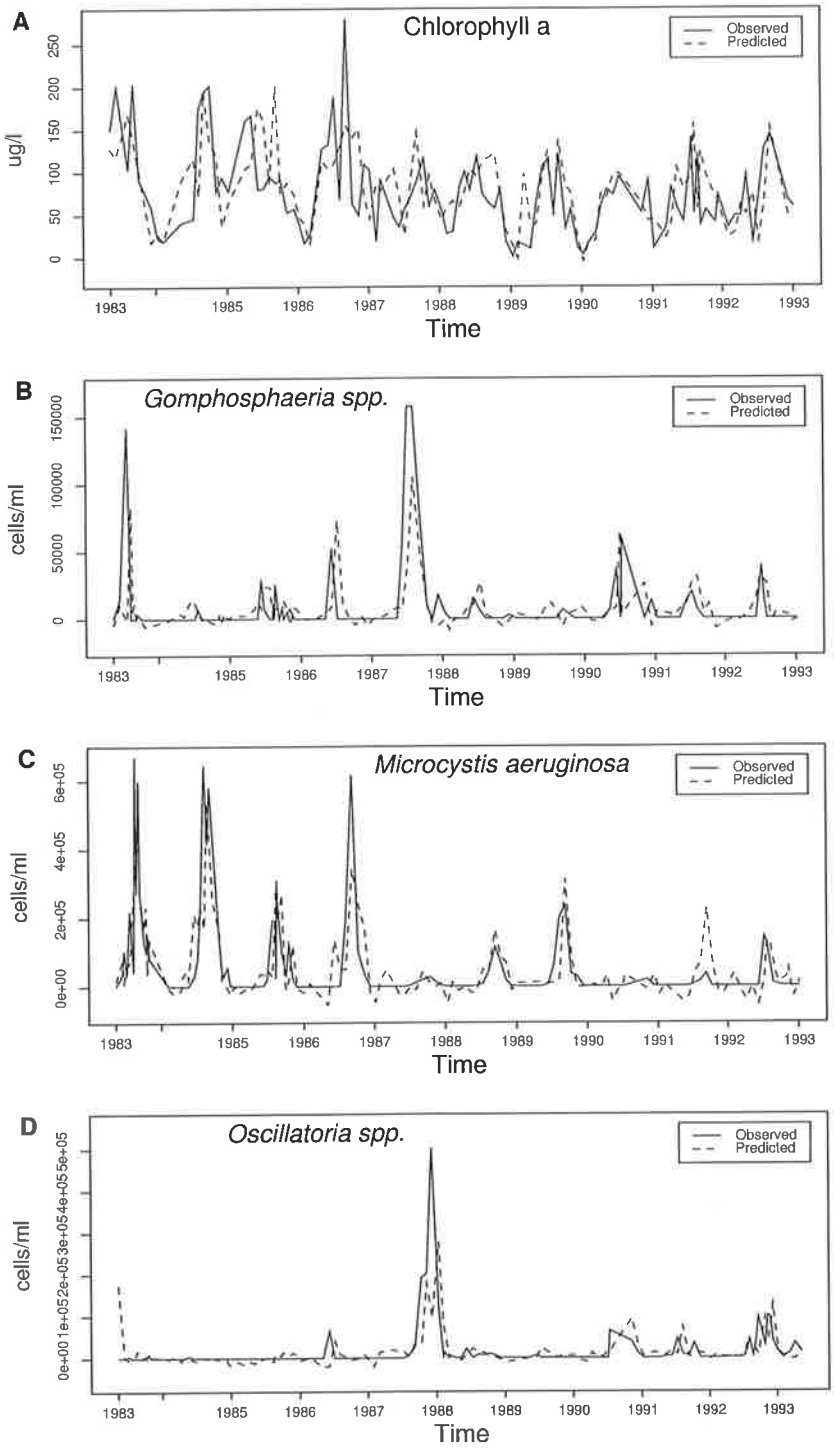

Figure M.4: Lake Kasumigaura. Specific input layer. **A** Chlorophyll *a*. **B** *Gomphosphaeria spp.* **C** *Microcystis aeruginosa*. **D** *Oscillatoria spp.*

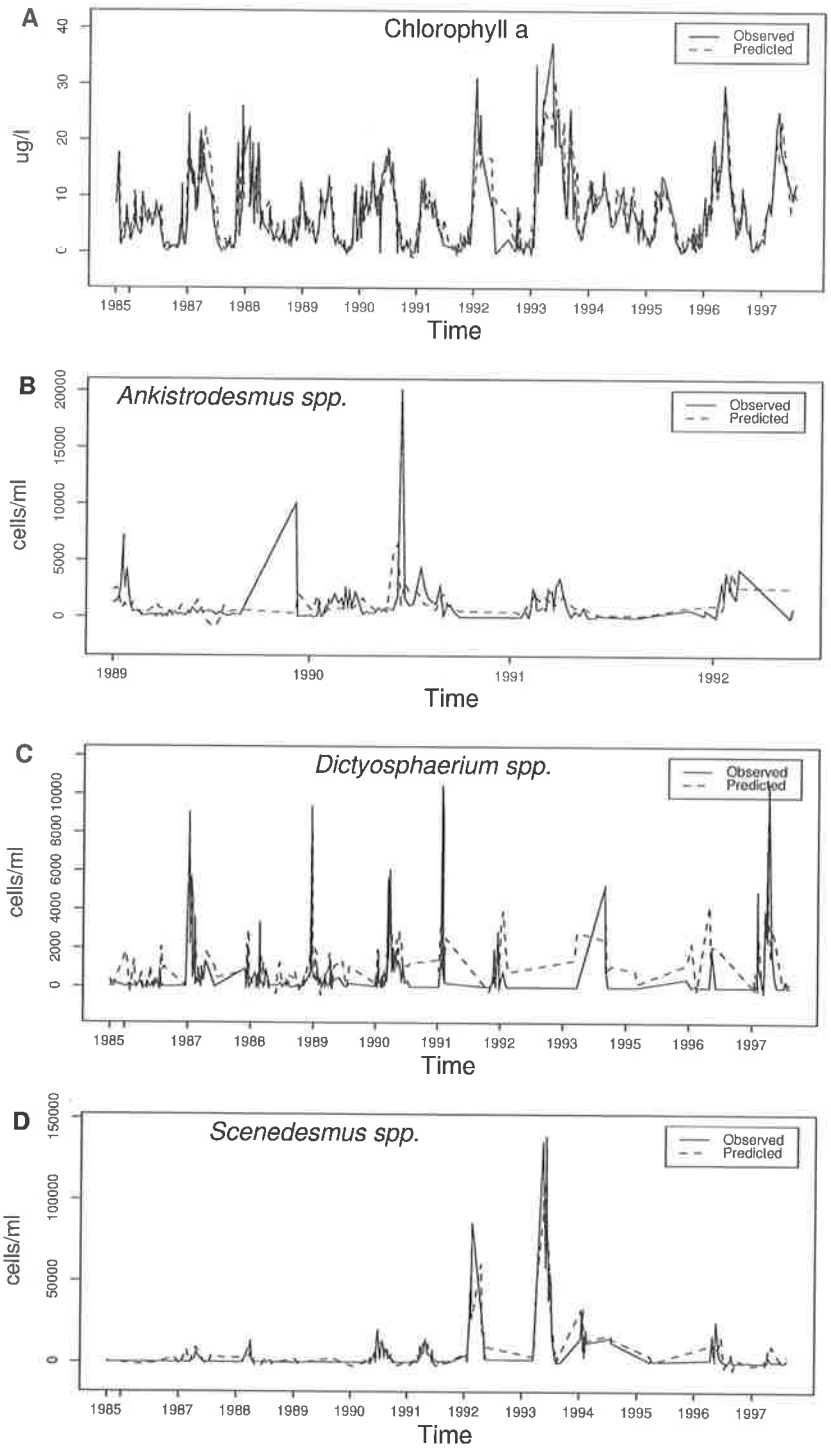

Figure M.5: Myponga Reservoir. Specific input layer. **A** Chlorphyll *a*. **B** *Ankistrodesmus spp.* **C** *Dictyosphaerium spp.* **D** *Scenedesmus spp.*

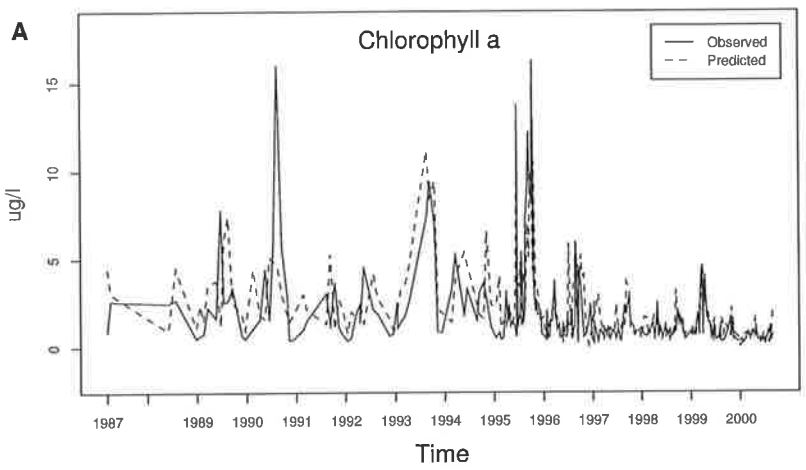

Figure M.6: Lake Soyang. Specific input layer. Chlorophyll *a*

Bibliography

Abu-Mostafa, Y. S. (1989). The Vapnik-Chervonenkis dimension: information versus complexity in learning. *Neural Computation 1*, 312–317.

Ackley, D., G. Hinton, and T. Sejnowski (1985). A Learning Algorithm for Boltzmann Machines. *Cognitive Science 9*, 147–169.

Adeli, H. and S.-L. Hung (1995). *Machine Learning, Neural Networks, Genetic Algorithms and Fuzzy Systems*. New York: John Wiley and Sons Inc.

Akaike, H. (1969). Fitting autoregressive models for prediction. *Annals of Institute of Statistical Mathematics 21*, 10–172.

Alpsan, D., M. Towsey, O. Ozdamar, A. C. Tsoi, and D. N. Ghista (1995). Efficacy of Modified Backpropagation and Optimisation Methods on a Real-world Medical Problem. *Neural Networks 8*(6), 945 – 962.

Amit, D. J. (1989). *Modeling Brain Function. The World of Attractor Neural Networks*. Cambridge, UK: University Press.

Andersen, T., M. Rimer, and T. Martinez (2001). Optimal Artificial Neural Network Architecture Selection for Bagging. In *Proceedings, 2001 INNS – IEEE International Joint Conference on Neural Networks*, Washington D.C., pp. 790–795. IEEE Press.

Aoki, I., T. Komatsu, and K. Hwang (1999). Prediction of response of zooplankton biomass to climatic and oceanic changes. *Ecological Modelling 120*, 261–270.

Armstrong, J. S. and F. Collopy (1992). Error measures for generalizing about forecasting methods: Empirical comparisons. *International Journal of Forecasting 8*, 69–80.

Aussem, A. and D. Hill (1999). Wedding connectionist and algorithmic modelling towards forecasting *Caulerpa taxifolia* development in the north-western mediterranean sea. *Ecological modelling 120*, 225–236.

Azevedo, S., W. Carmichael, E. Jochimsen, K. Rinehart, S. Lau, G. Shaw, and G. Eaglesham (2002). Human intoxication by microcystins during renal dialysis treatment in Caruaru – Brazil. *Toxicology 181–182*, 441–446.

Ball, G. R., J. Benton, D. Palmer-Brown, J. Fuhrer, L. Skärby, B. S. Gimeno, and G. Mills (1998). Identifying factors which modify the effects of ambient ozone on white clover (*Trifolium repens*) in Europe. *Environmental Pollution 103*, 7–16.

Barciela, R. M., E. García, and E. Fernández (1999). Modelling primary production in a coastal embayment affected by upwelling using dynamic ecosystem models and artficial neural networks. *Ecological Modelling 120*, 199–211.

Baxt, W. G. and H. White (1995). Bootstrapping confidence intervals for clinical input variable effects in a network trained to identify the presence of acute myocardial infarction. *Neural Computation 7*, 624–638.

Bellman, R. (1961). *Adaptive Control Processes: A Guided Tour*. Princeton University Press.

Benndorf, J. and F. Recknagel (1982). Problems of application of the ecological model SALMO to lakes and reservoirs having various trophic states. *Ecological Modelling 17*, 129–145.

Bertsekas, D. P. and J. N. Tsitsiklis (1996). *Neuro-Dynamic Programming*. Belmont MA: Athena Scientific.

Bobbin, J. and F. Recknagel (2001). Knowledge discovery for prediction and explanation of blue–green algal dynamics in lakes by evolutionary algorithms. *Ecological Modelling 146*, 253–262.

Bobbin, J. and F. Recknagel (2003). Predictive rules for phytopalnkton dynamics in freshwater lakes discovered by evolutionary algorithms. In F. Recknagel (Ed.), *Ecological Informatics. Understanding Ecology by Biologically–Inspired Computation*, pp. 291–311. Berlin: Springer–Verlag.

Braun, H. and M. Riedmiller (1992). Rprop: A fast adaptive learning algorithm. In *Proc. of the Int. Symposium on Computer and Information Science VII*.

Breiman, L. (1994, September). Bagging predictors. Technical Report 421, Department of Statistics, University of California, California.

Breiman, L. (1996a). Bagging predictors. *Machine Learning 24*(2), 123–140.

Breiman, L. (1996b). The heuristics of instability in model selection. *Annals of Statistics 2*, 2350–2383.

Brosse, S., J. L. Giraudel, and S. Lek (2001). Utilisation of non-supervised neural networks and principal component analysis to study fish assemblages. *Ecological Modelling 146*(1–3), 131–142.

Burch, M. (1993). The development of an alert levels and response framework for the management of blue-green algal blooms. In *Blue-green algal blooms*

- *new developments in research and management*, Adelaide, South Australia. Australian Centre for Water Quality Research and the University of Adelaide.

Burch, M. D. (1990). Algicidal control of algal blooms. In *Blue–green algae in Drinking and Receiving Waters*, Sydney, Australia, pp. 21–29.

Burch, M. D. and B. C. Nicholson (2000). Minimisation of cyanobacterial toxins in drinking water by reservoir management and water treatment. In *Proceedings of the International Water Resources Association's 10th World Water Congress. Accessed online from www.iwra.siu.edu*, Melbourne Australia 12–16 3/2000.

Burns, F. L. (1994). Case study: blue-green algal control in Australia by year-round automatic aeration. *Lake and Reservoir Management 10*(1), 61–67.

Cannon, A. and P. Whitfield (2002). Downscaling recent streamflow conditions in British Columbia, Canada using ensemble neural network models. *Journal of Hydrology 259*, 136–151.

Cheng, B. and D. M. Titterington (1994). Neural networks: a review from a statistical perspective. *Statistical Science 9*, 2–54.

Chon, T.-S., Y. S. Park, K. H. Moon, and E. Y. Cha (1996). Patternizing communities by using an artificial neural network. *Ecological Modelling 90*, 69–78.

Colasanti, R. (1991). Discussions of the possible use of neuronal network algorithms in ecological modelling. *Binary 3*, 13–15.

Connors, J., D. Martin, and L. Atlas (1994). Recurrent neural networks and robust time series prediction. *IEEEE Transactions Neural Networks 5*, 240–254.

Cortez, P., F. S. Allegro, M. Rocha, and J. Neves (2002, September). Real-Time Forecasting by Bio-Inspired Models. In M. Hamza (Ed.), *Proceedings of the Second IASTED International Conference on Artificial Intelligence and Applications (AIA 2002)*, Malaga, Spain, pp. 52–57. IASTED ACTA Press.

de la Maza, M. (1991). Splitnet. dynamically adjusting the number of hidden units in a neural network. In T. Kohonen, O. Makisara, O. Simula, and J. Kangas (Eds.), *Artificial Neural Networks*, Volume 1, pp. 647–651. Amsterdam, The Netherlands: Elsevier Science Publishers B. V.

Dillon, P. and F. Rigler (1974). The phosphorus-chlorophyll relationship in lakes. *Limnol. Oceanogr. 19*, 767–773.

Dokulil, M. T. and K. Teubner (2000). Cyanobacterial dominance in lakes. *Hydrobiologia 438*, 1–12.

Efron, B. and R. Tibshirani (1993). *An Introduction to the Bootstrap*. New York: Chapman and Hall.

Efron, B. and R. Tibshirani (1997). Cross–validation and the bootstrap: estimating the error rate of a prediction rule. *Journal of the American Statistical Association 92*, 548–560.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science 14*, 179–211.

Embrechts, M. A., F. Arciniegas, M. Ozdemir, C. M. Breneman, K. Bennett, and L. Lockwood (2001). Bagging neural network sensitivity analysis for feature reduction for in-silico drug design. In *Proceedings, 2001 INNS – IEEE International Joint Conference on Neural Networks*, Volume 4, Washington D.C., pp. 2478–2482. IEEE Press.

Everall, N. C. and D. R. Lees (1996). The use of barley-straw to control general and blue-green algal growth in a Derbyshire reservoir. *Water Resources 30*(2), 269–276.

Fahlman, S. E. (1988). Faster–learning variations on back-propagation: An empirical study. In T. J. Sejnowski, G. E. Hinton, and D. S. Touretzky (Eds.), *1988 Connectionist Models Summer School*. San Mateo, CA: Morgan Kaufmann.

Fahlman, S. E. and C. Lebiere (1990). The Cascade–Correlation learning architecture. In D. S. Touretzky (Ed.), *Advances in Neural Information Processing Systems*, Volume 2, pp. 524–532. San Francisco, CA: Morgan Kaufman.

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

Finnoff, W., F. Hergert, and H. G. Zimmermann (1993). Improving model selection by nonconvergent methods. *Neural Networks 6*, 771–783.

Flexer, A. (1995). Connectionists and statisticians, friends or foes? Technical Report OEFAI-95-06, Austrian Research Institute for Artificial Intelligence.

Foody, G. M. (1999). Applications of the self–organising feature map neural network in community data analysis. *Ecological Modelling 120*(2–3), 97–108.

Forsberg, C. and S.-O. Ryding (1980). Eutrophication parameters and trophic state indices in 30 Swedish waste–receiving lakes. *Arch. Hydrobiol. 89*, 189–207.

Freitas de Magalhẽs, V., R. Soares, and S. Azevedo (2001). Microcystin contamination in fish from the Jacarepaguá Lagoon (Rio de Janeiro, Brazil): ecological implication and human health risk. *Toxicom 39*, 1077–1085.

French, M. and F. Recknagel (1994). Modelling algal blooms in freshwaters using artificial neural networks. In P. Zannetti (Ed.), *Computer Techniques in Environmental Studies V. Vol. 2: Environmental Systems*, pp. 87–94. Boston: Computer Mechanics Publications.

Gallant, S. I. (1993). *Neural network learning and expert systems*. Cambridge Massachussetts: MIT Press.

Ganf, G. G. and R. L. Oliver (1982). Vertical separation of light and available nutrients as a factor causing replacement of green algae by blue–green algae in the plankton of a stratified lake. *Journal of Ecology 70*, 829–844.

Garson, G. D. (1991). Interpreting neural–network connection weights. *Artif. Intell. Expert 6*, 47–51.

Geman, S., E. Bienenstock, and R. Doursat (1992). Neural networks and the bias/variance dilemma. *Neural Computation 4*, 1–58.

Gori, M. and A. Tesi (1992). On the problem of local minima in backpropagation. *IEEE Transactions on pattern analysis and machine intelligence 14*, 76–86.

Government of South Australia (1962). *The Official Opening of Myponga Reservoir – Souvenir*. The Government of South Australia.

Gragnani, A., M. Scheffer, and S. Rinaldi (1999). Top-down control of cyanobacteria: A theoretical analysis. *The American Naturalist 153*(1), 59–72.

Györgyi, G. (1990). Inference of a rule by a neural network with thermal noise. *Phys. Rev. Lett. 64*, 2957–2960.

Hall, K., T. Murphy, M. Mawhinney, and K. Ashley (1995). Iron treatment for eutrophication control in Black Lake, British Colombia. *Lake Reservoir Manage. 9*(1), 114–117.

Harris, G. P. (1986). *Phytoplankton Ecology : structure, function and fluctuation*. London: Chapman and Hall.

Harvey, F. L. (1992). Destratification by mechanical mixing and the resultant effect on water quality in the myponga reservoir south australia, 1991–1992. Technical report, Australian Centre for Water Quality Research.

Hassibi, B. and D. G. Stork (1993). Second order derivatives for network pruning: Optimal brain surgeon. In S. J. Hanson, J. D. Cowan, and G. C. L. (Eds.), *Advances in Neural Information Processing Systems*, Volume 5, pp. 164–171. San Mateo, CA: Morgan Kaufman.

Hassoun, M. H. (1995). *Fundamentals of Artificial Neural Networks*. Cambridge, Massachusetts: MIT Press.

Haugh, L. and G. Box (1977). Identification of dynamic regression (distributed lag) models connecting two time series. *Journal of the American Statistical Association 72*(397), 121–130.

Haykin, S. S. (1994). *Neural Networks: a comprehensive foundation*. New York: McMillan.

Hebb, D. O. (1949). *The Organisation of Behaviour*. New York: Wiley.

Hecht-Nielsen, R. (1990). *Neurocomputing*. Addison-Wesley Publishing Company Inc.

Hinton, G. E. (1992). Connectionist learning procedures. In P. Mehra and B. W. Wah (Eds.), *Artificial Neural Networks: concepts and theory*. Los Alamitos, California: IEEE Computer Society Press.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences, USA 79*, 2554–2558.

Hornik, K. (1993). Some new results on neural network approximation. *Neural Networks 6*, 1069–1072.

Hosper, S. H. (1998). Stable states, buffers and switches: an ecosystem approach to the restoration and management of shallow lakes in the Netherlands. *Wat. Sci. Tech. 37*(3), 151–164.

Hwarng, H. and H. Ang (2001). A simple neural network for ARMA(p,q) time series. *Omega 29*, 319–333.

Jeong, K.-S., F. Recknagel, and G.-J. Joo (2003). Prediction and Elucidation of Population Dynamics of the Blue–green Algae *Microcystis aeruginosa* and the Diatom *Stephanodiscus hantzschii* in the Nakdong River–Reservoir System (South Korea) by a Recurrent Artificial Neural Network. In F. Recknagel (Ed.), *Ecological Informatics. Understanding Ecology by Biologically–Inspired Computation.*, pp. 195–213. Berlin: Springer–Verlag.

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

Jørgensen, S. E. (1999). State–of–the–art of Ecological Modelling with emphasis on development of Structural Dynamic Models. *Ecological Modelling 120*, 75–96.

Karul, C., S. Soyupak, A. Cilesiz, N. Akbay, and E. Germen (2000). Case studies on the use of neural networks in eutrophication modelling. *Ecological Modelling 134*, 145–152.

Kim, B., K. Choi, C. Kim, U.-H. Lee, and Y.-H. Kim (2000). Effects of the summer monsoon on the distribution and loading of organic carbon in a deep reservoir, Lake Soyang, Korea. *Water Resources 34*(14), 3495–3504.

Kim, B., J.-O. Kim, M.-S. Jun, and S.-J. Hwang (1999). Seasonal Dynamics of Phytoplankton and Zooplankton Community in Lake Soyang. *Korean Journal of Limnology 32*(2), 127–134.

Kim, B., J.-H. Park, G. Hwang, and C. Kwangsoon (1997). Eutrophication of Large Freshwater Ecosystems in Korea. *Korean Journal of Limnology 30*(Supplement), 512–517.

Kohonen, T. (1982). Self–organized formation of topologically correct feature maps. *Biol. Cybern. 43*, 59–69.

Krogh, A. and J. A. Hertz (1992). A simple weight decay can improve generalization. In J. E. Moody, S. J. Hanson, and R. P. Lippmann (Eds.), *Advances in Neural Information Processing Systems*, Volume 4, pp. 950–957. San Mateo, CA: Morgan Kaufman.

Kung, S. Y. (1993). *Digital Neural Networks*. New Jersey: PTR Prentice Hall Inc.

Lae, R., S. Lek, and J. Moreau (1999). Predicting fish yield of african lakes using neural networks. *Ecological Modelling 120*, 325–335.

Lawrence, S. and C. L. Giles (2000). Overfitting and Neural Networks. Conjugate Gradient and Backpropagation. In *International Joint Conference on Neural Networks, Como, Italy, July 24–27*, Los Alamitos, CA, pp. 114–119. IEEE Computer Society.

LeCun, Y. L., J. S. Denker, and S. A. Solla (1990). Optimal brain damage. In D. S. Touretzky (Ed.), *Advances in Neural Information Processing Systems*, Volume 2, pp. 598–605. San Francisco, CA: Morgan Kaufman.

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

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

Lek, S., J. L. Giraudel, and J. F. Guégan (2000). Neuronal networks: Algorithms and architectures for ecologists and evolutionary ecologists. In S. Lek and J. Guégan (Eds.), *Artificial Neuronal Networks. Application to Ecology and Evolution*. Berlin: Springer.

Lek, S. and J. F. Guégan (1999). Artificial neural networks as a tool in ecological modelling, an introduction. *Ecological Modelling 120*, 65–73.

Levin, A. U., T. K. Leen, and J. E. Moody (1994). Fast pruning using principal components. In J. D. Cowan, G. Tesauro, and J. Alspector (Eds.), *Advances in Neural Information Processing Systems*, Volume 6. San Mateo, CA: Morgan Kaufman.

Levine, E., D. Kimes, and V. Sigillito (1996). Classifying soil structure using neural networks. *Ecological Modelling 90*, 39–52.

Lewis, D. M., J. A. Elliott, M. F. Lambert, and C. S. Reynolds (2002). The simulation of an Australian reservoir using a phytoplankton community model: PROTECH. *Ecological Modelling 150*, 107–116.

Maier, H. and G. Dandy (2001). Neural network based modelling of environmental variables: A systematic approach. *Mathematical and Computer Modelling 33*, 669–682.

Maier, H., T. Sayed, and B. Lence (2000). Forecasting Cyanobacterial Concentrations Using B–Spline Networks. *Journal of Computing in Civil Engineering 14*(3), 183–189.

Maier, H., T. Sayed, and B. Lence (2001). Forecasting cyanobacterium *Anabaena* spp. in the River Murray, South Australia, using B–spline neurofuzzy models. *Ecological Modelling 146*(1–3), 85–96.

Maier, H. R. and G. C. Dandy (1997). Determining inputs for neural network models of multivariate time series. *Microcomputers in Civil Engineering 12*, 353–368.

Maier, H. R. and G. C. Dandy (2000). Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications. *Environmental Modelling and Software 15*, 101–124.

Maier, H. R., G. C. Dandy, and M. D. Burch (1998). Use of Artificial Neural Networks for Modelling Cyanobacteria *Anabaena spp.* in the River Murray, South Australia. *Ecological Modelling 105*, 257–272.

Masters, T. (1994). *Practical Neural Network Recipes in C++*. Boston: Academic Press Inc.

McAuliffe, T. F. and R. S. Rosich (1989). Review of artificial destratification of water storages in Australia. Technical report, Urban Water Research Association.

McClelland, J. L. and D. E. Rumelhart (1988). *Explorations in parallel distributed processing*. Cambridge, Mass: MIT Press.

McCulloch, W. S. and W. Pitts (1943). A logical calculus of ideas imminent in nervous activity. *Bulletin of Mathematical Biophysics 5*, 115–133.

McGill, R., J. W. Tukey, and W. A. Larsen (1978). Variations of box plots. *The American Statistician 32*(1), 12–16.

McKay, J. and A. Moeller (2001). Is risk associated with drinking water in Australia of significant concern to justify mandatory regulation. *Environmental Management 28*(4), 469–481.

Minsky, M. and S. Papert (1969). *Perceptrons; an introduction to computational geometry*. Cambridge, Mass.: MIT Press.

Møller, M. F. (1993). A scaled conjugate gradient algorithm for fast supervised learning. *Neural Networks 6*, 525–533.

Moody, J. E. (1991). Note on generalization, regularization, and architecture selection in nonlinear learning systems. In B. H. Juang, S. Y. Kung, and C. A. Kamm (Eds.), *Workshop on Neural Networks for Signal Processing*, Princeton, N.J., pp. 1–10. IEEE Signal Processing Society in cooperation with the IEEE Neural Networks Council.

Morgan, N. and H. Bourlard (1990). Generalization and parameter estimation in feedforward nets: Some Experiments. In D. S. Touretzky (Ed.), *Advances in Neural Information Processing Systems*, Volume 2, pp. 630–637. San Francisco, CA: Morgan Kaufman.

Mosteller, F. and J. W. Tukey (1977). *Data Analysis and Regression – a second course in statistics*. Reading, MA: Addison–Wesley.

Mur, L. R., O. M. Skulberg, and H. Utkilen (1999). Cyanobacteria in the environment. In I. Chorus and J. Bartram (Eds.), *Toxic Cyanobacteria in Water. A Guide to Public Health Consequences and Their Supplies. Who Series in Environmental Management*, pp. 15–40. London: Routledge.

MySQL AB (2002). MySQL Database Server Version 3.23.47. http://www.mysql.com.

Nejad, A. F. and T. D. Gedeon (1995). Bidirectional neural networks reduce generalisation error. *Lecture notes in Computer Science 930*, 543–550.

Nowlan, S. J. and G. E. Hinton (1992). Simplifying neural networks by soft weight–sharing. *Neural Computation 4*(4), 473–493.

Olden, J. D. (2000). An artificial neural network approach for studying phytoplankton succession. *Hydrobiologia 436*, 131–143.

Olden, J. D. and D. A. Jackson (2000). Torturing data for the sake of generality: How valid are our regression models? *Ecoscience 4*(7), 501–510.

Olden, J. D. and J. A. Jackson (2002). Illuminating the "black box": a randomization approach for understanding variable contributions in artificial neural networks. *Ecological Modelling 154*, 135–150.

Otsuki, A., M. Aizaki, and T. Kawai (1987). Long–term variations of three types of phosphorus concentrations in highly eutrophic shallow lake kasumigaura, with special reference to dissolved organic phosphorus. *Jpn. J. Limnol. 48*, S1–S11.

Özesmi, S. L. and U. Özesmi (1999). An artificial neural network approach to spatial habitat modelling with interspecific interaction. *Ecological Modelling 116*, 15–31.

Pandya, S. A. and R. B. Macy (1996). *Pattern Recognition with neural networks in C++*. Boca Raton, Florida: CRC Press.

Parker, D. B. (1982). Learning logic. Technical Report S81–64, File 1, Office of Technology & Licencing, Stanford University.

Paruelo, J. M. and F. Tomasel (1997). Prediction of functional characteristics of ecosystems: a comparison of artificial neural networks and regression models. *Ecological Modelling 98*, 173–186.

Pineda, F. J. (1987). Generalisation of back–propagation to recurrent neural networks. *Physical Review Letter 59*(19), 2229–2232.

Prechelt, L. (1998). Automatic Early Stopping Using Cross Validation: Quantifying the Criteria. *Neural Networks 11*(4), 761–767.

Recknagel, F. (2002). Simulation of aquatic food web and species interactions by adaptive agents embodied with evolutionary computation: a conceptual framework. *Ecological Modelling 170*, 291–302.

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

Recknagel, F., J. Bobbin, P. Whigham, and H. Wilson (2002). Comparative application of artificial neural networks and genetic algorithms for multivariate time–series modelling of algal blooms in freshwater lakes. *Journal of Hydroinformatics 4*(2), 125–133.

Recknagel, F., M. French, P. Harkonen, and K. Yabunaka (1997). Artificial neural network approach for modelling and prediction of algal blooms. *Ecological Modelling 96*(1-3), 11–28.

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

Recknagel, F. and H. Wilson (2000). Elucidation and prediction of aquatic ecosystems by artificial neuronal networks. In S. Lek and J. Guégan (Eds.), *Artificial Neuronal Networks: Application to Ecology and Evolution.* Berlin: Springer.

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

Reynolds, C. S. (1987). Cyanobacterial water–blooms. *Advances in Botanical Research 13*, 67–143.

Reynolds, C. S., S. W. Wiseman, and M. J. O. Clarke (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.

Riedmiller, M. and H. Braun (1992). RPROP- A fast adaptive learning algorithm. Technical report, Universitat Karlsruhe.

Robarts, R. D. and T. Zohary (1987). Temperature effects on photosynthetic capacity, respiration, and growth rates of bloom–forming cyanobacteria. *New Zealand Journal of Marine and Freshwater Research 21*, 391–399.

Rosenblatt, F. (1962). *Principles of Neurodynamics*. New York: Spartan.

Ruley, J. E. and K. A. Rusch (2002). An assessment of long–term post–restoration water quality trends in a shallow, subtropical, urban hypereutrophic lake. *Ecological Engineering 19*, 265–280.

Rumelhart, D. E., G. E. Hinton, and R. J. Williams (1986). Learning internal representations by error propagation. In D. E. Rumelhart, J. L. McClelland, and T. P. R. Group (Eds.), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, Volume 1: Foundations. MIT Press.

Ryding, S.-O. and W. Rast (Eds.) (1989). *The Control of Eutrophication of Lakes and Reservoirs*, Volume 1. Paris: UNESCO and The Parthenon Publishing Group.

Saarinen, S., R. Bramley, and G. Cybenko (1993). Ill-conditioning in neural network training problems. *SIAM Journal on Scientific Computing 14*(3), 693–714.

Sakamoto, M. (1966). Primary production by phytoplankton community in some Japanese lakes and its dependence on lake depth. *Arch Hydrobiol 62*, 1–28.

Sarle, W. S. (2001, July). Neural network faq, part 1 of 7: Introduction, periodic posting to the usenet newsgroup comp.ai.neural-nets.

Scardi, M. (2001). Advances in neural network modeling of phytoplankton primary production. *Ecological Modelling 146*, 33–45.

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

Schaap, M. G. and W. Bouten (1996). Modeling water retention curves of sandy soils using neural networks. *Water Resources Research 32*(10), 3033–3040.

Schleiter, I. M., D. Borchardt, R. Wagner, T. Dapper, K.-D. Schmidt, H.-H. Shmidt, and H. Werner (1999). Modelling water quality, bioindication and population dynamics in lotic ecosystems using neural networks. *Ecological Modelling 120*, 271–286.

Schreurs, H. (1992). *Cyanobacterial dominance. Relations to eutrophication and lake morphology*. Phd, Univ. Amsterdam.

Senate Standing Committee on Environment Recreation and the Arts (Aust.) (1993, December). Water resources, toxic algae. Technical report, The Parliament of the Commonwealth of Australia.

Shapiro, J. (1990). Current beliefs regarding dominance by blue–greens: the case for the importance of pH and CO_2. *Int. Revue Ges. Hydrobiol 24*, 38–54.

Sivonen, K. and G. Jones (1999). See chapter 2. In I. Chorus and J. Bartram (Eds.), *Toxic Cyanobacteria in Water. A Guide to Public Health Consequences and their Supplies. WHO Series in Environmental Management*. London.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society. Series A (General). 137*(2), 111–133.

Sun Microsystems Inc. (2001). Java JDK Version 1.3. http://java.sun.com.

Takahashi, Y. (1993). Generalization and approximation capabilities of multilayer networks. *Neural Computation 5*, 132–139.

Takamura, N. and M. Aizaki (1991). Change in Primary Production in Lake Kasumigaura (1986-1989) Accompanied by Transition of Dominant Species. *Jpn. J. Limnol 52*(3), 173–187.

Takamura, N., A. Otsuki, M. Aizaki, and Y. Nojiri (1992). Phytoplankton species shift accompanied by transition from nitrogen dependence to phosphorus dependence of primary production in Lake Kasumigaura, Japan. *Arch. Hydrobiol. 124*(2), 129–148.

Takamura, T., T. Iwakuma, and M. Yasuno (1987). Primary Production in Lake Kasumigaura, 1981–1985. *Jpn. J. Limnol. 48*, S13–S38.

Teubner, K., R. Feyerabend, M. Henning, A. Nicklisch, P. Woitke, and J. G. Kohl (1997). Alternative blooming of the *Aphanizomenon flos–aquae* or *Planktothrix argardhii* induced by the timing of the critical nitrogen:phosphorous ratio in hypertrophic riverine lakes. *Arch. Hydrobiol., Adv. Limnol. 54*, 325–344.

The R Development Core Team (2001). The R Package for Statistical Computing Version 1.4.1. http://www.r-project.org.

Theil, H. (1961). *Economic Forecasts and Policy*. Amsterdam: North–Holland Publishing Company.

Todorovski, T., S. Džeroski, and B. Kompare (1998). Modelling and prediction of phytoplankton growth with equation discovery. *Ecological Modelling 113*, 71–81.

Trimbee, A. M. and E. E. Prepas (1988). The effect of oxygen depletion on the timing and magnitude of blue–green algal blooms. *Verh. Internat. Verein. Limnol. 23*, 220–226.

Tukey, J. W. (1977). *Exploratory data analysis*. Reading, Mass.: Addison–Wesley Pub. Co.

Ueno, Y. and S. Nagata (1997). ELISA analysis of microcystins, algal hepatotoxins, in environmental water. *Toxicon 35*(4), 482–483.

Ueno, Y., T. Tsutsumi, A. Hasegawa, M. Watanabe, H. Park, G. C. Chen, G. Chen, and S. Yu (1996). Detection of microcystins, a blue–green algal hepatotoxin in drinking water sampled in Haimen and Fusui, endemic areas of primary liver cancer in China, by highly sensitive immunoassay. *Carcinogenesis 17*, 1317–1321.

University of Stuttgart (1999). Stuttgart Neural Network Simulator Version 4.1. http://www-ra.informatik.uni-tuebingen.de/SNNS/.

Urabe, J., T. B. Gurung, and T. Yoshida (1999). Effects of phosphorous supply on phagotrophy by the mixotrophic alga *Uroglena americana* (chrysophyceae). *Aquatic Microbial Ecology 18*(1), 77–83.

Van Der Smagt, P. and G. Hirzinger (1998). Solving the ill–conditioning in neural network learning. In G. Orr and Müller (Eds.), *Neural Networks: Tricks of the Trade*, Lecture Notes in Computer Science 1524, pp. 193–206. Springer–Verlag.

Vapnik, V. N. and A. Chervonenkis (1971). On the uniform convergence of relative frequencies of events to their probabilities. *Theory Prob. Appl. 16*, 264–280.

Velleman, P. F. and D. C. Hoaglin (1981). *Applications, Basics, and Computing of Exploratory Data Analysis*. Duxbury Press, Boston, Massachusetts.

Velzeboer, R. M. A., J. A. Cugley, and J. C. Patterson (1991). *Modelling optimum conditions for reservoir destratification using mechanical mixers*. Melbourne and Metropolitan Board of Works.

Vollenweider, R. (1976). Advances in defining critical loading levels for phosphorus in lake eutrophication. *Mem. Ist. Ital. Idrobiol. 33*, 53–83.

Vollenweider, R. and J. Kerekes (1982). *Eutrohpication of waters, monitoring, assessment and control*. Paris: OECD.

Vollenweider, R. A. (1970). Scientific fundamentals of the eutrophication of lakes and flowing waters, with particular reference to nitrogen and phosphorous as factors in eutrophication. Technical Report DAS/SST 68.27, Organisation for Economic Cooperation and Development.

Waibel, A. (1989). Modular construction of time–delay neural networks for speech recognition. *Neural Computation 1*, 39–46.

Walley, W. J. and V. N. Fontama (1998). Neural network predictors of average score per taxon and number of families at unpolluted river sites in Great Britain. *Wat. Res. 32*(3), 613–622.

Walter, M., F. Recknagel, C. Carpenter, and M. Bormans (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.

Wasserman, P. D. (1989). *Neural computing: theory and practice*. New York: Van Nostrand Reinhold.

Wei, B., N. Sugiura, and T. Maekawa (2001). Use of artificial neural network in the prediction of algal blooms. *Water Resources 35*(8), 2022–2028.

Weigend, A. S., D. E. Rumelhart, and B. A. Huberman (1991). Generalization by weight–elimination with applications to forecasting. In R. P. Lippmann, J. E. Moody, and D. S. Touretzky (Eds.), *Advances in Neural Information Processing Systems*, Volume 3, pp. 875–882. San Mateo, CA: Morgan Kaufman.

Weiss, M. and C. Kulikowski (1991). *Computer Systems that Learn: Classification and Prediction Methods from Statistics, Neural Networks, Machine Learning and Expert Systems*. San Mateo, California: M. Kaufmann.

Werbos, P. J. (1974). *Beyond Regression: new tools for prediction and analysis in the behavioural sciences*. PhD thesis, Harvard University.

Weston, B. (Ed.) (1987). *Murray–Darling Basin Environmental Resources Study*. Canberra, Australia: Murray–Darling Basin Ministerial Council.

Whigham, P. A. and F. Recknagel (2001). An inductive approach to ecological time series modelling by evolutionary computation. *Ecological Modelling 146*, 275–287.

Whitaker, J., J. Barica, H. Kling, and M. Buckley (1978). Efficacy of copper sulphate in the suppression of *Aphanizonenon flos-aquae* blooms in prairie lakes. *Environmental Pollution 15*, 185–194.

Whitehead, P. and G. Hornberger (1984). Modelling algal behaviour in the River Thames. *Water Research 18*(8), 945–953.

Whitehead, P. G., A. Howard, and C. Arulmani (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.

Widrow, B. and M. E. Hoff (1960). Adaptive switching circuits. *IRE WESTCON Connection Record 4*, 96–104.

Wilson, H. and F. Recknagel (1997). Advances in modelling and prediction of algal blooms in freshwater lakes by artificial neural networks. In *MODSIM 97, International congress on modelling and simulation*, Volume 4, Hobart, Australia, pp. 1771–1777.

Wilson, H. E. C. and F. A. Recknagel (2001). Towards a generic artificial neural network model for dynamic predictions of algal abundance in freshwater lakes. *Ecological Modelling 146*(1–3), 69–84.

Winder and Cheng (1995). Quantification of Factors controlling the Development of *Anabaena circinalis* Blooms. Research Report 88, Urban Water Research Association of Australia.

Yabunaka, K., M. Hosomi, and A. Murakami (1997). Novel application of a back-propagation artificial neural network model formulated to predict algal bloom. *Wat. Sci. Tech. 36*(5), 89–97.

Young, W. J., F. M. Marston, and R. Davis (1996). Nutrient exports and land use in Australian catchments. *Journal of Environmental Management 47*, 165–183.

Zevenboom, W. and L. R. Mur (1980). N–fixing cyanobacteria: why they do not become dominant in Dutch hypertrophic lakes. In J. Barica and L. R. Mur (Eds.), *Hypertrophic Ecosystems. Developments in Hydrobiology*, Volume 2, pp. 123–130. The Hague, The Netherlands: Dr W. Junk Publishers.