

Predicting Estuarine Algal Blooms Utilising Neural Network Modelling-A Preliminary Investigation

¹Coad, P., ²Cathers B. and ³Van Senden, D.

¹Hornsby Shire Council, Hornsby, ²University of New South Wales, ³Manly Hydraulics Laboratory,
E-Mail: pcoad@hornsby.nsw.gov.au

Keywords: algal bloom; neural network; coastal resource management

EXTENDED ABSTRACT

Algal blooms are prevalent within the Berowra Estuary. When algal species are present in high numbers they pose serious problems for commercial and recreational users of the estuary. Management authorities require an understanding of the relationship between the incidence of algal blooms and the environmental conditions required to initiate and promote these populations. An Artificial Neural Network (ANN) is currently being developed to predict the occurrence and risk of algal blooms within the Berowra Estuary.

Modelling the algal dynamics for this project is based on a unique data set, for South Eastern Australia, obtained from a deployed probe which monitors Chlorophyll-a (Chl-a), temperature and salinity at 15 minute intervals. Distinguishing features of the present study are that it is being conducted in an estuarine environment with prediction horizons of the order of hours to several weeks. This is in contrast to previous studies which are more commonly set within freshwater environments with the relevant time scales ranging from biweekly to seasonal.

Preliminary network development for this project has utilised the back-propagation training algorithm and the sigmoid activation function. A multilayer perceptron architecture containing an input, hidden and output layer was selected. Data pre-processing and division into training, selection and test subsets occurred prior to being incorporated into the network. Prediction outputs have been generated which aim to provide predictions for 1, 3 and 7 days in advance. Preliminary analysis of the data indicates the best predictive results (i.e. lowest selection error) are obtained with models with the lowest number of variables. Specifically,

time-lagged Chl-a concentrations provide the best data set from which a prediction is made. This suggests initially that antecedent algal concentrations within the previous week are the most significant variable to be used when predicting future Chl-a levels. However, it is acknowledged that with further refinement of internal network geometries and potential alteration to the data preprocessing techniques this may not be the case. This paper outlines initial model results and compares each individual model on its predictive ability whilst maintaining constant internal model geometries between models. Future improvements to the models developed in this paper are expected.

Prediction of Chl-a within the estuarine environment is a suitable application of ANNs. This predictive tool provides opportunities for proactive rather than reactive management regimes with regard to mitigating the effects of estuarine algal blooms. Essential to the implementation and adoption of a proactive strategy is the requirement for a specified degree of certainty in the model outputs, an understanding of problematic algal concentrations and their duration. These requirements are essential to ensure logistics staff and financial support are maintained for an algal bloom early warning system.

1.0 INTRODUCTION

The occurrence of algal blooms within the estuarine ecosystem threatens both recreational and commercial pursuits. Generally, algal blooms occur when a favourable set of environmental conditions exist. Characteristic of blooms within the Berowra Estuary is their occurrence over relatively short time scales- of the order of days to a few weeks. These blooms often lead to the discoloration of estuarine water and in some instances, dissolved oxygen depletion, fish kills and potential shellfish poisoning which may lead to closure of the estuary.

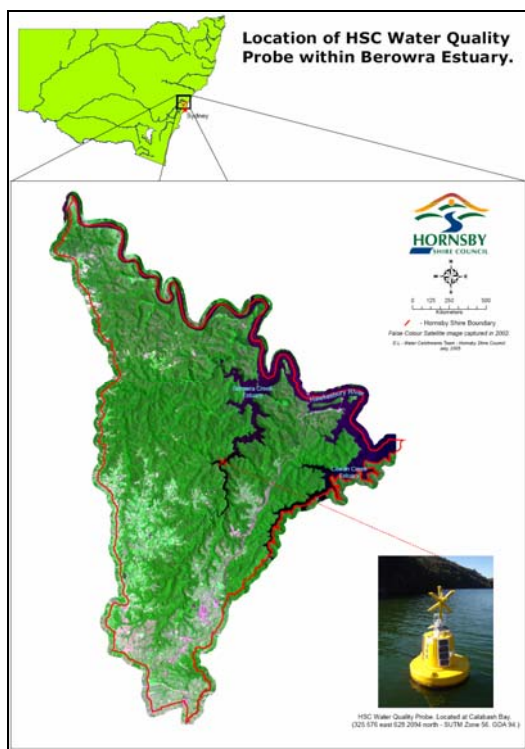


Figure 1 The Berowra Estuary

Current knowledge on the interactions between environmental variables and the resultant ecological algal response to these variables has been studied and largely incorporated into process based mathematical models. Whilst the causality of an algal bloom is well known the actual responding algal dynamic to these causalities is difficult to predict due to the non-linear relationships between environmental variables and algal response. These process based models do not provide adequate forecasting capabilities due to the uncertainty in the kinetic rate coefficients and the complexity

of existing deterministic two or three dimensional models (which are often coupled with hydrodynamic models) (Lee and Huang et al. 2003). It is not the intent of this paper to provide an ecological model of algal bloom dynamics, rather it is to provide a framework for developing a predictive model which utilises Artificial Neural Network (ANN) modelling. The main advantage of this approach is that ANNs are able to model non-linear, dynamic and noisy data, especially when the underlying physical/biological relationships are not fully understood (Lee and Huang et al. 2003).

In providing predictive capabilities to environmental managers, a new management paradigm is created. Predictive tools provide managers with the opportunity to *proactively* manage natural resources rather than *reactively*. Benefits arising from this new management paradigm can result in improved economic and time efficiencies with regard to monitoring programs, staff resourcing, and response times to algal blooms. This project will determine whether management of algal blooms can be assisted by predicting algal blooms 1, 3 and 7 days in advance using an ANN.

2.0 INSTRUMENTATION

Coordination of probe deployment, verification and calibration has been principally undertaken by Manly Hydraulics Laboratory and Hornsby Shire Council. The probe instrumentation, YSI™ 6820 sonde, for this project has been deployed in its current location above a deep hole (approximately 13m in depth) between Calabash Point and Cunio Point in the Berowra Estuary. This site was selected as previous estuary process studies indicate that typical bloom characteristics were shown to have peak Chl-a concentrations around Calabash Bay (MHL 1998).

The YSI™ sonde is deployed at 0.5m depth and is connected to a data logger and mobile phone. Measurements of salinity, temperature and Chl-a are taken every 15 minutes. The data is available online to the public via a web link on the Hornsby Shire Council website (www.hornsby.nsw.gov.au).

To limit potential problems of fouling during deployment in the estuary, the probe has an automated wiper attached. The wiper completes a cleaning cycle of the probe optics prior to each measurement being taken to reduce the occurrence of marine fouling leading to a biasing

of the results. Routine cleaning and probe change-overs every 3 weeks further reduce problems of fouling and vandalism.

3.0 INPUT VARIABLES

Selection of network parameters is undertaken through iterative testing of a number of network scenarios which contain a variety of input variables. The aim is to provide a robust model based on the lowest number of input variables, with a modest data requirement.

Inputs considered for the ANN include nutrients, solar radiation, Chl-a, water temperature and salinity. Nutrients are delivered to the estuary from catchment inflows which include discharges from two Sewerage Treatment Plants (STP). These STP's contribute more than 25% of the total phosphorus load and 97% of the total nitrogen load to the estuary (MHL 1998). Despite these nutrient loads entering the estuary, hydrodynamic investigations suggest that the highly variable light field and spring and neap tide variations in salinity dispersion control the algal biomass distribution. SALMON-Q model results indicate that light and salinity dispersion were more important than nutrient limitation for controlling algal biomass and possible blooms (MHL 1998).

Tidal characteristics also play an important role in determining the presence of phytoplankton which require reasonably stable conditions (ie long residence times) to reach bloom proportions.

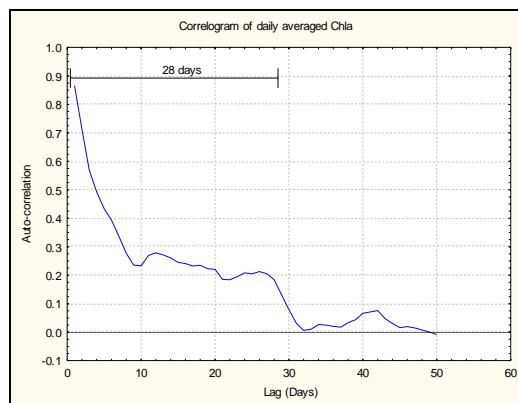


Figure 2 Correlogram of daily averaged Chl-a

Figure 2 illustrates that in the period 05/07/2002 to 01/01/2003, autocorrelation of daily-averaged Chl-a indicates Chl-a concentrations are significantly correlated for time lags of up to 28 days. The significant autocorrelation of Chl-a

within the first 7 days is considered to be the result of the flushing characteristics at the probe location. As the probe is located above a deep hole a water residence time of approximately 7 days is maintained (MHL 1998).

Water temperature is also considered an important variable controlling algal biomass and subsequent bloom conditions. Temperature is considered within the development of these initial models with consideration given to diurnal and seasonal patterns. These patterns indicate that the warmest diurnal temperatures occur during the afternoon and that the warmest seasonal temperatures are associated with summer months. These warm periods are mostly associated with the highest Chl-a measurements.

Therefore, variables considered to be most influential for the prediction of algal blooms include; time lagged Chl-a, water temperature and salinity. Tidal range data as an input variable will be considered in future model developments, at the time of writing this paper the data was not available.

4.0 NETWORK ARCHITECTURE AND LEARNING ALGORITHM

Being conceptually based on biological nervous systems, ANNs consist of a large number of highly interconnected processing elements. Specifically, they attempt to mimic the fault tolerance and the learning capacity of biological neural systems by modelling a low-level structure of the brain.

ANNs contain artificial neurons which receive a number of inputs (either from original data or from the output of other neurons in the network). Each of these inputs comes via a connection that has a strength (or weight); these weights correspond to the synaptic efficiency in a biological neuron. Each neuron also has a single threshold value. The weighted sum of the inputs is formed, and the threshold subtracted, to compose the activation of the neuron. The activation signal is passed through an activation function (also known as the sigmoid transfer function(1)) to produce the output of the neuron.

$$f(x) = \frac{1}{1 - e^{-x}} \quad (1)$$

In utilising this function the trained network is able to undertake non-linear interpolations in order to provide a reasonable output in response to a range of inputs (Maier 1995; Lee and Huang

et al. 2003). This function is considered appropriate for algal bloom predictions as the onset of algal blooms and associated growth is determined by highly non-linear processes, such as nutrient uptake by phytoplankton, the light limitation factor and photo-inhibition (Lee and Huang et al. 2003).

Neural networks learn the input/output relationship through training which in this study utilises the back propagation algorithm. This algorithm uses the data to adjust the network's weights and thresholds so as to minimise the error in the predictions on the training set. Once the network is properly trained, it has learned to model the (unknown) function that relates the input variables to the output variables, and can be used to make predictions where the output is not known (StatSoft 2004).

The back propagation training algorithm is commonly used within many ANNs and is described within numerous texts (Haykin 1994; Bishop 1995; StatSoft 2004). Benefits in using this algorithm are that it has lower memory requirements than most algorithms, and usually reaches an 'acceptable' error level relatively quickly, although it can be very slow to converge properly on an error minimum (StatSoft 2004). An on-line version of back propagation (2) is used within this investigation, that is, it calculates the local gradient of each weight with respect to each case during training. Weights are updated once per training case.

The update formula is (StatSoft 2004):

$$\Delta w_{ij}(t) = \eta \delta_j o_i + \alpha \Delta w_{ij}(t-1) \quad (2)$$

w_{ij} - the weights between the input and hidden layer or between the hidden and output layers (depending on the local error gradient being used)

t - the epoch number

η - the learning rate

δ_j - the local error gradient

α - the momentum coefficient

o_i - the output of the i 'th unit

Thresholds are treated as weights with $o_i = -1$.

Within this formula the local error gradient calculation depends on whether the unit into which the weights feed is in the output layer or the hidden layers. If the local gradients are in the output layers then they are the product of the

derivatives of the network's error function and the units' activation functions. Local gradients in the hidden layers are the weighted sum of the unit's outgoing weights and the local gradients of the units to which these weights connect (StatSoft 2004).

5.0 DATA PRE-PROCESSING

Data pre-processing is essential to remove spurious data which can result from probe malfunctions. Caution is given to readings where consecutive 15 minute Chl-a readings have exceeded 10 μ g/L and are subsequently removed. For temperature and salinity data, values were removed where there was a 1 $^{\circ}$ C and 10ppt difference respectively in consecutive data values. These thresholds for excluding data values have been set subjectively and do require further attention. Values removed were replaced with an alternate value through linear interpolation.

6.0 INPUT SELECTION

An optimisation of the input variables was undertaken to select those parameters considered most useful for prediction of Chl-a. The procedure started with the model containing all parameters and then removal of one parameter at a time until Chl-a was the only input parameter. For the purpose of this initial investigation three models were developed based on this procedure:

- Model-1 Chl-a, Temp and Sal
- Model-2 Chl-a, Temp
- Model-3 Chl-a

To enable consistency with comparisons between models, all internal model coefficients and geometries were held constant in all models. Thus, the three models presented here maybe capable of further improvement with appropriate adjustment of the internal coefficients and geometries (e.g learning and training parameters, initialisation of weights, etc).

7.0 NETWORK PERFORMANCE

In assessing the performance of the various models, use is made of several test statistics. These test statistics include the Root Mean Square Error (RMSE), Standard Deviation Ratio (S.D. Ratio) and Pearson-R correlation coefficient (r^2).

Of particular interest is the prediction error, S.D. Ratio. This ratio is calculated by comparing the prediction data error standard deviation with either the training or test data standard deviation. A ratio significantly below 1.0 is indicative of good regression. Furthermore, a value less than 0.1 indicates very good regression performance. Care needs to be taken when using this statistic to select networks, as good performance rates based on the training set can be deceptive and may actually be indicating “over-learning” (which is similar to “over fitting” in linear models).

In comparing the three models, the ‘best’ prediction performance, lowest error, and ‘best’ correlations between input and predicted data was achieved by model-3 which utilised Chl-a data only. Subsequent improvement in model predictions was not achieved by adding more environmental variables to this model (ie temperature and salinity).

Table 1 Input and Model Selection

Model	RMS Error		S.D. Ratio Performance		Correlation r ²	
	Train	Test	Train	Test	Train	Test
1. Chla, Temp, Sal	0.116	0.115	0.925	0.941	0.422	0.248
2. Chla, Temp	0.077	0.076	0.838	0.838	0.547	0.550
3. Chla	0.043	0.052	0.383	0.443	0.925	0.897

This does not suggest that salinity and light are not important ecological considerations in ecological models, rather it indicates that these parameters are potentially redundant in predictive models which utilise ANNs. Specifically, it appears that all the past ecological information required to make a prediction is embodied in the historic Chl-a concentrations (Lee and Huang et al. 2003). This makes ecological sense as the presence of algal biomass is the result of (or highly correlated with) the previous behaviour of the phytoplankton population 2 weeks prior.

To further extrapolate the use of various parameters within the neural network, a sensitivity analysis was undertaken.

8.0 SENSITIVITY ANALYSIS

A sensitivity analysis was conducted on the inputs to the ANN to determine which inputs are considered most beneficial. Sensitivity analysis ranks the variables according to the deterioration in modeling performance that occurs if that variable is no longer available to the model. Care is to be taken in interpreting sensitivity results as it is assumed that all the input variables are independent and ignores the possibilities that there may be in fact subtle interdependencies between variables (StatSoft 2004).

Table 2 Sensitivity analysis

Model	Variable		
	Chla	Temp	Sal
1. Chla, Temp, Sal	50.79	1.17	0.96
2. Chla, Temp	75.73	1.40	-
3. Chla	120.54	-	-

The measure of sensitivity is determined as a ratio of the error with missing value substitution to the original error. Therefore, based on the assumption that by removing some of the information required by the network to make a prediction (ie one of the inputs) then it is reasonable to expect some deterioration in error to occur. Thus, the more sensitive the network is to a particular input, the greater the deterioration expected, and hence the greater the ratio.

From Table 2 it is possible to see the effect of removing various variables. Firstly, salinity within Model-1 recorded a value less than 1 indicating that this parameter contains no discernable information that is useful for prediction of Chl-a. That is, the network error actually decreases when the training data mean is substituted. Removal of the salinity variable improved the network errors on the test data. Further removal of temperature from the model input improved the error ratio as shown by the result for Model-3 when only Chl-a data is used. Thus Model-3 is selected as it has the lowest selection error and is able to produce the most accurate predictions (Table 1) based on the fewest variables (Table 2).

9.0 NETWORK PREDICTIONS

The objective of this project is to predict the values of Chl-a, given known input variables. From Figure 3 it can be seen that Model-3 is generally mimicking the pattern of Chl-a, however, the magnitude of the Chl-a concentration is being well represented at low Chl-a values only.

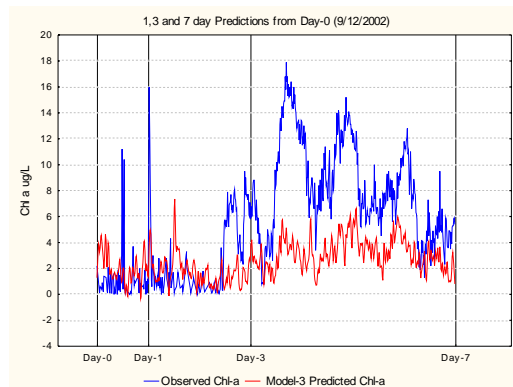


Figure 3 Model-3 network prediction scenario at 1,3 and 7 days in advance

Reasons for this are most probably related to the effect from output scaling and extrapolation. In using a sigmoid activation function (also known as a transfer function) there is a scaling effect whereby the independent variable which has an initial input range from $-\infty$ to ∞ is transformed to a output range of between 0 and 1. That is, the activation function can accept input in any range but produces output within a limited range. Thus to ensure that the network's output will be in reasonable range a scaling function is then utilised. The scaling function used in this project finds the minimum and maximum values of a variable in the training data, and performs a linear transformation (using a shift and a scale factor) to convert the values into the target range (typically [0.0, 1.0]). The net effect is that a 0.0 output activation level in the network is translated into the minimum value encountered in the training data, and a 1.0 activation level is translated into the maximum training data value. Consequently, the network is able to interpolate between the values represented in the training data. However, extrapolation outside the range encountered in the training set is limited as the network's output will be constrained to lie within this range. This can be considered to be both beneficial and also restrictive (StatSoft 2004). Beneficial, in terms that it prevents the network

from making predictions beyond the range of the training set and thus limiting extrapolation. If it is considered restrictive, a linear activation function can be used to enable extrapolation beyond the range of the training data set, however, this can lead to increasing the risk of spurious model results. The implications of using various scaling and activation functions require further investigation.

10.0 PROACTIVE MANAGEMENT REGIMES

It is anticipated that with the development of an algal bloom prediction model, environmental managers will be able to undertake proactive management strategies. Therefore, by being forewarned of an impending algal bloom, management actions can be undertaken prior to the occurrence of an algal bloom. This is contrary to the current management paradigm of reactive management, whereby a response to an algal bloom is made only when they occur. Integration of predictive model outputs into proactive management regimes will assist in minimising the impacts of algal blooms and will provide adequate warning to public and private sectors of an impending algal bloom. Early detection of an algal bloom provides managers with essentially more options in dealing with the bloom situation undertaken (Nancarrow and Wood 2000).

Pivotal to the implementation and adoption of a proactive strategy, based on a predictive model, environmental managers require the following:

- a specified degree of certainty/confidence in the network predictions 1, 3 and 7 days in advance,
- an understanding of problematic Chl-a levels and subsequent alert levels, and
- an understanding of the potential duration of high Chl-a levels.

These requirements are essential to ensure logistics, staff and financial support are maintained for an algal bloom early warning system.

11.0 FUTURE DEVELOPMENT

Future modifications of the three models presented requiring investigation include:

- Use of different activation functions (eg linear, hyperbolic)

- Use of different scaling functions depending on management requirements
- Use of different network architectures
- Use of longer term data sets. A 6 month data set has been used for this study.
- Use of Bureau of Meteorology forecast data to assist with model predictions
- Use of tidal range data and solar radiation as input variables

12.0 CONCLUSIONS

The use of neural networks for the prediction of estuarine Chl-a levels is currently being investigated for the Berowra Estuary. The multilayer perceptron architecture has been developed utilising a backpropagation training algorithm, logistic activation function and minimum/maximum scaling function. In using this framework the 'best' prediction results were obtained, based on time lagged Chl-a data as the only inputs. This suggests that information required to make future Chl-a predictions is encapsulated within the historical Chl-a concentrations. The exact causative parameter for inducing an algal bloom is not known and cannot be deduced from the predictive models presented within this paper. Causation of algal blooms is best addressed through the interpretation of ecological models.

The concept of proactive management regimes is presented to address the often missing association between model outputs and management response. Central to a management response is an indication of confidence in model outputs and understanding model limitations.

In being able to forecast algal blooms it is envisaged that environmental managers will be able to apply the information to:

- Improve monitoring efficiencies, by monitoring only when problematic concentrations occur.
- Inform recreational users of potential public health risks.
- Inform oyster growers and other commercial operators of potential bloom concentrations.
- Determine which environmental conditions are most useful in predicting

the occurrence of problematic algal concentrations.

This paper presents preliminary results and it is considered that improved models can be made through adjustments of internal network parameters and geometries. Namely, alterations to weight optimisation, learning and momentum rates, network architectures and stopping criterion need further consideration. Further refinement of these ANNs will improve both the predictive capacity of the ANNs and the managerial confidence in the network predictions.

13.0 ACKNOWLEDGEMENTS

This study is supported by Hornsby Shire Council, University of New South Wales and Manly Hydraulics Laboratory. The authors wish to thank Dr Ross McPherson for comments and support for this project. The assistance of Miss Kristy Guise and Mr David Leggett in probe deployment and GIS support is also gratefully acknowledged.

14.0 REFERENCES

- Bishop, C. 1995. *Neural networks for pattern recognition*, Oxford University Press.
- Haykin, S. 1994. *Neural Networks*. Englewood Cliffs, NJ, Prentice-Hall.
- Lee, J. H. W., Y. Huang, et al. 2003. *Neural network modelling of coastal algal blooms*. *Ecological Modelling* **159**(2-3): 179-201.
- Maier, H. R. 1995. *A review of Artificial Neural Networks*. Department of Civil and Environmental Engineering. The University of Adelaide.
- Manly Hydraulic Laboratory MHL, 1998. *Berowra Creek Estuary Processes Study Estuarine Water Quality*. NSW Department of Commerce.
- Nancarrow, S. and J. Wood 2000. *Algal Contingency Plan*. Sydney Metropolitan/South Coast Regional Algal Coordinating Committee, NSW Department of Land and Water Conservation.
- StatSoft, Inc. 2004. *STATISTICA (data analysis software system), version 7*.