Neural Network Based Modelling of Environmental Variables: A Systematic Approach

H. R. MAIER
Engineering Adviser, Western Samoa Water Authority
G. C. DANDY
Associate Professor, Department of Civil and Environmental Engineering, University of Adelaide

Abstract: Artificial Neural Networks (ANNs) of the back-propagation type are a useful tool for modelling environmental systems. They have already been successfully used to predict salinity, nutrient concentrations and incidences of blue-green algae. These successes, coupled with their suitability for modelling complex systems, have resulted in an increase in their popularity and their application in an ever increasing number of areas. They are generally treated as black box models that are able to capture underlying relationships when presented with input and output data. In many instances, little consideration is given to potential input data and the internal workings of ANNs. This can result in inferior model performance and an inability to accurately compare the performance of different ANN models. Back-propagation networks employ a modelling philosophy that is similar to that of statistical methods in the sense that unknown model parameters (i.e. connection weights) are adjusted in order to obtain the best match between a historical set of model inputs and corresponding outputs. Consequently, the principles that are considered good practice in the development of statistical models should be considered. In this paper, a systematic approach to the development of ANN based forecasting models is presented, which is intended to act as a guide for potential and current users of back-propagation networks. Issues that need to be considered in the model development phase are discussed and ways of addressing them presented. The major areas covered include data transformation, the determination of appropriate model inputs, the determination of an appropriate network geometry, the optimisation of connection weights and validation of model performance.

1. INTRODUCTION

In recent years, Artificial Neural Networks (ANNs) have become a popular and useful tool for modelling environmental systems. They have already been successfully used to simulate the export of nutrients from river basins [Clair and Ehrman, 1996], to forecast salinity [DeSilets et al., 1992], to predict incidences of blue-green algae [Maier and Dandy, 1997a], and are being considered for a variety of other applications. Many environmental modellers are "experimenting" with ANNs on datasets for which the use of more conventional techniques (e.g. regression) has been unsuccessful. However, a large proportion of users are not "experts" in the use of ANNs and tend to treat them as a "black box". Data pre-processing, methods for determining adequate model inputs and the internal workings of ANNs are seldom considered in the model building process. This can result in inferior model performance and an inability to accurately compare the performance of different ANN models.

In this paper, guidelines for developing ANN forecasting models are presented to assist current and potential users of back-propagation neural networks. The concepts presented are illustrated with two case studies: the forecasting of salinity in the River Murray at Murray Bridge, South Australia, 14 days in advance [Maier and Dandy, 1996a, 1997c] and the forecasting of incidences of a species group of the cyanobacterium *Anabaena* spp. in the River Murray at Morgan [Maier and Dandy, 1997b] four weeks in advance.

2. ANNS: A MODELLING TOOL

ANNs provide a means of computation inspired by the structure and operation of the brain and central nervous system. They operate as a parallel computer, which consists of a number of processing elements (PEs) that are interconnected. Typically, the PEs are arranged in layers; an input layer, one or more hidden layers and an output layer (Figure 1). The input from each PE in the previous layer (x_i) is multiplied by a connection weight (w_{ji}). These connection weights are adjustable and may be likened to the coefficients in statistical models. At each PE, the weighted input signals are summed and a threshold value (θ_j) is added. This combined input (I_j) is then passed through a non-linear transfer function (f(.)) to produce the output of the PE (y_j). The output of one PE provides the input to the PEs in the next layer. This process is summarised in (1) and (2) and illustrated in Figure 1.

$$I_j = \sum w_{ji} x_i + \theta_j \qquad \text{summation} \qquad (1)$$

$$y_j = f(I_j) \qquad \text{transfer} \qquad (2)$$

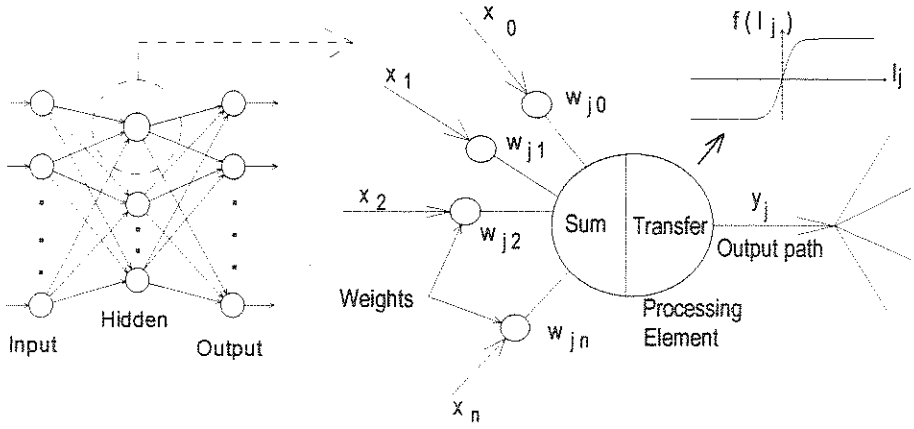

Figure 1: Operation of a Typical Artificial Neural Network

ANNs are well suited to environmental modelling as they are non-linear [Chakraborty et al., 1992], relatively insensitive to data noise [Tang et al., 1991] and perform reasonably well when limited data are available [Tang et al., 1991]. When ANNs are used for the prediction of environmental variables, the modelling philosophy employed is similar to that used in the development of more conventional statistical models. In both cases, the purpose of the model is to capture the relationship between a historical set of model inputs and corresponding outputs. This is achieved by repeatedly presenting examples of the input / output relationship to the model and adjusting the model coefficients (i.e. the connection weights) in an attempt to minimise an error function between the historical outputs and the outputs predicted by the model.

3. THE MODEL DEVELOPMENT PROCESS

As discussed in Section 2, the modelling philosophy used to develop ANN and conventional statistical models is similar. Consequently, the principles that are considered good practice in the development of statistical models should be given careful consideration. The major areas that should be considered include data transformation, the choice of adequate model inputs, the choice of an appropriate network geometry, parameter estimation and model validation.

3.1 Data Transformation

In any model development process, familiarity with the available data is of the upmost importance. Issues in relation to the statistical distribution of the input data, and the effects of trends, seasonal variation and heteroscedasticity are of major importance when more traditional statistical techniques are being considered. However, they are generally considered less important in the development of ANN models.

A normally distributed data set is a prerequisite when traditional regression or ARMA (AutoRegressive Moving Average) type models are being developed. This is a severe restriction when modelling environmental data, as they are often not normally distributed, and their nature is such that it is extremely difficult, if not impossible, to find suitable transformations to normality. It is suggested in the literature (e.g. Burke and Ignizio [1992]) that ANNs overcome this problem, as the probability distribution of the input data does not have to be known. However, this issue needs to be investigated more fully.

Another requirement when developing time series models of the ARMA type is that the input data have to be stationary. The effect of stationary and non-stationary input data on ANN models was investigated by Maier and Dandy [1996b]. Their findings indicate that ANNs have the ability to cater for non-stationarities in time series data with the aid of their hidden layer nodes. The data used in the study exhibited irregular seasonal variation, but did not contain any trends or heteroscedasticity. Consequently, the ability of ANNs to deal with trends and heteroscedasticity in the data needs further investigation.

The data used for the salinity case study included daily salinities at Murray Bridge and the upstream sites of Mannum, Morgan, Waikerie and Loxton as well as daily flow at Lock 1 Lower (approx. 100 km upstream of Murray Bridge). All data were available from 1987 to 1991.

The data used for the blue-green algae case study included weekly values of turbidty, colour, temperature, total phosphorus, soluble phosphorus, oxidised nitrogen and total iron at Morgan as well as weekly flows at Lock 7 (approx. 150 km upstream of Morgan). All data were available from 1983/84 to 1992/93.

The time series used in both case studies exhibited non-regular seasonal variations and were non-normally distributed. However, they did not contain any trends or heteroscedasticity. Consequently, it was not considered necessary to transform the data.

3.2 Determination of Model Inputs

In this step, it has to be decided which input variables (z_1, z_2, ..., z_n) to include in the model, as well as which lags (e.g. $z_{j,t-1}$; $z_{j,t-2}$; ..., $z_{j,t-N}$; j=1, 2, ..., n) to use for each of

these.

Choice of Variables:

The choice of input variables is generally based on *a priori* knowledge of causal variables in conjunction with inspections of time series plots of potential inputs and outputs. If the relationship to be modelled is less well understood, cross-correlation analysis can be used. In the salinity case study, analytical methods for determining appropriate input variables were deemed unnecessary, as the underlying processes (i.e. salt transport and saline groundwater accessions) are well understood. The input variables chosen include salinities at Murray Bridge, Mannum, Morgan, Waikerie, Loxton and flow at Lock 1 Lower.

The mechanisms responsible for incidences of blue-green algae, on the other hand, are not well understood. Consequently, various input variables were tried. All available variables (i.e. turbidity, colour, temperature, flow, total phosphorus, soluble phosphorus, oxidised nitrogen and total iron) were considered as potential input variables. Initially, eight models were developed, each using only one of the available input variables. Subsequently, seven models were developed, combining the variable that resulted in the best forecast when only one input variable was used with each of the remaining variables. This procedure was repeated using models with three input variables, four input variables etc., until the addition of any extra variables did not improve model performance. This process resulted in the inclusion of flow, temperature and colour inputs.

Choice of Lags:

In the development of conventional time series models, analytical procedures are generally used to determine which lagged inputs to include from each variable. This is done by evaluating the strength of the relationship between the output time series and the potential input time series at various lags. The lags of the input time series which have a significant influence on the output time series are then selected as model inputs. Most analytical approaches are based on the method of Haugh and Box [1977], which uses cross-correlation analysis.

Analytical approaches are generally not used to determine the inputs for multivariate ANN models. The main reason for this is that ANNs belong to the class of data driven approaches, whereas conventional statistical methods are model driven [Chakraborty et al., 1992]. In model driven approaches, the structure of the model has to be determined first, which is done with the aid of the analytical approach mentioned above, before the unknown model parameters can be estimated. Data driven approaches, on the other hand, have the ability to determine which model inputs are critical, so there is no need for *a priori* rationalisation about relationships between variables.

However, presenting a large number of inputs to ANN models, and relying on the network to determine the critical model inputs, usually increases network size. This has a number of disadvantages, such as increasing training time, increasing the amount of data required to efficiently estimate the connection weights and increasing the number of local minima in the error surface. This is particularly true for complex problems, where the number of potential inputs is large, and where no *a priori* knowledge is available to suggest possible lags at which strong relationships exist between the output time series and the input time series.

Consequently, there are distinct advantages in using an analytical technique to help determine which lags of the input variables should be included in multivariate ANN models. Maier and Dandy [1997c] have evaluated the suitability of the method of Haugh and Box [1977] and a neural network based approach for the above. They found that both methods were suitable, although the neural network based approach was preferred, as it was quicker and simpler to use.

The neural network based method involves the development of n bi-variate ANN models, one for each of the input variables chosen. Each model relates lagged inputs (i.e. at times t-1, t-2, ..., t-N) from one of the input variables to the output variable. The value of N is chosen so that the lags of the input time series that exceed N are not suspected to have any significant effect on the output time series. The strength of the relationship between the output variable and each of the input variables at the different lags is then determined with the aid of sensitivity analyses. As part of the sensitivity analyses, each of the inputs is increased by a certain percentage (e.g. 5%) in turn, and the change in the output caused by the change in the input is calculated. The sensitivity of each input is given by:

$$\text{Sensitivity} = \frac{\%\ \text{change in output}}{\%\ \text{change in input}} \times 100 \qquad (3)$$

Plots of the sensitivities at various lags are then inspected to decide which lags should be included. No fixed level is used to distinguish between significant and non-significant inputs. Instead, the sensitivities are used as a guide to decide which inputs should be chosen by applying some degree of judgement.

For the salinity case study, six bi-variate models were developed, one for each of the six input variables. The output variable was salinity at Murray Bridge at time t+13 (i.e. 14 days in advance) in each case. The model inputs chosen using the above procedure are shown in Table 1. A typical plot of sensitivities is shown in Figure 2. As can be seen from Table 1, inputs at lags 1 to 7 were selected based on the sensitivities shown in Figure 2.

For the blue-green algae case study, eight bi-variate models were developed, one for each of the eight input variables. The output variable was concentrations of *Anabaena* spp. at time t+3 (i.e. 4 weeks in advance) in each case. The lags of each of the potential input variables that were found to have a significant effect on concentrations of *Anabaena* spp. are shown in Table 2. It should be noted that these lagged inputs were used in the process for determining appropriate input variables described above.

Table 1: Input Lags Chosen for the Salinity Case Study

Variable	Location	Lags of Inputs (days)	Total No.
Salinity	Murray Bridge	1, 2	2
Salinity	Mannum	1, 2	2
Salinity	Morgan	1, 2	2
Salinity	Waikerie	1, 2, ..,4	4
Salinity	Loxton	1, 2, .., 7	7
Flow	Lock 1 Lower	1, 2, .., 8	8

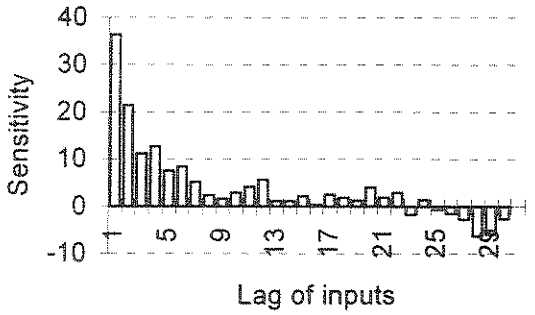

Figure 2: Typical Sensitivities of Salinity Inputs from Loxton

Table 2: Input Lags Chosen for the Blue-Green Algae Case Study

Variable	Lags of Inputs (days)	Total No.
Turbidity	1, 2, ..,8	8
Colour	1, 2, ..,4	4
Temperature	1	1
Flow	1, 2, ..,15	15
Tot. Phosphorus	1, 2	2
Sol. Phosphorus	1, 2, ..,5	5
Oxidised Nitrogen	1, 2, .., 7	7
Tot. Iron	1, 2, .., 21	21

3.3 Choice of Network Geometry

Network geometry is generally defined by the number of hidden layer nodes and the number of nodes in each of these layers. It determines the number of model parameters that need to be estimated. If there is an insufficient number of parameters, it may be difficult to obtain convergence during training, as the network may be unable to obtain an adequate fit to the training data. On the other hand, if too many parameters are used in relation to the number of available training samples, the network may lose its ability to generalise. In addition, keeping the number of parameters to a minimum reduces the computational time needed for training.

In most instances, the use of one hidden layer is sufficient. In fact, it has been shown that ANNs with one hidden layer can approximate any continuous function [Cybenko, 1989]. The optimum number of hidden layer nodes is generally found using a trial and error approach. However, there are some general guidelines which may be followed. Hecht-Nielsen [1987] suggests the following upper limit for the number of hidden layer nodes in order to ensure that ANNs are able to approximate any continuous function:

$$N^H \leq 2N^I+1 \quad (4)$$

where

N^H = number of hidden layer nodes
N^I = number of inputs

However, in order to ensure that the networks do not overfit the training data, the relationship between the number of training samples and network size also needs to be considered. Rogers and Dowla [1994] recommend the following upper limit for the number of hidden layer nodes to satisfy the above criteria:

$$N^H \leq N^{TR}/(N^I+1) \quad (5)$$

where

N^{TR} = number of training samples

Consequently, the upper limit for the number of hidden layer nodes may be taken as the smaller of the values for N^H obtained using (4) and (5). However, in many instances, good performance can be obtained with fewer numbers of nodes.

In the blue-green algae case study, the above guidelines were used to determine network geometries for all models developed. The geometry of the final model selected (i.e. the model using flow, temperature and colour data as inputs) was 20-17-1 (number of inputs - number of hidden nodes - number of outputs). The effect of using 5, 10, 23, 30 and 35 hidden layer nodes was also investigated. The different geometries were found to have negligible impact on model performance. The same was found for the salinity case study, where three different geometries were trialed (25-5-1, 25-15-1 and 25-30-1).

3.4 Parameter Estimation

In the parameter estimation, or "training", phase, the connection weights are adjusted in order to obtain the best fit to the training data. The back-propagation algorithm [Rumelhart et al., 1986] is by far the most popular method of optimising the connection weights and will be discussed here. The back-propagation training process involves the following basic steps:

1. The connection weights are assigned small, arbitrary values.
2. A training sample is presented to the network, producing a network output.
3. The global error function is calculated:

$$E = \frac{1}{2}\sum (o_d - o_p)^2 \quad (6)$$

where

E = global error function
o_d = desired (historical) output
o_p = output predicted by network

4. The connection weights (w) are adjusted using the gradient descent rule of optimisation:

$$\Delta w(t) = \sum_{s=1}^{\varepsilon} -\eta \frac{\partial E}{\partial w} + \mu \Delta w(t-1) \qquad (7)$$

where

s = training sample presented to network
η = learning rate
μ = momentum value

The number of training samples presented to the network between weight updates is called the epoch size (ε).

Steps 2 to 4 are repeated until certain stopping criteria are met. For example, training may be stopped when a fixed number of training samples have been presented to the network or when there is no further improvement in the forecasts obtained using an independent test set.

The way the generalisation ability of a network, as measured by the Root Mean Squared Error (RMSE) between the predicted and historical values of an independent test set, changes as training progresses is a function of the size of the steps taken in weight space (Figure 3). When small steps are taken, the RMSE decreases slowly and steadily until a local minimum in the error surface has been reached (point A, Figure 3). Continued training results in small oscillations in RMSE, as the network jumps from one side of a local minimum to the other (region A - B, Figure 3). When larger steps are taken, the local minimum is reached more quickly (point C, Figure 3), but continued training can result in large oscillations in the RMSE, or even divergent behaviour (region C - D, Figure 3). Clearly, the former network behaviour is more desirable, although very small step sizes should be avoided, as they increase training time.

The way generalisation ability changes as training progresses is highly problem dependent. The absolute step sizes that should be selected to achieve the desired network behaviour needs to be determined for each case study. In order to optimise model performance, it is vital to know at what learn count a local minimum in the error surface is reached and what the magnitude of the oscillations in the forecasting error will be if training is continued.

This requires three data sets: a training set, a test set and a validation set. The test set is used to evaluate the generalisation ability of the network at various learn counts for a variety of step sizes. The validation set is used to assess the performance of the model once the training phase has been completed. For many applications, the data available are limited. In such cases, the amount of data available for training should be maximised. This can be achieved by conducting preliminary studies in which a subset of the training data is used to determine network behaviour for a number of step sizes. The information obtained from these trials can be used to select an appropriate step size and how many training samples should be presented to the network in the training phase. Using this information, all of the training data can then be used and a fixed number of training samples presented to the network.

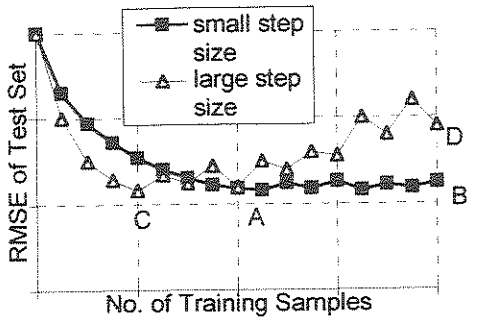

Figure 3: Change in Generalisation Ability as Training Progresses

The size of the steps taken in weight space during training is a function of a number of internal network parameters including the learning rate, momentum value, error function, epoch size and gain of the transfer function [Maier and Dandy, 1997d]. The same step size can be achieved by using different combinations of the above parameters. As discussed above, appropriate step sizes, and hence appropriate combinations of network parameters, have to be determined by trial and error.

For both case studies, limited data were available and trials were conducted using a subset of the training data [Maier and Dandy, 1997b, 1997d]. As a result of these trials, a learning rate of 0.02 was used for the salinity case study. For the blue-green algae case study, the corresponding value was 0.004. A momentum value of 0.6, an epoch size of 16, the quadratic error function and the hyperbolic tangent transfer function were used for both case studies. The number of training samples presented to the network was 100,000 in the salinity case study and 80,000 in the blue-green algae case study.

3.5 Model Validation

Once the training process has been completed, model performance needs to be validated using data that have not been used in the training phase. For both case studies, the latest available year of data was used for this purpose, thus simulating a real-time forecasting situation. A plot of the 14 day forecast of salinity at Murray Bridge for 1991 obtained using the ANN model is shown in Figure 4. Similarly, the four week forecast of concentrations of *Anabaena* spp. at Morgan for 1992/93 are shown in Figure 5. The results obtained indicate that ANNs are a useful tool for forecasting environmental variables.

4. CONCLUSIONS

Back-propagation neural networks have the potential to be a useful tool for modelling environmental variables. In order to optimise their performance, a systematic approach needs to be adopted in the model development phase. The following issues need to be given consideration:

1. Data transformation: Research carried out to date indicates that there is no need to transform data which

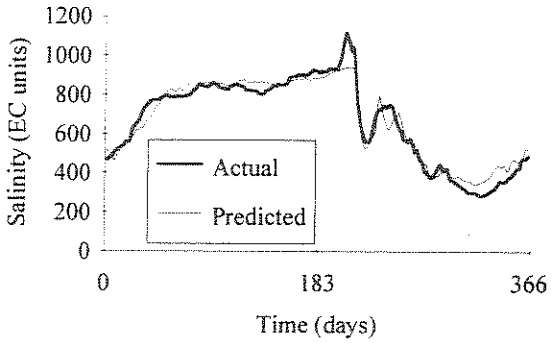

Figure 4: 14 Day Forecast of Salinity at Murray Bridge (1991)

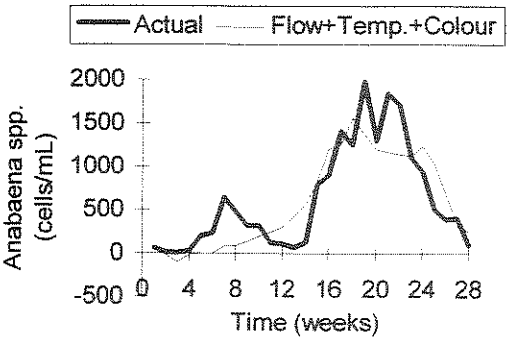

Figure 5: Four Week Forecast of *Anabaena* Concentrations at Morgan (Nov. 1992 to May 1993)

are not normally distributed and which exhibit non-regular seasonal variation. However, the ability of ANNs to deal with data containing trends and heteroscedasticity has not yet been investigated.

2. The determination of appropriate model inputs: The method of Haugh and Box [1977] and a neural network based approach [Maier and Dandy, 1997c] have been found to be suitable tools for determining appropriate model inputs.
3. The choice of an adequate network geometry: The relationship between the number of inputs and the number of hidden nodes and the relationship between the latter and the number of available training samples need to be examined using guidelines given in the literature.
4. Network behaviour during the parameter estimation phase: Trials need to be conducted to determine at what learn count a local minimum in the error surface is reached, and what the oscillations in the RMS forecasting error are with continued training, when different step sizes are taken in weight space. This assists with choosing appropriate network parameters (e.g. learning rate, momentum, epoch size, error function and transfer function) and how many training samples to present to the network for a particular case study.
5. Model validation: Model performance should be assessed using data that have not been used during training.

REFERENCES

Burke, L.I. and Ignizio, J.P., Neural networks and operations research: an overview, *Comp. and Operations Res.*, 19(3/4), 179-189, 1992.

Chakraborty, K., Mehrotra, K., Mohan, C.K. and Ranka, S., Forecasting the behaviour of multivariate time series using neural networks, *Neural Networks*, 5, 961-970, 1992.

Clair, T.A. and Ehrman, J.M., Variations in discharge and dissolved organic carbon and nitrogen export from terrestrial basins with changes in climate: a neural network approach, *Limonl. Oceanogr.*, 41(5), 921-927, 1996.

Cybenko, G., Continuous valued neural networks with two hidden layers are sufficient, Technical Report, Computer Science Department, Tufts Univ., Medford, MA, 1988.

DeSilets, L., Golden, B., Wang, Q. and Kumar, R., Predicting salinity in the Chesapeake Bay using backpropagation, *Comp. and Operations Res.*, 19(3/4), 227-285, 1992.

Haugh, L.D. and Box, G.E.P., Identification of dynamic regression (distributed lag) models connecting two time series. *J. American Statist. Assoc.*, 72(397), 121-130, 1977.

Hecht-Nielsen, R., Kolmogorov's mapping neural network existence theorem, paper presented at First IEEE International Joint Conference on Neural Networks, San Diego, California, June 21-24, 1977.

Maier, H.R. and Dandy, G.C., The use of artificial neural networks for the prediction of water quality parameters., *Water Resour. Res.*, 32(4), 1013-1022, 1996a.

Maier, H.R. and Dandy, G.C., Neural network models for forecasting univariate time series, *Neural Network World*, 5/96, 747-771, 1996b.

Maier, H.R. and Dandy, G.C., Modelling cyanobacteria (blue-green algae) in the River Murray using artificial neural networks, *Math. and Comput. in Simulation*, in press, 1997a.

Maier, H.R. and Dandy, G.C., Use of artificial neural networks for modelling the incidence of cyanobacteria *Anabaena* spp. in River Murray, South Australia., submitted to *Ecol. Model.*, 1997b.

Maier, H.R. and Dandy, G.C., Determining inputs for neural network models of multivariate time series., *Microcomp. in Civil Eng.*, in press, 1997c.

Maier, H.R. and Dandy, G.C., The effect of internal parameters and geometry on the performance of back-propagation neural networks, submitted to *Env. Model. and Software*, 1997d.

Rogers, L.L. and Dowla, F.U., Optimization of groundwater remediation using artificial neural networks with parallel solute transport modeling, *Water Resour. Res.*, 30(2), 457-481, 1994.

Rumelhart, D.E., Hinton G.E. and Williams, R.J., Learning internal representations by error propagation, in *Parallel Distributed Processing*, Vol.1, edited by D.E. Rumelhart and J.L McClelland, MIT Press, 1986.

Tang, Z., deAlmeida, C. and Fishwick, P.A., Time series forecasting using neural networks vs. Box-Jenkins methodology. *Simulation*, 57(5), 303-310, 1991.