Advances in Modelling and Prediction of Algal Blooms in Freshwater Lakes by Artificial Neural Networks

Hugh Wilson and Friedrich Recknagel. University of Adelaide, Department of Environmental Science and Management, Roseworthy, South Australia 5371

Abstract: This paper presents a number of modifications to artificial neural network models for improving predictions of algal blooms. These include a method for finding the optimum hidden layer configuration, 10 fold cross-validation to increase representation of data in the training and validation sets, and controlled training to increase the probability of finding the global optimum training error. These enhancements have succeeded in improving the performance of neural network phytoplankton models in two key areas: i) the model is valid for a broader spectrum of algal species, and ii) a more accurate estimation of the model's performance when applied to new data is obtained. The case study used was Lake Kasumigaura, Japan.

1. INTRODUCTION

The potential of a new type of empirical model using neural networks trained on historical water quality databases has been demonstrated by French and Recknagel (1994), Recknagel et al. (1997a), and Recknagel et al. (1997b). These studies showed that neural networks are capable of predicting algal blooms with superior temporal resolution than other types of phytoplankton growth models such as empirical steady state models, deterministic models, time series models and fuzzy models (Recknagel et al., 1997a). Furthermore, these neural network models are capable of predicting the species of phytoplankton responsible for the bloom where as previous models could at best only predict assemblages of species

The feed-forward neural network phytoplankton model pioneered by the above authors is comprised of three layers of artificial neurons. Water quality parameters deemed to be forcing functions comprise the nodes of the input layer of the neural network. These parameters typically include nutrient concentrations, light intensity, chlorophyll a concentration, zooplankton populations and temperature conditions. The output layer nodes of the network represent a quantity measure of the relevant phytoplankton species. Connection weights between the layers of nodes are set by backpropagation training using a historical time series of water quality and algal population vectors. When training is complete the connection weights are frozen, and the generalisation performance is estimated using the split-plot validation method. This method tests the predictive ability of the trained network on independent validation set comprising of one or more years of data excluded from the training set.

A number of modifications to the methodology used by French and Recknagel (1994), Recknagel et al. (1997a) and Recknagel et al. (1997b) will be introduced in this paper. These include i) more careful optimisation of the hidden layer, ii) 10-fold cross-validation instead of conventional split-plot validation, and iii) a more complete exploration of input layer dimensionality. The aim of these modifications has been to produce a model which is valid for a wider cross-section of algal species.

2. MATERIALS AND METHODS

2.1 Materials - Lake Kasumigaura Database

Lake Kasumigaura (Japan) is a shallow hypertrophic lake which suffers frequent severe algal blooms. It is fully mixed all year round, has a wide temperature range, and is highly unstable with regards to nutrient loads. Hence it is characterised by high stochasticity of the environmental variables which affect growth of phytoplankton. These properties make it an ideal acid test for a phytoplankton population modelling system. The lake is well studied, and a comprehensive database of water quality and algal quantity observations spanning the period 1981 to 1993 is available. This case study has been used to benchmark progress in neural network modelling applications by Recknagel et al. (1997a), and Recknagel et al. (1997b).

Missing values in the water quality record has forced previous neural network studies by the above authors to consider a limited selection of parameters measured from 1984 to 1993. By not investigating all the inputs available it is possible that one or more particularly useful driving variables were

Table 1: 6 Input Layer Permutations

NN No.	Time span	# Inputs	# Records	Inputs
K1	1981 - 93	4	222	PO4-P, DIN, Water Temp, Light
K2	1981 - 92	9	182	K1 + Si, Rain, Radiation, *Diaphanosoma, Bosmina*
K3	1981 - 89	12	137	K2 + *Rotifera, Cladocera, Copepoda*
K4	1984 - 93	9	157	K1 + NO3-N, Secchi Depth, Water Depth, Dissolved Oxygen, pH, Chl-a
K5	1984 - 92	14	117	K4 + Si, Rain, Radiation, *Diaphanosoma, Bosmina*
K6	1984 - 89	17	72	K5 + *Rotifera, Cladocera, Copepoda*

excluded. For the experiments presented in this paper, six permutations of the input parameters were devised to allow all the available water quality parameters to be used as inputs (see table 1). These input layer structures are indicative of the spans of the data for each parameter through the time series. By using all the data available there is no chance that useful driving variables are being omitted from the model. However there may be an increased risk of overfitting by the introduction of irrelevant, or noisy inputs. This risk can be minimised by careful optimisation of the hidden layer as discussed below.

2.2 Methods

2.2.1 Background

Previous implementations of artificial neural networks for predicting phytoplankton blooms used a process of trial and error supplemented by rules of thumb to find the optimum configuration of the hidden layer. In this study, the effect that hidden layer configuration has on model performance is considered in greater depth. The number of hidden nodes in the hidden layer is one way to control the complexity of the function approximated by a neural network (Smith, 1993). Weiss and Kulikowski (1991), Masters (1993), Sarle (1997) and others make general assumptions about the effect of the number of hidden nodes, and hence model complexity, on training and validation performance. As hidden nodes are added, the model gains complexity causing the training error to fall as the model generates a more accurate map of the training database. The validation error may fall at first, stabilise at an optimum, and then rise as overfitting or noise modelling starts to take effect. Optimising the number of hidden nodes is a simple matter of locating the model complexity between upper and lower bounds which offers the best validation performance. Comparing validation results of a number of networks with differently sized hidden layers is the most reliable way to identify the best compromise in model complexity (Masters, 1993).

French and Recknagel (1994), Recknagel et al. (1997a) and Recknagel et al. (1997b) used the split plot method of estimating the generalisation or interpolative performance of neural network phytoplankton models. This validation method is statistically inefficient in situations where data is scarce, as the need to exclude a sufficiently large amount of data for validation reduces training set representation. An alternative form of validation which makes more efficient use of data is 10-fold cross-validation. The dataset is randomly split into 10 subsets. 10 networks are created using each of the 10 subsets for validation and the remaining 9 subsets for training. The generalisation error estimate of the model is calculated by averaging the error obtained for all 10 validation sets. Cross-validation provides an almost bias free estimate of generalisation performance (Weiss and Kulikowski, 1991; Ripley, 1996). 90% training set representation increases the probability of producing a good model. 100% validation set representation reduces the effect of variance on the estimate of generalisation performance. Figure 1 illustrates an example of split-plot validation as used by French and Recknagel (1994), Recknagel et al. (1997a) and Recknagel et al. (1997b). Two years of data, e.g. 1986 and 1988, have been excluded from training for validation. Figure 2 illustrates a single cross-validation training run. A small random sample (approx. 10% for 10-fold cross-validation) of records over the entire dataset have been excluded for validation purposes. This process is repeated 10 times during which time all of the data is used exactly once in the validation set.

Previous studies of neural network phytoplankton models by French and Recknagel (1994), Recknagel et al. (1997a), and Recknagel et al. (1997b) interpolated lake data to produce a daily time series from a database with a varying temporal resolution of 1 day to several months. In this study, the use of cross-validation introduced a problem in using interpolated databases. Independence of training and validation data is an essential requirement to unbiased estimation of the validation error (Masters, 1993). Cross-validation requires random drawing of validation datasets from the total database. If the time series has been augmented with interpolated values, it is impossible to prevent a randomly selected validation set from containing values which have been calculated from values in the training set during the interpolation process. Hence the independence of the two sets is destroyed.

2.3 Implementation

Networks with 1, 3, 5, 7, 10, and 20 hidden nodes arranged in a single hidden layer were trained for each of the experiments outlined in table 1. While this is by no means a complete exploration of the effect of hidden nodes on model performance, it is expected that it will be enough to demonstrate the relationship between network validation performance and the number of hidden nodes.

If the neural network is not trained to a global optimum due to premature stopping of training, or entrapment at local optima, the configuration of the hidden layer is no longer the sole determinant of the complexity of the function being

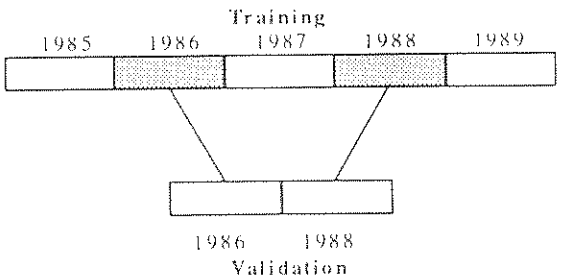

Figure 1: Split Plot Validation

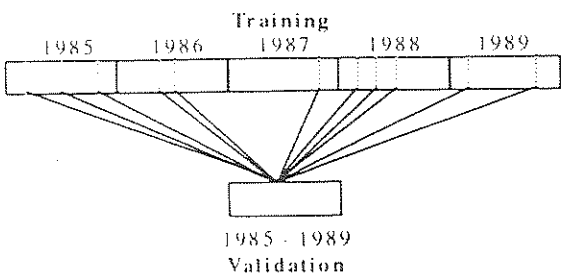

Figure 2: One Cross-Validation run

approximated by the network. This source of variance needs to be eliminated to make a meaningful comparison of hidden layer configurations. As yet there is no learning rule which guarantees that the global optimum can be found every time a network is trained. The only method of improving peace of mind is to increase the probability that the global optimum will be found. In this case study a number of enhancements to conventional backpropagation were applied to achieve this aim.

The "Neural" software package (bundled with Masters, 1993) was used to train all networks. This software employs the conjugate gradient optimiser instead of the more usual backpropagation learning rule. Optimisation algorithms are reputed to lead to faster training than backpropagation (Masters 1993; Ripley, 1996; Sarle, 1997), and have an additional advantage in that there is no need to set the step size or a momentum coefficient.

The Neural package (Masters 1993) has an additional feature where it employs a simulated annealing algorithm when convergence has been achieved which attempts to shake the system out of a local optimum. Simulated annealing is an optimisation algorithm which takes a series of random steps of decreasing size in the search space, and retains those steps which lead to lower error. The networks were trained for as many iterations as necessary to converge on an optimum solution.

As a final less elegant guard against local optima, each network was trained to convergence five times using different random initialisations of connection weights. Out of these five runs, the set of network weights which yielded the best training error measurement was saved for validation as this weight vector is the closest to the global optimum.

Unlike the water quality time series, records of algal quantity are unbroken for the entire duration of the time series. This allowed the 8 most important phytoplankton species in terms of bloom intensity to be used as outputs for all 6 input structures used (see table 1). Neither time nor previous phytoplankton populations were considered as inputs to the models in this study. The neural networks were trained to predict the phytoplankton population in cells per ml which existed at the same time each measurement is taken. Uninterpolated data was used. 10-fold cross-validation was used to estimate the validation performance.

3. RESULTS AND DISCUSSION

3.1 Summary of Validation Results

The fit performance of the validation runs was evaluated visually. Performance is evaluated according to the following criteria : i) the timing of predictions matches the timing of observed bloom events, ii) the magnitude of predictions is proportional to the magnitude of observed events, and iii) the absence of significant false predictions (ie correct prediction of null quantity). Figures 3 to 5 illustrate how this subjective evaluation works. The dotted line indicates the neural network prediction on independent data. Figure 3 depicts an example where the model fails to predict the timing of the actual bloom events. Figure 4 illustrates an example where the timing of a bloom is well predicted, but the magnitudes are not proportional to the actual events. In addition there is a significant false prediction. Figure 5 illustrates a case where all three criteria are well matched.

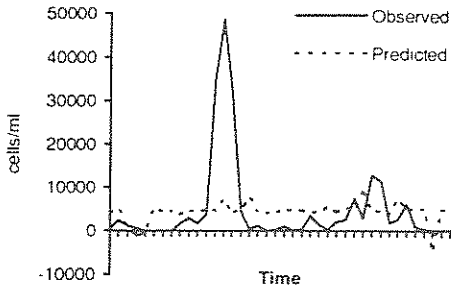

Figure 3: *Cyclotella* - Poor timing

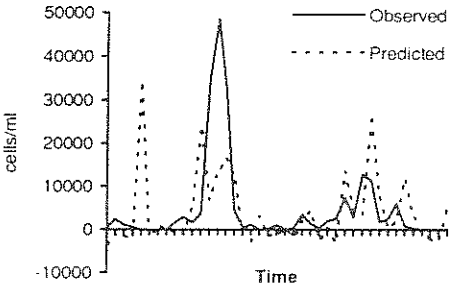

Figure 4: *Cyclotella* - Good timing but poor magnitudes and a false prediction

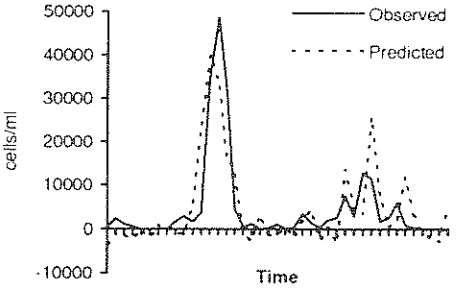

Figure 5: *Cyclotella* - All criteria well handled

Figures 6 to 13 plot observed and predicted cell counts versus time for the 8 species of phytoplankton considered by the model in validation mode. These graphs show that the model is capable of predicting the timing of nearly all bloom events for all 8 species considered by the model. The predictive performance of the model according to the other two criteria (ie magnitude and false predictions) varies according to species. It can be observed that the species performance falls into three distinct groups. Very good predictions of quantity are available for *Microcystis*, *Anabaena*, *Oscillatoria*, and *Ochromonas* (figures 6 to 9). For each of the species there is a good compromise of the timing, magnitude, and false prediction performance measures. Quite good predictions are available for *Cyclotella*, and *Gomphosphaeria* (figures 10

and 11). The timing and false prediction aspects of performance are very good, but the ability of the model to predict magnitude of blooms is not so good. Fair results are available for *Phormidium* and *Synedra* (figures 12 and 13). Timing of bloom predictions is very good, but there is a significant problem with magnitudes and false predictions. Note that the optimum predictive performance for each algal species was achieved with different hidden and input layer configurations.

3.2 Comparisons with Previous Work

The value of the modifications to the neural network phytoplankton model presented in this study can be illustrated by comparing the validation performance with previous models for Lake Kasumigaura developed by Recknagel et al. (1997a) and Recknagel et al. (1997b).

Recknagel et al. (1997b) developed a neural network model which proved to be valid for five problem blue green algae species in Lake Kasumigaura - *Microcystis*, *Oscillatoria*, *Phormidium*, *Gomphosphaeria*, and *Anabaena*. This wide species generality was obtained through selection of a set of training data (corresponding to a year) which produced a network with the best performance on the validation set. Because the validation set was used to identify training data which yielded the best results, it can be argued that the independence between training and validation sets is reduced. As a result, the validation result achieved is likely to be an optimistically biased estimate of the model's ability to generalise to new data. The validation results presented in this paper have been achieved using an unbiased validation procedure. Hence there is a higher degree of confidence that the predictive performance demonstrated can be repeated on newly sampled data. Weiss and Kulikowski (1991) and Ripley (1996) provide some insight into the degree of confidence which can be placed in a validation result.

Recknagel et al. (1997a) on the other hand used an independent validation set. Because of this, the results of this investigation give a better idea of how the model might perform when confronted with new data. The trained neural net in this study obtained good performance in terms of the criteria specified in section 3.1 for a single species - *Microcystis*. Blooms of the other two species considered by the model (*Oscillatoria* and *Phormidium*) were not predicted. It can be concluded that modifications to the methodology introduced in this study have increased the species generality of the model, as good predictions in validation are now available for 4 species, and fair predictions were achieved with the remaining 4 species.

3.3 Effect of Hidden Layer Configuration

Table 2 shows the effect of the number of hidden nodes on validation performance. Results are illustrated for the input layer configuration which yielded the best performance for each species. The performance score is calculated from the number of criteria outlined above in section 3.1 which are satisfied. As an explanation of this scoring system, the example validations illustrated in figures 3 to 5 are considered again. The validation illustrated in figure 3 would score 0 as none of the performance criteria are satisfied. Figure 4 would gain a score of 1 due to the correct prediction of the timing of the observed events. Figure 5 obtains the maximum possible score of 3 as the model correctly predicts the timing and magnitude of blooms. Additionally, there are no significant false predictions.

Table 2: Effect of # Hidden Nodes on Validation Performance (* denotes that result is graphed - see figures 6 to 13)

No. Hidden Nodes	1	3	5	7	10	20	Input
Anabaena	1	3*	2	2	1	0	K3
Cyclotella	0	0	2*	2	1	1	K4
Gomphosphaeria	0	1	1*	1	1	1	K5
Microcystis	2	2	2	2	3	3*	K5
Ochromonas	2	2	1	3*	1	1	K6
Oscillatoria	0	1	2	2	3*	0	K5
Phormidium	0	0	0	0	1*	0	K4
Synedra	0	0	0	0	0	1*	K5

In general validation performance improved with added hidden nodes, reached an optimum, and then declined. This result corresponds with the assumptions described in section 2.2.

These results indicate that validation performance varies according to hidden layer configuration. This observation has noteworthy implications. Firstly, it is demonstrated that careful optimisation of the hidden layer can have a profound effect on neural net performance and therefore should be taken seriously. Secondly, the hidden layer configuration is an important source of variation in validation performance which should probably be accounted for before investigation of other potential sources of variation such as input layer configuration or data representation. If hidden layers are not optimised, it is possibly dangerous to conclude that discrepancies in validation performance between two trained networks are due to experimental differences such as hidden layer configuration or data representation.

3.4 Problems to be Addressed

There are a number of areas in which further work is required:

(1) Validation is assessed in a subjective visual manner. A more objective method of assessing validation will result in more meaningful conclusions.

(2) The model makes same day predictions of algal quantity. The ability to make forecasts of blooms up to several weeks in advance will be of more use from an operational management point of view.

(3) The relative effects of each of the modifications introduced in this study in improving the species generality over results by Recknagel et al. (1997a) have not been determined. More experiments are needed to quantify the importance of each of these modifications (ie. hidden and input layer configuration, removal of

interpolation, and increased training set representation through cross-validation).

(4) The modifications introduced in this study have increased the computational expense of experiments over previous neural network phytoplankton models - the so-called training bottleneck.

4. CONCLUSION

This study demonstrates that a neural network model is capable of making predictions of quantity of a broad spectrum of problem algal species in a freshwater lake from water quality data. This work represents a stepping stone to the ultimate aim of developing a model which can make short term forecasts of algal blooms.

Specific changes in methodology which have lead to the improvement in model validity have not yet been quantified. However, the benefit of careful hidden layer optimisation has been demonstrated. An important message from these results is that the hidden layer optimisation should be a high priority when investigating the effects of other changes in neural network structure such as input layer configuration or data representation.

5. REFERENCES

French, M., and F. Recknagel, Modeling algal blooms in freshwaters using artificial neural networks, in *Computer Techniques in Environmental Studies V*, edited by P. Zannetti, pp 87-94, Computational Mechanics Publications, Boston, 1994.

Maren, A., C. Harston, R. Pap, *Handbook of Neural Computing Applications*, Academic Press Inc., San Diego, California, 1990.

Masters, T., *Practical Neural Network Recipes in C++*, Academic Press Inc., Harcourt Brace Jovanavich Publishers, 493 pp., Boston, 1993.

Recknagel, F., M. French, P. Harkonen, and K. Yabunaka, Artificial neural network approach for modelling and prediction of algal blooms. *Ecological Modelling*, 96(1-3), 11-28, 1997a.

Recknagel, F., T. Fukushima, T. Hanazato, N. Takamura, and H. Wilson, Modelling and prediction of phyto- and zooplankton dynamics in Lake Kasumigaura by artificial neural networks. *Lakes and Reservoirs*, 1997b (in press).

Ripley, B., *Pattern Recognition and Neural Networks*, Cambridge University Press, 403 pp., New York, 1996.

Sarle, W., ai-faq/neural-nets/part3 [Online, accessed 27 June 1997]. Available FTP: ftp.sas.com Directory: pub/neural File: FAQ3.html

Smith, M., *Neural Networks for Statistical Modelling*, Van Nostrand Reinhold, 235 pp., New York, 1993.

Weiss, M. and C. Kulikowski, *Computer Systems that learn: Classification and Prediction methods from statistics, neural nets, machine learning and expert systems*, M Kaufmann, 223 pp., San Mateo, California, 1991.

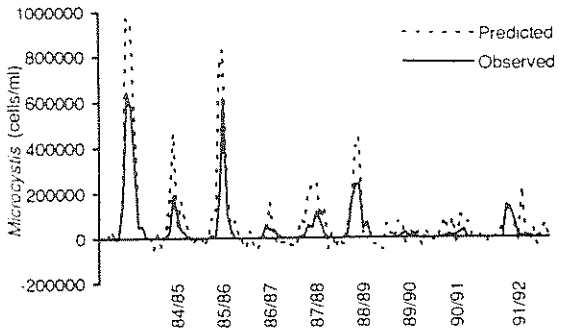

Figure 6: *Microcystis* (20 Hidden nodes; inputs = K5) *

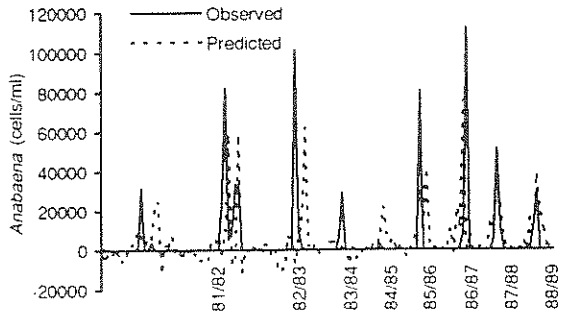

Figure 7: *Anabaena* (3 Hidden nodes; inputs = K3) *

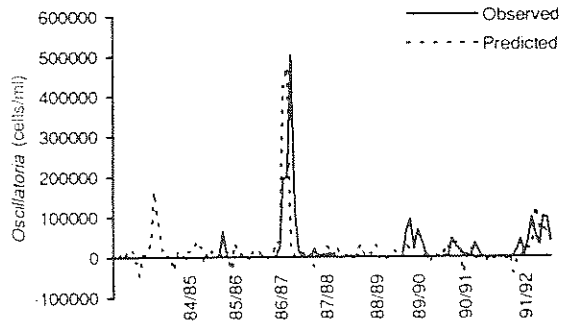

Figure 8: *Oscillatoria* (10 Hidden nodes; inputs = K5) *

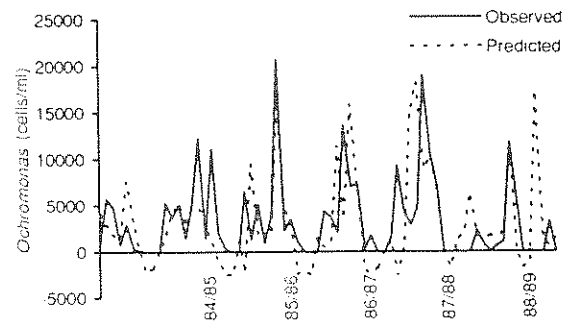

Figure 9: *Ochromonas* (7 Hidden nodes; inputs = K6) *

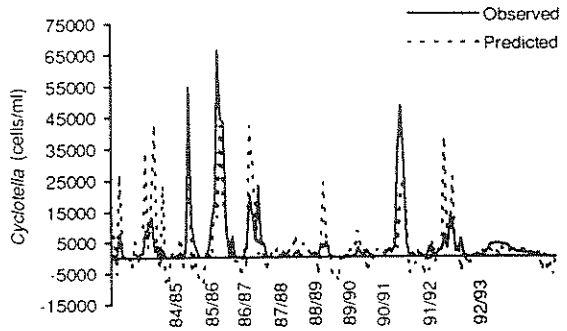

Figure 10: *Cyclotella* (5 Hidden nodes; inputs = K4) *

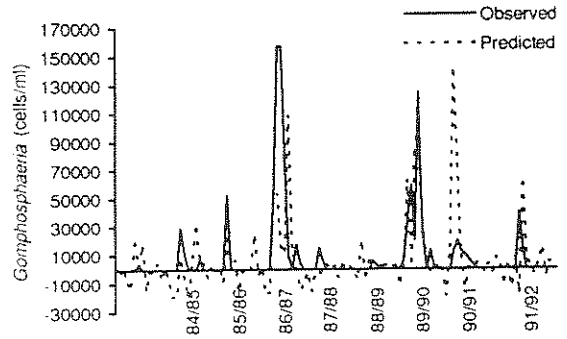

Figure 11: *Gomphosphaeria* (5 Hidden nodes; inputs = K5) *

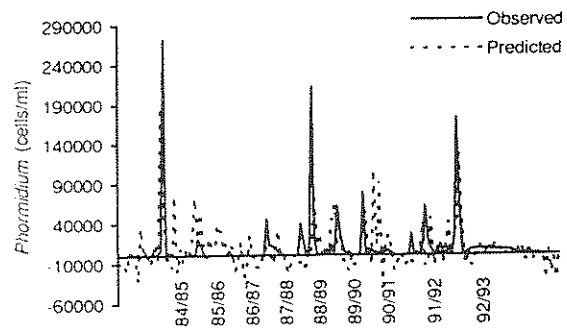

Figure 12: *Phormidium* (10 Hidden nodes; inputs = K4) *

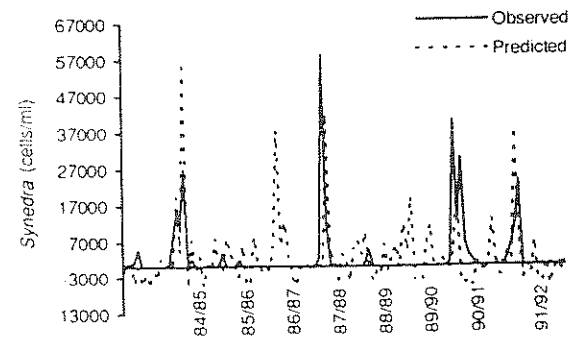

Figure 13: *Synedra* (20 Hidden nodes; inputs = K5) *

* : See table 2 for scoring of these results.