DOI: 10.1007/s00267-005-0074-9

# Time Series Forecasting of Cyanobacteria Blooms in the Crestuma Reservoir (Douro River, Portugal) Using Artificial Neural Networks

**LUIS OLIVA TELES***
**VITOR VASCONCELOS**
Departamento de Zoologia e Antropologia
Faculdade de Ciências da Universidade do Porto
Praça Gomes Teixeira
4099-002 Porto, Portugal

**LUIS OLIVA TELES**
**ELISA PEREIRA**
**MARTIN SAKER**
**VITOR VASCONCELOS**
CIIMAR Centro Interdisciplinar de Investigação Marinha e Ambiental
Rua dos Bragas 289
4050-123 Porto, Portugal

ABSTRACT / In this work, time series neural networks were used to predict the occurrence of toxic cyanobacterial blooms in Crestuma Reservoir, which is an important potable water supply for the Porto region, located in the north of Portugal. These models can potentially be used to provide water treatment plant operators with an early warning for developing cyanobacteria blooms. Physical, chemical, and biological parameters were collected at Crestuma Reservoir from 1999 to 2002. The data set was then divided into three independent time series, each with a fortnightly periodicity. One training series was used to "teach" the neural networks to predict results. Another series was used to verify the results, and to avoid over-fitting of the data. An additional independently collected data series was then used to test the efficacy of the model for predicting the abundance of cyanobacteria. All of the models tested in this study incorporated a prediction time (look-ahead parameter) equal to the sampling interval (two weeks). Various lag periods, from 2 to 52 weeks, were also investigated. The best model produced in this study provided the following correlations between the target and forecast values in the training, verification, and validation series: 1.000 ($P$ = 0.000), 0.802 ($P$ = 0.000), and 0.773 ($P$ = 0.001), respectively. By applying this model to the three-year data set, we were able to predict fluctuations in cyanobacteria abundance in the Crestuma Reservoir, with a high level of precision. By incorporating a lag-period of eight weeks, we were able to detect secondary fluctuations in cyanobacterial abundance over the annual cycle.

Modelling the growth of a population can be complicated by a lack of knowledge on the ecology of individual species. In cases where the prediction of population growth is more important than identification of the underlying processes affecting the population growth, data-driven inductive models are a useful alternative to many theoretical models. Neural networks (NNs) are a type of data-driven inductive model, inspired by the functioning of the brain and nervous system. NNs are able to ''learn'' from and generalize, based on experience (Zhang and others 1998).

KEY WORDS: Cyanobacteria; Neural network; Forecasting; Modelling; Water quality management; Artificial reservoir; Eutrophication

Published online May 19, 2006

*Author to whom correspondence should be addressed; *email:* loteles@fc.up.pt

Traditionally, NNs were designed to mimic some of the tasks performed by the human brain. However, the number of uses for NNs is expanding rapidly, and in recent years an increasing number of engineers and scientists have attempted to use NNs for environmental modeling, in preference to more conventional statistical techniques (Maier and Dandy 1998).

Recently, many researchers have used NN models for forecasting time series data. These studies have shown that NN models are often superior to traditional linear forecasting models (Yun and others 1998). There are several features of NNs that make them valuable and attractive for forecasting. Firstly, in contrast to the traditional model-based methods, NNs are nonparametric data-driven self-adaptive methods and, as a result, incorporate few *a priori* assumptions. NNs are able to learn from examples and respond to subtle functional relationships within the data, even when the underlying relationships are unknown or difficult to describe (Zhang and others 1998). Secondly, another

 

important feature of NNs is that they can generalize. After ''learning'' the data presented to them, NNs can often infer or predict an event or occurrence, even if the sample input data is ''noisy'' (Zhang and others 1998). Thirdly, NNs can approximate nonlinear multivariate functions with high accuracy (Hornik 1991). This is a very important feature since the number of possible nonlinear patterns is, in general, very large for real-world problems (Zhang and others 2001). In addition, NNs can be used when only limited data sets are available (Maier and Dandy 1998), when the diversity of data is great, and when the relationships between causes and effects are vague (Schultz and others 2000).

Initially, NNs were regarded by many people as ''black boxes.'' However, in recent years, there has been increasing interest in the way that trained NNs process different combinations of input data. The most widely used method is known as input sensitivity analysis. This type of analysis can be used to determine the isolated and combined importance of all of the input variables in an NN (Statsoft 2001). Recknagel (2001) found that sensitivity analysis, using trained NN models, can provide useful information regarding the relationship between variables in natural systems.

In this study, we used a generalized regression neural network (GRNN) for the time series forecasting of cyanobacteria abundance in Crestuma Reservoir (Douro River, Portugal), using data collected fortnightly over a three-year period. The use of NNs for modelling cyanobacteria bloom populations is quite recent (French and Recknagel 1994). Most of the models that have been used in previous studies have involved explanatory or causal forecasting. Using these types of models, future events cannot be predicted, even in cases where time series data are used.

Cyanobacteria are a naturally occurring component of the phytoplankton community. These organisms can dominate the aquatic biota and form blooms under favourable growth conditions (Maier and others 1998). Cyanobacteria blooms are costly in terms of water treatment, and are often a risk for public health due to the production of toxic substances. These blooms lead to a general degradation in the quality of freshwater ecosystems (Bobbin and Recknagel 2001a,b).

GRNNs can potentially be used to provide water treatment plant operators with an ''early warning'' tool for the detection of cyanobacteria. It is likely that many of the deleterious effects of cyanobacteria blooms might be prevented or minimized if population growth of these organisms could be predicted at an early stage (Recknagel and others 1997). Toxicological surveys carried out at Crestuma Reservoir have shown that this water body is periodically dominated by *Microcystis aeruginosa*, a species that produces a group of hepatotoxic peptides known as microcystins (Vasconcelos and others 1993) as well as *Aphanizomenon*, a taxon known to produce neurotoxic compounds including several saxitoxin analogues (Ferreira and others 2001).

## Materials and Methods

### Study Site

Crestuma Reservoir is located 23 km from the mouth of the Douro River, Portugal, and is one of the largest hydrographical basins on the Iberian Peninsula. The reservoir is used for energy production, recreational activities (including swimming and sailing), and provides the primary potable water supply for approximately 2,000,000 inhabitants in the Porto region. The reservoir has a length of 44 km and a volume of $110 \times 10^6$ m$^3$.

### Data Collection

Data used in the development of the prediction model were collected fortnightly by ADP (Águas do Douro e Paiva S.A.) from 2000 to 2002. A summary of the data collected over the study period is shown in Table 1. None of the variables, except the phytoplankton data required transformation in order to normalize the variances. In general, NNs do not require input data to be transformed, since the probability distribution of the input data were not affect the model output (Maier and Dandy 2001). Nevertheless, the phytoplankton data were $\log_2$ transformed to reduce the scale range. This transformation was found to give better results than the model produced using untransformed data. The data set was then divided into two independent time series, both with fortnightly periodicity. One ''training series'' was used to ''teach'' the neural network to predict results. The other ''verification series'' was used to cross-validate and avoid overfitting of the data.

Several of the parameters measured in 2000 and 2001 were not recorded in 2002. Due to differences in the availability of data, the data set was divided into two separate studies. In one study, the NN was trained and verified using data collected from 2000 to 2001. In the other study, the NN was trained and verified using data collected over the entire study period from 2000 to 2002.

For the independent validation of the final model, data collected between July and December of 1999

Table 1. Physical, chemical and biological characteristics measured in the Crestuma Reservoir from 2000 to 2002

| Parameter | Notation | Units | Transformation | Time series 2000–2001 | Time series 2000–2002 |
|---|---|---|---|---|---|
| Physical and chemical | | | | | |
| Colour | COL | Scale Pt-Co | $1{\cdot}x^{-1}$ | * | |
| Turbidity | TURB | NTU | $1{\cdot}x^{-1}$ | * | * |
| Water Temperature | TEMP | °C | – | * | * |
| pH | pH | Scale Sorensen | – | * | * |
| Alkalinity | ALK | $mg{\cdot}L^{-1}$ $CaCO_3$ | – | * | |
| Conductivity | COND | $\mu S{\cdot}cm^{-1}$ | – | * | * |
| Oxidability | OXID | $mg{\cdot}L^{-1}$ | $1{\cdot}x^{-1}$ | * | |
| Dissolved oxygen | DO | % saturation | $1{\cdot}x^{-1}$ | * | * |
| Chlorides | Cl | $mg{\cdot}L^{-1}$ | – | * | |
| Nitrate-$NO_3$ | $NO_3$ | $mg{\cdot}L^{-1}$ | $1{\cdot}x^{-1}$ | * | * |
| Sulphate-$SO_4$ | $SO_4$ | $mg{\cdot}L^{-1}$ | – | * | |
| Soluble iron | Fe | $mg{\cdot}L^{-1}$ | $1{\cdot}x^{-1}$ | * | * |
| Total iron | T Fe | $mg{\cdot}L^{-1}$ | $1{\cdot}x^{-1}$ | * | |
| Total suspended soils | TSS | $mg{\cdot}L^{-1}$ | $1{\cdot}x^{-1}$ | * | * |
| Solar radiation | RAD | $Kj{\cdot}m^{-2}$ | – | * | * |
| Water evaporation | EVP | $mm{\cdot}day^{-1}$ | – | * | * |
| Atmospheric precipitation | PRC | $mm{\cdot}day^{-1}$ | $1{\cdot}x^{-1}$ | * | * |
| Discharge | DISCH | $M^3.s^{-1}$ | $1{\cdot}x^{-1}$ | * | * |
| Minimum air temperature | Tmin | °C | – | * | * |
| Maximum air temperature | Tmax | °C | – | * | * |
| Retention time | R time | Day | – | * | * |
| Biological | | | | | |
| Cyanobacteria | CYAN | $cells{\cdot}ml^{-1}$ | $Log_2$ (x+1) | * | * |
| Chlorophytes | CHLOR | $cells{\cdot}ml^{-1}$ | $Log_2$ (x+1) | * | * |
| Cryptophytes | CRYPT | $cells{\cdot}ml^{-1}$ | $Log_2$ (x+1) | * | * |
| Crysophytes | CRYS | $cells{\cdot}ml^{-1}$ | $Log_2$ (x+1) | * | * |
| Diatoms | DIAT | $cells{\cdot}ml^{-1}$ | $Log_2$ (x+1) | * | * |
| Others taxa | O taxa | $cells{\cdot}ml^{-1}$ | $Log_2$ (x+1) | * | * |
| Total phytoplankton | PHYTO | $cells{\cdot}ml^{-1}$ | $Log_2$ (x+1) | * | * |

*Available parameters in the time series respective.

were used. This included the important phases of cyanobacterial growth and bloom development.

### Data Analyses

Statistical analyses were carried out using the commercially available software package Statistica, Version 6 (Statsoft 2001). The regression statistics used to evaluate the prediction accuracy of the models were as follows: Data Mean, average value of the target output variable; Data S.D., standard deviation of the target output variable; Error Mean, average error (residual between target and actual output values) of the output variable; Error S.D., standard deviation of errors for the output variable; Abs. E. Mean, average absolute error (difference between target and actual output values) of the output variable; RMS error, root mean square error; S.D. Ratio, the error/data standard deviation ratio (generally a S.D ratio of 0.1 or lower indicates good regression performance); Correlation, the standard Pearson-R correlation coefficient between the target and actual output values.

### Neural Network for Time Series Forecasting

NNs are data-driven self-adaptive computing models that rely on few *a priori* assumptions regarding model format (Zhang and Hu 1998). These unique features make them valuable for solving many practical forecasting problems. Most time series forecasting models assume that there is an underlying process from which data are generated, and that a future value in a time series is dependent on past and current observations. NNs can be used to identify the underlying pattern or autocorrelation structure within a time series, even when the factors influencing the system are unknown or too complex to describe (Zhang and Hu 1998).

A NN typically consists of an input layer, hidden layers, and an output layer. The neural network undergoes a learning process using an error conver-

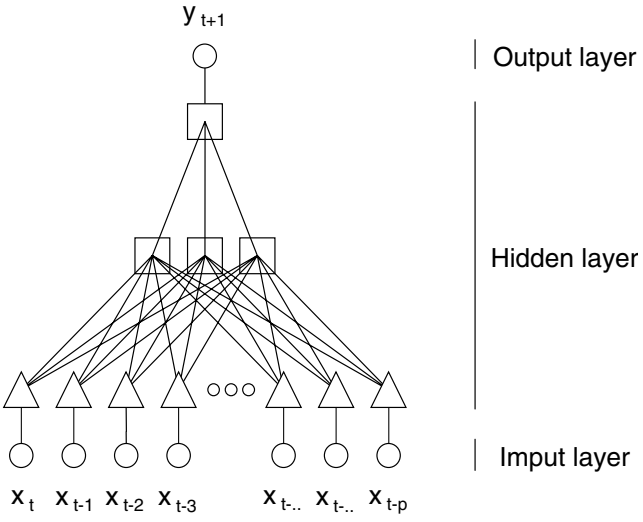


**Figure 1.** Fully connected feed-forward neural network used for time series forecasting.

gence technique (Recknagel 2001). Once provided with data for input and output nodes, the NN determines the weighted connections between the input and output nodes using ''neurons.'' Neurons are defined as interconnected computing elements that are located in hidden layers. They are fed to a nonlinear function, which is typically sigmoid or hyperbolic, by the sum of their inputs either coming from the input nodes (feed-forward) or from the output nodes (feedback). After being processed with the nonlinear function, the value of a neuron is multiplied by a weighting factor. Each neuron has a separate weight parameter for each of the connections with input and output nodes. This value is known as the firing rate. A learning algorithm then adjusts the strength of the interconnections between neurons in order to minimize the output error. The output error is defined as the sum of the difference between the actual output vector and the desired output vector. The weighted values remain fixed in the hidden layer, and the neural network can then be used for predictions. Figure 1 shows an example of a fully connected neural network consisting of one hidden layer.

For explanatory or causal forecasting problems, the NN input data usually consists of independent predictor variables, and a dependent output variable. In this sense, the NN is functionally equivalent to a nonlinear regression model. On the other hand, for an extrapolative or time series forecasting problem, the input data consists of past observations and data series, and the output is typically a future predicted value. Thus the NN is equivalent to a nonlinear autoregressive model for time series forecasting problems. Therefore, both predictor variables and time-lagged observations can easily be incorporated into one NN model (Zhang and others 1998). Figure 1 represents a typical feed-forward NN used for time series forecasting. For a more detailed description of NNs, see Bishop (1995) and Haykin (1999).

### The Model Development Process

The regression NN for time series forecasting was produced using a Pentium II processor with 160.0 MB of RAM, and the commercially available software package Statistica Neural Networks, Version 4.0 F (Statsoft 2000). Regression analyses in the Statistica Neural Networks package employ a Bayesian kernel-based type estimation. These types of NNs are usually referred to as generalized regression neural networks (GRNNs) (Speckt 1991; Patterson 1996; Bishop 1995). Statistica Neural Networks supports feed-forward NNs, which are fully connected between successive layers. The default selections were used for network geometry (number of regression layer nodes), network parameters (PSP and activation functions), pre- and post-data processing (normalization, conversion, and missing value substitution methods), and error function. The prediction time (look-ahead parameter) was equal to the sampling interval (i.e., 2 weeks). Several different lag times, between 2 to 52 weeks, were also investigated. Radial centres were assigned based on the results of the K-means algorithm. The selection of input variables and other factors are discussed in the next section.

### Selection of Input Variables

In any prediction model, definition of the input variables is of paramount importance. However, in most applications involving NNs, little attention is

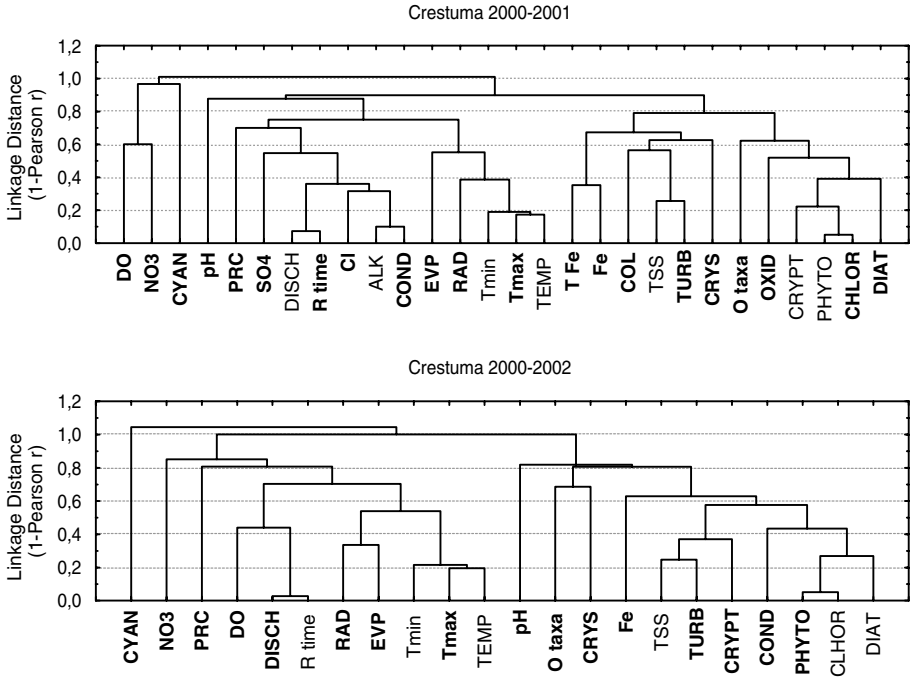


**Figure 2.** Tree diagram showing the relationships between variables. Tree diagram constructed using the weighted pair-group average linkage rule. Variables marked in bold were selected for inclusion in the time series forecasting model.

given to this task. According to Maier and Dandy (2000), the main reason for this is that NNs belong to the class of data-driven approaches that have the ability to determine which model inputs are critical. As a result, there is no need for *a priori* determination of the relationships between variables. In contrast, the excess use of input variables may have a negative influence on the NN model because it decreases the processing speed and affects the redundancy contained in the different variables. A model with too many parameters may result in data overfitting. Such models fit well to sample (or training) data but are of little use for forecasting (Qi and Zhang 2001).

To avoid those problems, it is necessary to carefully select the variables to use in the NN model. The selection of variables is based on acquired *a priori* knowledge. However, the physical and chemical factors responsible for the incidence of cyanobacteria are not well understood (Maier and Dandy 2001). The selection of input variables for the two studies (Crestuma 2000–2001 and Crestuma 2000–2002), was, therefore, carried out using cluster analysis with all of the available variables (Figure 2). Some variables were transformed prior to the analysis (Table 1) to equalize the variances. For each cluster, a representative variable was selected (Figure 2). Variables were chosen that (1) gave the best correlation with the principal components of the phytoplankton groups, and (2) contained the lowest number of missing values.

Before performing the cluster analysis, we selected an appropriate time interval for determining the hydrological and climatic parameters that gave the best correlations with the principal components of the phytoplankton groups for the samples collected between 2000 and 2002.

### Sensitivity Analysis

Sensitivity analyses were carried out using Statistica Neural Networks, Version 4.0 F (Statsoft 2000). This analysis was used to provide information on the relative significance of the input variables in each of the GRNN models. Sensitivity was defined as the square root of the mean squared error (RMS error). This value indicates the performance of a network if the variable under consideration is omitted from the analysis. As a result, more important variables have high RMS error values, indicating that the network is affected to a greater extent when they are not included. If the ratio between the RMS error and the baseline error (i.e., the RMS error of the network if all variables are available) is less than or equal to one, then omitting the variable from the model either has no effect on the performance of the network, or enhances it.

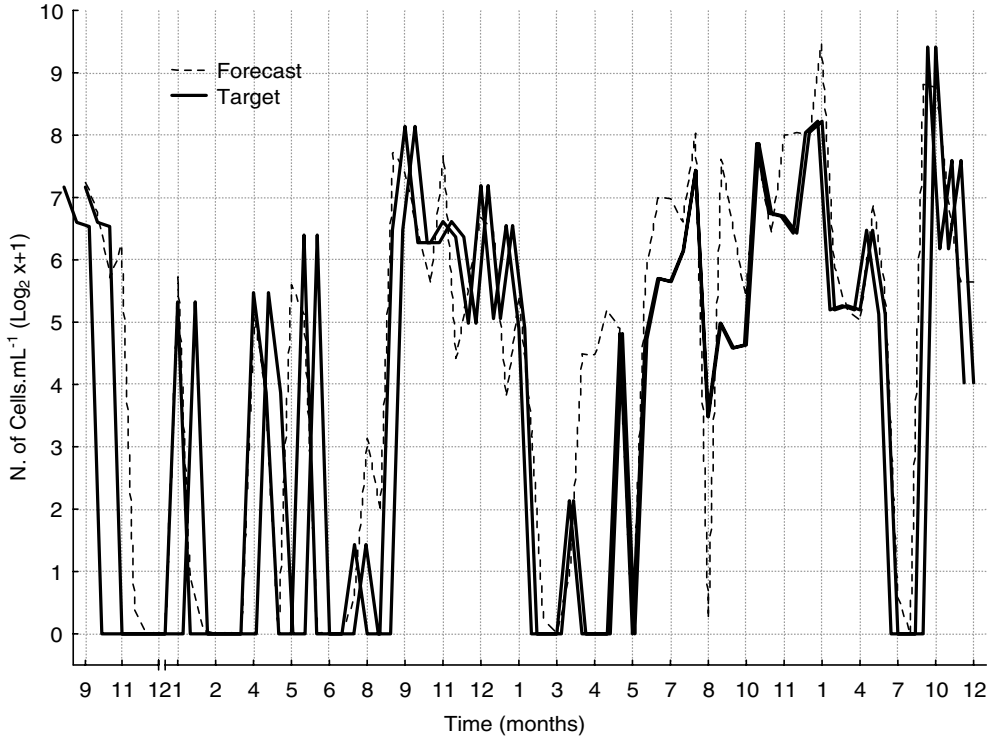


**Figure 3.** Validation (first time set) and verification (second time set) time series used for predicting the abundance of cyanobacteria with the GRNN 2C model.

### Estimation of the Best Time Lags

In a time series forecasting problem, the number of input nodes corresponds to the number of lagged observations used to discover the underlying pattern in a time series and the number of observations required to forecast future values (Zhang and others 1998). Consequently, selection of the best time lag is very important. In this work, the appropriate time lag was selected by using the GRNN with a range of different time lags. There are several methods that can be used to determine the appropriate time lag. However, the NN-based approach is considered to be one of the best (Maier and Dandy 2001). All of the GRNNs used in this study had the same structure and, therefore, the variables included in this NN were chosen based on the results of the multifactorial analysis (Figure 2). A time lag of 8 weeks was selected for both studies (Crestuma 2000–2001 and Crestuma 2000–2002). This time lag resulted in GRNNs with the best correlations between the expected and observed values.

### Search for the Best Smoothing

In time series analysis, the general purpose of smoothing is to expose the major patterns or trends in a data series in the presence of minor fluctuations (random noise) (Statsoft 2000). Visually, smoothing can be used to transform a jagged line pattern into a smooth curve with the aid of smoothing coefficients, which in most cases are between 0.1 and 100. For the training of the GRNNs, smoothing coefficients were chosen that provided the best correlation between the expected and observed values in the verification time series. A smoothing coefficient of 0.18 was selected for both of the studies (Crestuma 2000–2001 and Crestuma 2000–2002).

### Search of the Best Regression Model

A search for the best regression model (GRNN) was carried out by training the NNs with the training series and then searching for the combination of input variables that would produce the best regression in the verification series. With the exception of the input variables, all of the parameters used for training and regression were the same as those indicated in the previous section.

The search for the best variable combination was performed backwards and forwards in a stepwise manner until the best correlation between expected and observed values of the verification time series was obtained. The criteria for inclusion or exclusion of each of the variables was based on their sensitivity in

Table 2. Results obtained for regression analyses of several of the best models obtained using different categories of variables from 2000 to 2001

| | Models | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | GRNN 1A | | | GRNN 1B | | | GRNN 1C | | |
| Variable type | Physical and chemical | | | Phytoplankton taxa | | | All variables | | |
| Input variables | DO pH COND ALK CI $SO_4$ EVP DISCH RAD $NO_3$ TURB Fe TFe OXID | | | CHLOR CRYPT CRYS PHYTO | | | CHLOR DO pH COND ALK CI $SO_4$ EVP DISCH RAD $NO_3$ TURB Fe TFe OXID | | |
| Serial time | Train. | Verif. | Valid. | Train. | Verif. | Valid. | Train. | Verif. | Valid. |
| Data mean | 3.877 | 3.489 | 3.017 | 3.877 | 3.489 | 3.017 | 3.877 | 3.489 | 3.017 |
| Data S.D | 2.955 | 3.033 | 3.625 | 2.955 | 3.033 | 3.625 | 2.955 | 3.033 | 3.625 |
| Error mean | 0.000 | 0.480 | 1.914 | −0.001 | −0.136 | −0.154 | 0.000 | 0.475 | 2.188 |
| Error S.D | 0.170 | 1.469 | 2.988 | 0.433 | 2.885 | 4.425 | 0.170 | 1.484 | 3.136 |
| Abs. E. mean | 0.034 | 0.918 | 2.213 | 0.192 | 2.156 | 3.618 | 0.034 | 0.946 | 2.462 |
| RMS error | 0.169 | 1.532 | 3.457 | 0.429 | 2.860 | 4.267 | 0.169 | 1.545 | 3.731 |
| S.D. ratio | 0.058 | 0.484 | 0.824 | 0.147 | 0.951 | 1.221 | 0.058 | 0.489 | 0.865 |
| Correlation: | | | | | | | | | |
| r | 0.998 | 0.878 | 0.592 | 0.990 | 0.457 | −0.094 | 0.998 | 0.876 | 0.539 |
| N | 48 | 52 | 14 | 48 | 52 | 14 | 48 | 52 | 14 |
| *P* | 0.000 | 0.000 | 0.026 | 0.000 | 0.001 | 0.750 | 0.000 | 0.000 | 0.047 |
| RMS error* | 59.9 | 88.0 | 53.2 | 15.3 | 70.1 | 91.5 | 59.9 | 88.4 | 55.5 |

Train., Verif., and Valid.: training, verification, and validation series, respectively.
*Unlogged values.

Table 3. Results obtained for regression analyses of several of the best models obtained using different categories of variables from 2000 to 2002

| | Models | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | GRNN 2A | | | GRNN 2B | | | GRNN 2C | | |
| Variable type | Physical and chemical | | | Phytoplankton taxa | | | All variables | | |
| Input variable | DO pH COND Tmax $NO_3$ TURB Fe DISCH | | | CHLOR CRYS PHYTO | | | CHLOR Otaxa DO pH COND Tmax $NO_3$ TURB Fe DISCH | | |
| Serial time | Train. | Verif. | Valid. | Train. | Verif. | Valid. | Train. | Verif. | Valid. |
| Data mean | 4.095 | 3.756 | 3.017 | 4.095 | 3.756 | 3.017 | 4.095 | 3.756 | 3.017 |
| Data S.D | 3.008 | 3.071 | 3.625 | 3.008 | 3.071 | 3.625 | 3.008 | 3.071 | 3.625 |
| Error mean | 0.000 | 0.913 | 0.784 | 0.016 | -0.052 | -0.419 | 0.000 | 0.697 | 0.725 |
| Error S.D | 0.002 | 1.999 | 2.455 | 1.073 | 2.713 | 4.097 | 0.000 | 1.904 | 2.332 |
| Abs. E. mean | 0.001 | 1.413 | 1.519 | 0.642 | 2.131 | 3.699 | 0.000 | 1.215 | 1.234 |
| RMS error | 0.002 | 2.183 | 2.492 | 1.064 | 2.693 | 3.970 | 0.000 | 2.014 | 2.361 |
| S.D. ratio | 0.001 | 0.651 | 0.677 | 0.357 | 0.884 | 1.130 | 0.000 | 0.620 | 0.643 |
| Correlation: | | | | | | | | | |
| r | 1.000 | 0.771 | 0.740 | 0.937 | 0.517 | -0.288 | 1.000 | 0.802 | 0.773 |
| N | 61 | 65 | 14 | 61 | 65 | 14 | 61 | 65 | 14 |
| *P* | 0.000 | 0.000 | 0.002 | 0.000 | 0.000 | 0.319 | 0.000 | 0.000 | 0.001 |
| RMS error* | 0.6 | 149.3 | 45.4 | 93.1 | 105.6 | 89.4 | 0.0 | 97.6 | 42.9 |

Train., Verif., and Valid.: training, verification, and validation series, respectively.
*Unlogged values.

the series. Only the variables with the highest sensitivity were included in the analysis. The analysis was carried out using three different subgroups of input variables: (1) all variables, (2) physical and chemical variables only, and (3) phytoplankton variables only. In this last group, we included all identified groups of phytoplankton. During the forward stepwise search, we considered all of the input variables that were excluded during the backward stepwise search, as well as all of the variables that were not selected.

## Results and Discussion

The results obtained for the different time series (training, verification, validation) using the regression models developed for the Crestuma 2000–2201 and Crestuma 2000–2002 data sets are shown in Tables 2 and 3, respectively. The validation results were obtained by training the model with the verification and training series. Tables 2 and 3 also show the variables that were used in each of the models. The models produced using this method were found to be useful for predicting the training series data, with an accuracy of close to 100% (correlation coefficients between 0.937 and 1.000).

For the verification series, the best model obtained with the Crestuma 2000–2001 data series was found to be GRNN 1A. Using this model, a very high correlation was found between the observed and predicted values ($R = 0.878$, $P = 0.000$). In the verification series, the data set employing phytoplankton data only gave the worst results for both the Crestuma 2000–2001 and Crestuma 2000–2002 data series.

Even in the verification series, for the years in which a complete data set was available (Crestuma 2000–2001), the model was not improved by the inclusion of phytoplankton variables in the data set. The models produced using physical and chemical variables only (GRNN 1A) gave better results than those that incorporated both types of variables (GRNN 1C). The GRNN 1A model was also found to be better than the model that included data from the years 2000–2002 (GRNN 2A), despite the longer time period over which the data were collected. From these results, we can conclude that the physical and chemical variables included in the first model had more information content that was relevant for the prediction of the cyanobacterial blooms in the Crestuma Reservoir.

For the Crestuma 2000–2002 data set, the best model obtained for the verification series was found to be GRNN 2C. This model contained the same variables as the best model obtained using only the physical and chemical variables (GRNN 2A), but with the inclusion of two phytoplankton groups (Chlorophytes and ''Other taxa''). These two variables compensated to some extent for the absence of physical and chemical data in the data series, but were not as useful, in terms of performance, as the GRNN 1A model.

In the validation test, the best model was found to be GRNN 2C. This model gave a good correlation between observed and predicted values ($R = 0.773$, $P = 0.001$). In this validation test, the models that used phytoplankton variables only (GRNN 1B and GRNN 2B) were not useful for prediction of cyanobacteria abundance.

The models that incorporated physical and chemical variables, or those that additionally contained phytoplankton data provided better validation results for the Crestuma 2000–2002 data set than for the Crestuma 2000–2001 data set. For the verification results, however, the opposite occurred. From these observations, we can conclude that the two years of data collected during the Crestuma 2000–2002 study were not enough to give the model good ''generalization'' ability. We can also conclude that if the GRNN 1A model, which gave the best verification results (correlation = 0.878), could be trained using the same time period as the Crestuma 2000–2002 series, it would potentially provide even better results. The validation test carried out using the GRNN 2C model (with the Crestuma 2000–2002 data series), gave very good results considering the low number of training years. Clearly, three years is a very short time period in a hydrological time scale. It is also important to note that the data set used for the development of the GRNN 2C model lacked some of the variables that were previously (from the Crestuma 2000–2001 data set) shown to be useful for predicting cyanobacterial blooms.

Most prediction models that are commonly used in ecological studies of phytoplankton dynamics are not able to predict future values (e.g., Bobbin and Recknagel 2001a; Bobbin and Recknagel 2001b; Jeong and others 2001; Scardi, 2001; Wei and others 2001; Whigham and Recknagel 2001a; Whigham and Recknagel 2001b; Karul and others 1999; Karul and others 2000; Scardi and Harding 1999; Whitehead and others 1997). However, only one of these models (Jeong and others 2001) provided a level of accuracy greater than that of the GRNN 2C model, namely, in a study of the phytoplankton dynamics in the Nakdong River (Korea) where an RMS error of less than 0.003 was achieved in the validation series.

Figure 3 shows the time series data obtained for the Crestuma Reservoir from 2000 to 2002, together with the predicted values produced using the GRNN 2C model. A high degree of overlapping can be seen between the observed and predicted values. The model was able to predict all of the density oscillations. It is also important to note that, in most cases, the peaks in cyanobacterial abundance obtained using the model were of similar magnitude to those in the original data set. The time lag was in all cases less than 15 days, which corresponded to the time unit used for data collection.

The sensitivity analysis carried out using the GRNN 2C model showed the relative importance of each of the input variables presented in Figure 4. The results of the verification series indicated that all of the vari-

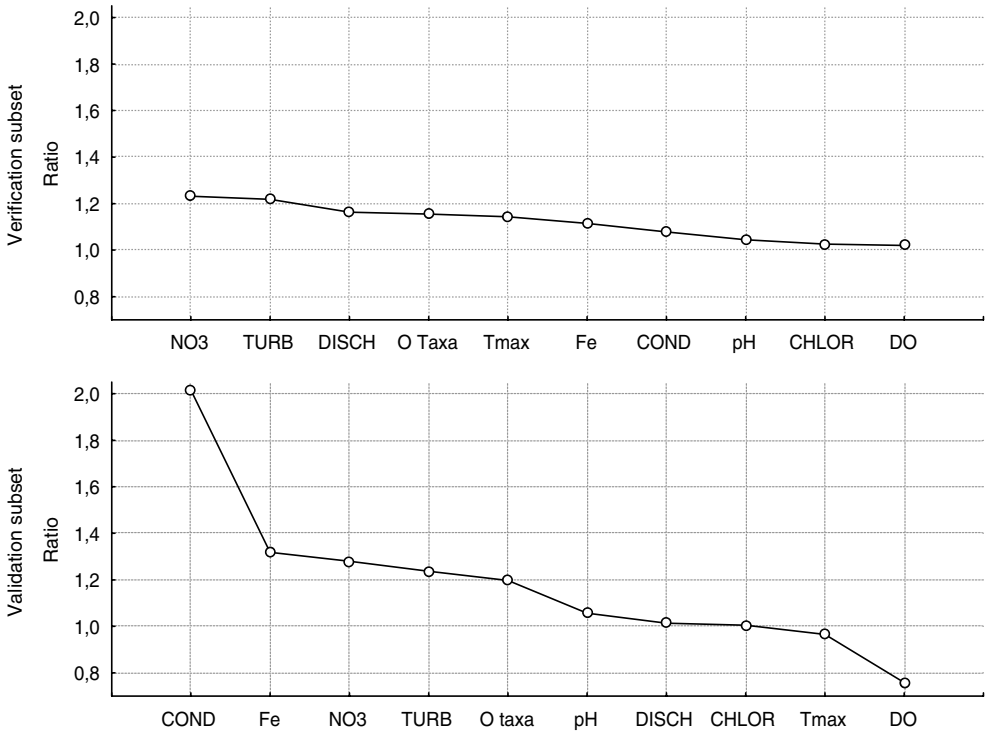


**Figure 4.** Sensitivity analysis carried out using the best trained model (GRNN 2C).

ables included in the model were equally important, and all were characterized by ratios of greater than 1. This shows that the exclusion of any of these variables would significantly reduce the efficiency of the model and, therefore, affect the ability of the model to predict the verification data.

In the validation series, there were clear differences in the relative importance of each of the variables (Figure 4). The ratios of conductivity (COND) and dissolved iron (Fe) increased significantly. In contrast, the ratio of the other variables such as discharge (DISCH), maximum air temperature (Tmax), and dissolved oxygen (DO) decreased. The ratio of the last two variables (Tmax and DO) was low enough that they could be excluded from the model without significantly affecting the predictive efficiency since the ratios of these variables were less than one.

The ecology of freshwater systems is affected, to a large extent, by the hydrological regime (Moss 1998). In many cyanobacteria and algal bloom prediction models based on river systems (Maier and Dandy 2001; Maier and others 2001; Recknagel and others 1997), discharge has been found to be the most important prediction parameters. In other lake-based models, the best prediction parameters have been found to be pH and nutrient concentrations (Recknagel and others 1997; Bobbin and Recknagel 2001; Wei and others 2001). In this study, the importance of ionic substances (COND) and nutrients ($NO_3^-$ and $Fe^{3+}$) was found in the validation test carried out using the GRNN 2C model (Figure 4). These factors, combined with discharge rates indicate that the Crestuma Reservoir is a semi-lentic system.

The density of cyanobacteria during years characterized by summertime blooms and low cyanobacteria density during the winter would suggest that the best time lag might be 12 months. Nevertheless, the best time lag was found to be two months. It is possible that lower time lags increase the sensitivity of the model for detecting secondary fluctuations during summer and winter time periods. In addition, shorter lag periods can be more useful for detecting the conditions that precede cyanobacterial blooms. The two-month lag detected in the data set does not include the annual fluctuation, and we believe that this is explained by the annual variation in physical and chemical variables. Using climatic, hydrological, and limnological data as input variables, Jeong and others (2001) showed that phytoplankton biomass was best predicted with a three-day time lag. Empirical studies have shown that NNs are better at forecasting monthly and quarterly time series than at forecasting yearly data. According to Zhang and others (1998), this is due to the fact that monthly and quarterly data contain more irregularities

(e.g., seasonality, cyclicity, nonlinearity, and noise) than yearly data. NNs are able to efficiently detect the underlying pattern masked by noisy factors in a complex system.

The development of cyanobacteria blooms in freshwater lakes and rivers is known to be highly irregular. It is, therefore, impractical to use prediction models based exclusively on the identification of simple patterns such as in the classical ARIMA (Auto-Regressive Integrated Moving Average) models. As a result, the greatest possible number of covariables should be included in an attempt to explain the underlying patterns. These covariables can be used to complement the information provided in the time series data. In this work, the inclusion of the information regarding cyanobacterial density did not contribute to a better performance of the predictive models. Surprisingly, the best predictive model that was obtained in the verification series (GRNN 1A) did not include any information related to phytoplankton abundance. The results of this study indicate that a good knowledge of the physical and chemical characteristics of the environment where cyanobacteria grow provides the best prediction of their abundance.

The differences between the errors obtained for the training series and the validation/verification series (Tables 2 and 3) suggest that the regression model could be improved with the inclusion of data obtained from other years.

## Conclusions

The accurate prediction of fluctuations in the abundance of cyanobacteria in a potable water supply (Crestuma Reservoir) was carried out using data collected over a three-year period. This prediction was performed without incorporating information on the abundance of cyanobacteria in the data input. Physical and chemical characteristics of the reservoir were found to be the best predictive variables of cyanobacteria abundance. Application of an eight-week time lag allowed us to detect secondary fluctuations in cyanobacteria density over the annual cycle. Although the Crestuma Reservoir is characterised by a high flow, it displays semi-lentic characteristics. The results obtained with the final prediction model are very promising. However, the inclusion of a greater number of variables and collection of data over an extended time period would undoubtedly improve the accuracy of the model.

## Literature Cited

Bishop, C. 1995. Neural networks for pattern recognition. Oxford University Press, Oxford, 482 pp.

Bobbin, J., and F. Recknagel. 2001a. Inducing explanatory rules for the prediction of algal blooms by genetic algorithms. *Environmental International* 27:237–242.

Bobbin, J., and F. Recknagel. 2001b. Knowledge discovery for prediction and explanation of blue-green algal dynamics in lakes by evolutionary algorithms. *Ecological Modelling* 146:253–262.

Ferreira, F. M. B., J. M. F. Soler, M. L. Fidalgo, and P. Fernández-Vila. 2001. PSP toxins from *Aphanizomenon flos-aquae* (cyanobacteria) collected in the Crestuma-Lever reservoir (Douro river, northern Portugal). *Toxicon* 39(6):757–761.

French, M., and F. Recknagel.. 1994. Modelling of algal blooms in freshwaters using artificial neural networks. Pages 87–94 *in* Zaneti P (ed.), Computer techniques in environmental studies V. Environmental systems. Vio. II. Computational Mechanics Publications, Boston.

Haykin, S. 1999. Neural networks: a comprehensive foundation. 2nd edition. Prentice-Hall, Inc., New Jersey, 842 pp.

Hornik K. 1991. Approximation capability of multilayer feedforward networks. *Neural Networks* 4:251–257.

Jeong, K.-S., 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(1–3):115–129.

Karul, C., S. Soyupak, and C. Yurteri. 1999. Neural network models as a management tool in lakes. *Hydrobiologia* 408/409:139–144.

Karul, C., S. Soyupak, A. F. Çilesiz, N. Akbay, and E. Germen. 2000. Case studies on the use of neural networks in eutrophication modelling. *Ecological Modelling* 134:145–152.

Maier, H. G., and G. C. Dandy. 1998. The effect of internal parameters and geometry on the performance of back-propagation neural networks: an empirical study. *Environmental Modelling and Software* 13:193–209.

Maier, H. G., 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. G., and G. C. Dandy. 2001. Neural network based modelling of environmental variables: a systematic approach. *Mathematical and Computer Modelling* 33:669–682.

Maier, H. G., G. C. Dandy, and M. Burch. 1998. Use of artificial networks for modelling cyanobacteria *Anabaena* spp. in the River Murray, South Australia. *Ecological Modelling* 105:257–272.

Maier, H. G., T. Sayed, and B. J. Lence. 2001. Forecasting cyanobacterium *Anabaena* spp. in the River Murray, South Australia, using B-spline neurofuzzy models. *Ecological Modelling* 146:85–96.

Moss, B. 1998. Ecology of fresh waters: Man and medium, past and future. 3rd edition. Blackwell Science, Oxford, 557 pp.

Patterson, D 1996. Artificial neural networks: Theory and applications. Prentice Hall, Singapore.

Qi, M., and G. P. Zhang. 2001. An investigation of model selection criteria for neural network time series forecasting. *European Journal of Operational Research* 132(3):666–680.

Recknagel, F. 2001. Applications of machine learning to ecological modelling. *Ecological Modelling* 146:303–310.

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

Scardi, M. 2001. Advances in neural network modelling of phytoplankton primary production. *Ecological Modelling* 146:33–45.

Scardi, M., and L. Harding. 1999. Developing an empirical model of phytoplankton primary production: a neural network case study. *Ecological Modelling* 120:213–223.

Schultz, A., R. Wieland, and G. Lutze. 2000. Neural networks in agroecological modelling: stylish application or helpful tool? *Computers and Electronics in Agriculture* 29:73–97.

Speckt, D. F. 1991. A generalized regression neural network. *IEEE Transactions on Neural Networks* 2(6):568–576.

Statsoft. 2000. STATISTICA Neural Networks (Release 4.0 F), Tulsa, OK.

Statsoft. 2001. STATISTICA (Version 6) Tulsa, OK.

Vasconcelos, V. M., W. W. Evans, W. Carmichael, and M. Namikoshi. 1993. Isolation of microcystin-LR from a *Microcystis* (Cyanobacteria) bloom collected in the drinking water reservoir for Porto, Portugal. *Journal of Environmental Science and Health* 28(9):2081–2094.

Wei, B., N. Sugiura, and T. Maekawa. 2001. Use of artificial neural network in the prediction of algal blooms. *Water Research* 35(8):2022–2028.

Whigham, P. A., and F. Recknagel. 2001a. Predicting chlorophyll-*a* in freshwater lakes by hybridising process-based models and genetic algorithms. *Ecological Modelling* 146:243–251.

Whigham, P. A., and F. Recknagel. 2001b. An inductive approach to ecological time series modelling by evolutionary computation. *Ecological Modelling* 146:275–287.

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.

Yun, S.-Y., S. Namkoong, J.-H. Rho, S.-W. Shin, and J.-U. Choi. 1998. A perfomance evaluation of neural network models in traffic volume forecasting. *Mathematical and Computer Modelling* 27(9–11):293–310.

Zhang, G. P., and M. Y. Hu. 1998. Neural network forecasting of the British pound/US dollar exchange rate. *Omega, International Journal of Management Science* 26(4):495–506.

Zhang, G. P., B. E. Patuwo, and M. Y. Hu. 1998. Forecasting with artificial networks: The state of the art. *International Journal of forecasting* 14:35–62.

Zhang, G. P., B. E. Patuwo, and M. Y. Hu. 2001. A simulation study of artificial neural networks for nonlinear time-series forecasting. *Computers and Operation Research* 28:381–396.