9.09 Artificial Neural Network Modeling of Phytoplankton Blooms and its Application to Sampling Sites within the Same Estuary

H-Y Kang and RA Rule, University of Washington, Seattle, WA, USA
PA Noble, Alabama State University, Montgomery, AL, USA

Abstract

Artificial neural networks (ANNs) are useful tools for modeling complex ecosystems because they can predict how ecosystems respond to changes in environmental variables (e.g., nutrient inputs). In addition, ANNs can be used to discover relationships among variables, which aids in the understanding of ecosystem function. ANN models were used to predict phytoplankton blooms in three different sites within the same salt marsh estuary located in South Carolina. We (1) compared ANN models with different architectures, (2) applied sensitivity analysis to identify the importance of input variables, and (3) compared the results from ANN modeling to those obtained using linear models.

9.09.1 Introduction

Artificial neural networks (ANNs) are useful tools for recognizing patterns in complex, nonlinear data sets such as those associated with ecological and biological data as demonstrated by the many articles published in *Ecological Modelling* and the *Journal of Microbiological Methods* (see Almeida and Noble, 2000). They are advantageous over conventional (linear-based) statistical methods because ANNs can deal with the inherent variability associated with biological data, and therefore better recognize patterns in data and make improved predictions of response variables than conventional methods.

9.09.1.1 What Are ANNs?

ANNs are computer programs that consist of networks of neurons that receive information from scaled input variables or other neurons, make independent computations, and pass their output to other neurons in a network (Noble and Tribou, 2007). Each neuron in a network is an autonomous entity composed of weights (that are used to weigh the value of the data received), a bias term (that prevents divisions by zero), and a transfer function (that passes the value of the neuron forward to the next neuron in the network) (**Figure 1**). The reason that scaling of the variables is necessary is because it ensures that the values are within an appropriate range for the transfer function (e.g., for

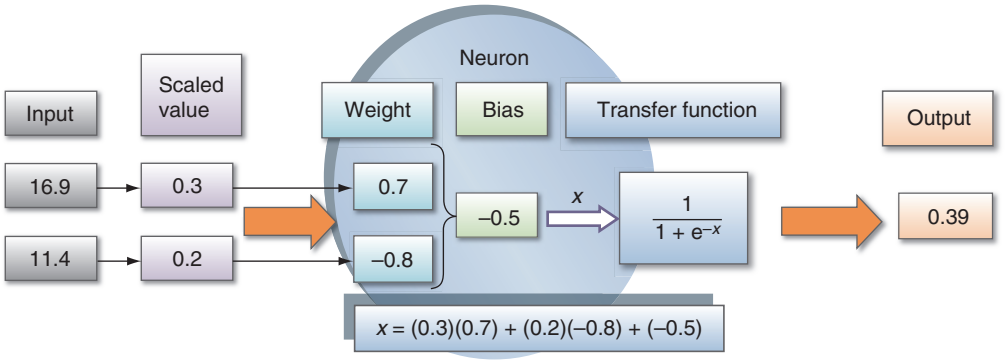

Figure 1 Computation of a single neuron. A neuron contains a weight term for each input variable, a bias term, and a transfer function. It receives input data and then computes output data.

the log-sigmoid transfer function, the value of x must be between 0 and +1). Two scaling methods are typically used: (1) the input and output values are set to a maximum of one and minimum of 0 (or −1) or (2) the input and output values are standardized to a mean of 0 and standard deviation of 1. As shown in Figure 1, the scaled inputs are multiplied by corresponding weights, the product is added to the bias, and then put into a transfer function, such as the log-sigmoid function:

$$S(x) = \frac{1}{(1 + e^{-x})} \quad [1]$$

The final output neuron is computed using the equation

$$H_j = S\left[c_j + \sum_{i=1}^{i=n}(a_{ij}x_i)\right] \quad [2]$$

where H_j is the output of j neurons, a_{ij} the weight of each scaled input variables, c_j the bias term of j neuron, n the number of input variables, and x_i the scaled input variables. In Figure 1, the computed output is $1/(1+e^{-[(0.3)(0.7)+(0.2)(-0.8)+(-0.5)]})$ or 0.39.

A neural network consists of many neurons that are organized into an input, a hidden, and an output layer. Figure 2 shows a simple ANN structure having three inputs (one for each variable), one hidden layer with three neurons, and one output neuron. The value of the output layer is computed as

$$Y_k = d_k + \sum_{i=1}^{nH}(b_{jk}H_j) \quad [3]$$

where Y_k is the normalized output variables, b_{jk} the weight of each neurons, d_k the bias term of k output, and n_H the number of neurons. Whereas the number of neurons in the input and output layers is fixed by the data, the number of neurons in the hidden layer is defined by the user, and typically ranges from one to the total number of input variables. In this study, the number of hidden neurons was calculated by experimentally determining the minimum number of hidden neurons needed to yield the highest R^2 for the linear relationship between predicted and actual outputs of the response variable.

9.09.1.2 Training and Testing of an ANN Model

Once the architecture has been determined, the ANN is trained using a set of known inputs and outputs and the weights and bias for each neuron in a network are adjusted so that they collectively learn the relationships based on the training set. The adjustment of the weights and biases is accomplished by minimizing the error between the predicted output and the actual response variable using an error function, such as Levenberg–Marquardt. The adjusted weights and biases can then be used to recalculate the predicted output of other data sets. The iterative process of adjusting (and readjusting) the

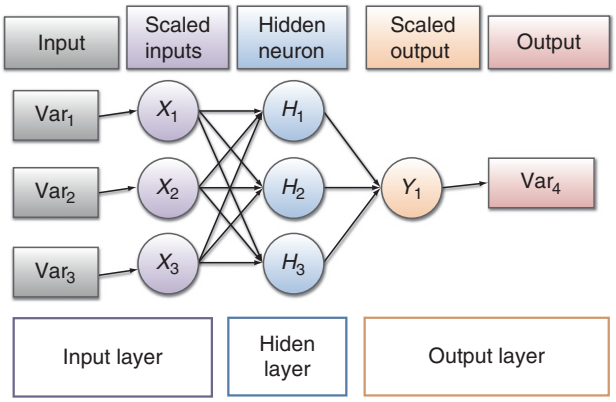

Figure 2 The architecture of a simple artificial neural network (ANN). This ANN contains three input variables, one hidden layer containing three neurons, and one output layer with one output variable. A complex ANN contains more input variables and more hidden and outer neurons than presented in this example.

weights and biases is referred to as error back-propagation (Rumelhart et al., 1986; Bishop, 1995). Typically, back-propagation (followed by forward propagation) continues until a global error minimum is attained. Once an ANN has been properly trained, the adjusted weights and biases can be used to provide information on the functional relationships among variables.

The data set used by the ANN is typically divided randomly into two subsets: one subset is used for training and testing of the ANN model and the other subset is used for validating the ANN model. Detailed discussions of ANN models can be found in the following articles: Basheer and Hajmeer (2000), Bishop (1995), Hagan (1995), Noble and Tribou (2007), Smith (1996), and Rao and Rao (1993).

9.09.1.3 ANN Models for Predicting and Understanding Harmful Algal Blooms

It is well established that excessive algal growth can be harmful to aquatic ecosystems because algae discolor the water, cause the depletion of dissolved oxygen, and are responsible for fish kills. The causality and dynamics of algal blooms are very complex and not entirely understood, particularly in estuarine ecosystems. Although deterministic models based on physical–biological data have contributed to a better understanding of plankton food webs (e.g., Baird et al., 2001), their predictive potential still remains relatively uncertain (Sarkar and Chattopadhayay, 2003). The development of an ANN model that can accurately predict the occurrence of phytoplankton blooms is highly desired by environmental engineers and managers who administer wildlife and water resources (Maier and Dandy, 2000; Lee et al., 2003). Moreover, an understanding of the mechanisms promoting or preventing phytoplankton blooms, such as red tides, could lead to better ways to control and/or minimize the effects of excessive algal growth (Sarkar and Chattopadhayay, 2003; Goethals et al., 2007). Identifying the ecological variables involved in these mechanisms will facilitate the development of broad hypotheses underlying our understanding of eutrophication and harmful algal blooms (Smayda, 1997; Rabouille et al., 2001).

This study focuses on the prediction of phytoplankton biomass using long-term ecological research (LTER) data sets and an ANN modeling approach. The objectives were, first, to present an ANN modeling approach, which can accurately predict the response (output) variable (chlorophyll concentration) and, second, to apply the model to other ecological sites in order to assess its overall performance. Specifically, we will (1) compare the prediction results of ANN models having different network architectures, (2) apply sensitivity analysis to identify the relative importance of input variables, and (3) compare the modeling results to those obtained using conventional linear regression models. Sensitivity analysis will be used to determine which of the input variables contribute most to predictions. We will demonstrate that the selection of input variables based on sensitivity analysis is essential for improving the accuracy of ANN model predictions.

9.09.2 Materials and Methods

9.09.2.1 Source Data

The LTER data sets were downloaded from the Baruch Institute Data Archives. The data from the following three sites within the North Inlet (South Carolina) estuary were used in this study: Oyster Landing (OL), Town Creek (TC), and Clambank (CB) (**Figure 3**). The TC site is located at the mouth, west of North Inlet (NI) estuary and is more strongly influenced by the coastal zone than the other sites. The OL site is farthest from

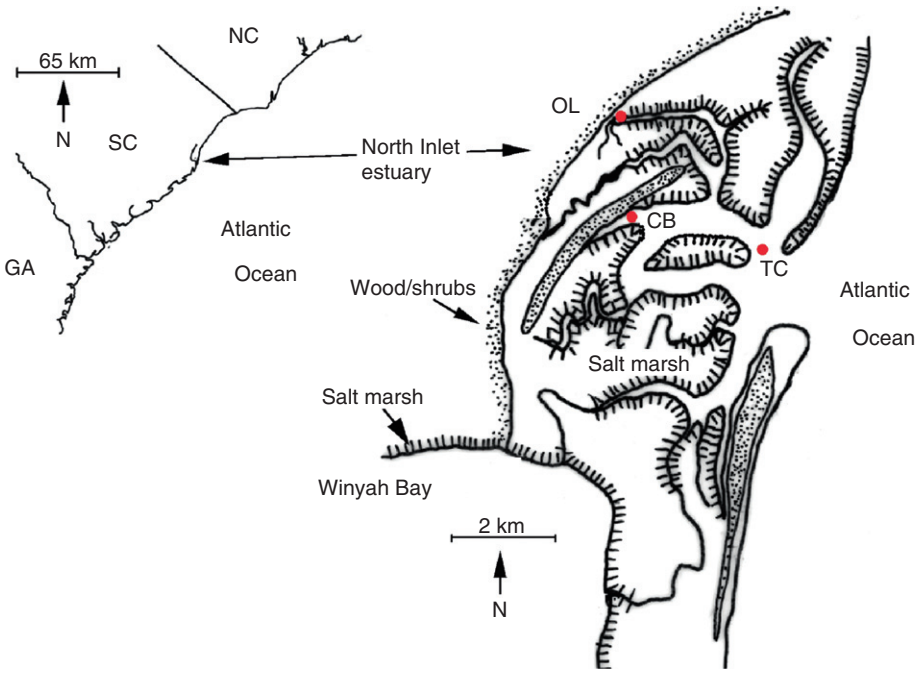

Figure 3 Map of North Inlet estuary, South Carolina. Red dots indicate sampling sites. Station 1, Oyster Landing; Station 2, ClamBank; Station 3, Town Creek.

the ocean and is influenced by the salt marsh and tidal mixing. The CB site is located between OL and TC sites and is closest to Winyah Bay, which makes it sensitive to fluctuations of salinity and suspended sediment (Gardner et al., 1989).

The LTER data sets consist of daily monitoring samples collected from 30 December 1982 to 24 March 1989 at approximately 10.00 a.m. There were 17 variables in the LTER archive (**Table 1**). In this study, chlorophyll *a* (CHLA) concentration was used as an indicator of phytoplankton biomass and the ANNs were trained to predict CHLA using the other ecosystem variables. The tide data were obtained from the National Oceanic and Atmospheric Administration (NOAA) for OL, NI estuary. **Table 1** also shows the accuracy of the variables obtained from the archives.

We excluded daily samples having numerous missing values because they might interfere with training the ANN models. In order to ensure that the ANN would not be affected by local patterns of the data, the chronological daily data were randomized. The randomized data set consisted of 2261 daily sampling points for the OL site. This data set was separated into two subsets: one for training and testing ($n = 1134$ samples) of the ANN model and the other for validating the ANN model ($n = 1127$ samples). It is important to recognize that the number of data points used for each analysis is dependent on variable input (i.e., some variables had missing data).

The data from the CB ($n = 1912$) and TC ($n = 3378$) sampling sites were used to assess whether the ANN models developed for the OL data set could accurately predict CHLA.

9.09.2.2 ANN Modeling

ANN modeling was conducted using the Neural Nets tool in SAS JMP® v8.0. The random holdback cross-validation method was used, with 50% of the data used for training and the remainder used for testing the model. In this software package, the user defines how many hidden neurons are used in the ANN model. We used a range of hidden neurons to develop models (i.e., from one to the number of input variables) with the goal of finding the minimum number of hidden neurons needed to yield the highest R^2 for the linear relationship between predicted and actual outputs. We would expect a one-to-one relationship for predicted and actual outputs (i.e., $R^2 = 1$). The software package automatically normalizes the input and output variables and generates an R^2 of the predicted and actual outputs by using standard nonlinear least-squares regression methods of the combined training and testing data sets. The SAS software also provides the weights and bias term of each neuron, so that a user can extract this information and develop a standalone model (i.e., the equations) in a spreadsheet program (i.e., Microsoft Excel).

After ANN modeling, the weight and bias terms of each neuron were extracted and the equations were rebuilt in a spreadsheet (i.e., Microsoft® Excel). In a spreadsheet, we standardized the data to a mean of 0 and standard deviation of 1. We first reconfirmed the equation with the R^2 of training and testing obtained from JMP® using the training data set and then computed the R^2 using the validation data set.

Table 1 Summary of input variables from Baruch Institute Data Archives

Abbreviation	*Name and accuracy value*
TNW	Total nitrogen whole $\pm$ 1 µmol^{-1}
TNF	Total nitrogen filtered $\pm$ 1 µmol^{-1}
TPW	Total phosphorus whole $\pm$ 1 µmol^{-1}
TPF	Total phosphorus filtered $\pm$ 1 µmol^{-1}
OP	Ortho phosphate $\pm$ 0.1 µmol^{-1}
NH4	Ammonia $\pm$ 0.1 µmol^{-1}
NN	Nitrate–nitrite $\pm$ 0.1 µmol^{-1}
DOC	Dissolved organic carbon $\pm$ 0.1–1 mg l^{-1}
TSS	Total suspended sediments $\pm$ 0.001 mg l^{-1}
ISS	Inorganic suspended sediments $\pm$ 0.001 mg l^{-1}
OSS	Organic suspended sediments $\pm$ 0.001 mg l^{-1}
SAL	Salinity $\pm$ 2 parts per thousand
TIDE	Tide elevation
WTEMP	Water temperature $\pm$ 1° C
SECCHI	Secchi $\pm$ 0.1 m
PHAEO	Phaeophytin $\pm$ 0.1 µg l^{-1}
CHLA	Chlorophyll a $\pm$ 0.1 µg l^{-1}

9.09.2.3 Sensitivity Analysis

Sensitivity analysis was used to identify which of the input variables most contributed to chlorophyll predictions (Noble et al., 2000). Once these variables were identified, the objective was to determine if ANN models developed using these input variables improve the predictability of chlorophyll concentration over ANN models developed using all input variables. In other words, do highly sensitive input variables improve chlorophyll concentration predictability?

The sensitivity of output variable (Y^*) was determined by putting the minimum or maximum values of one input variable into the ANN model while keeping all other input variables to their mean values. The input and output variables were standardized, denoted by an asterisk, to a mean of 0 and standard deviation of 1. $Y^*(0)$ denotes the output variable of all input variable by their mean values. $Y^*(\max)$ denotes the output variable of one input variable of interest at its maximum value. $Y^*(\min)$ denotes the output variable of the input variable of interest at its minimum value. The difference of output Y^* from its mean value $Y^*(0)$ reflects how sensitive the input variable is. The percentage of relative sensitivity was calculated by the sensitivity of the input variable of interest divided by the sum of the absolute sensitivity of all the input variables, that is, $(Y^* - Y^*(0))_i / \Sigma \text{ absolute } (Y^* - Y^*(0))_i$.

9.09.3 Results

9.09.3.1 Modeling Chlorophyll Concentration

The prediction potential of an ANN model was determined by the R^2 of the linear regression between actual and predicted chlorophyll concentrations (CHLA). Our analyses produced two R^2 values, one representing the R^2 determined using the combined training and testing data sets, henceforth referred to as R^2_{tv} and the other representing the R^2 obtained using a second data set, henceforth referred to as R^2_t. The R^2_t represents the true predictability of the ANN model because the second data set was not used in developing the model.

9.09.3.1.1 Prediction of chlorophyll concentration using physical and chemical data

To determine if physical and chemical input variables alone could predict chlorophyll concentration (i.e., excluding input variables that were directly linked to the biology; e.g., SECCHI), the following input variables were used: WTEMP, SAL, TIDE, OP, NH4, and NN. As shown in Table 2, water temperature, salinity, and tide level (trials #1 and #2) provided better predictions of chlorophyll concentration than orthophosphate, ammonia, and nitrite/nitrate (trial #3) (R^2_{tt} of 0.70 vs. 0.25; R^2_t of 0.68 vs. 0.18).

9.09.3.1.2 Improvement of ANN performance using input time delays

Due to biological processing (i.e., the time needed for algae to respond to changing environmental conditions), an input variable might have a delayed effect on a response variable (e.g., chlorophyll). To determine if any of the input variables came under this category, we systematically assessed the effects of time delays on chlorophyll concentration for all variables. We rationalized that the ANN model predictions of the response variable might be improved by using values from the previous sampling day rather than the value of an input variable from the same sampling day. This was accomplished by considering the correlations between individual input variables and chlorophyll concentration collected on the same day (CHLA), on the following day (CHLA+1), and on the next following day (CHLA+2).

For most input variables, the highest correlation was obtained for variables collected on the same day as the response variable (CHLA) (Table 3). For example, total nitrogen whole (TNW) was more correlated to chlorophyll concentration (CHLA) obtained on the same sampling day ($r = 0.66$) than chlorophyll concentration on the following day (CHLA+1, $r = 0.59$), or the next following day (CHLA+2, $r = 0.52$). However, for input variables NH4, NN, and SAL, the correlations slightly improved using 1- and 2-day time delays (see Table 3). These findings are consistent with the notion that phytoplankton (as represented by chlorophyll concentration) required time to respond to changes in ammonia, nitrate/nitrite concentrations, and salinity in the environment.

We tested the effects of time delays for ammonia and nitrate/nitrite concentration on ANN model predictions of chlorophyll concentration. The results revealed no significant

Table 2 ANN modeling of chlorophyll a using physical, chemical, and/or ecological inputs (see Table 1 for abbreviations/acronyms)

Input variables	# Hidden Neurons	Output	R_{tt}^2	R_t^2	Trial #
WTEMP, SAL, TIDE, OP, NH4, NN	3	CHLA	0.71	0.68	1
	4		0.71	0.66	
	5		0.73	0.67	
	6		0.75	0.64	
WTEMP, SAL, TIDE	3	CHLA	0.70	0.68	2
OP,NH4,NN	3	CHLA	0.25	0.18	3
WTEMP,SAL,TIDE,OP,NH4-2,NN-2	3	CHLA	0.72	0.66	4
	4		0.73	0.66	
	5		0.74	0.66	
	6		0.75	0.64	
WTEMP, SAL, TIDE, OP, NH4, NN, TPF, TNF, ISS	3	CHLA	0.77	0.68	5
	4		0.76	0.67	
	5		0.78	0.68	
	6		0.80	0.66	
	7		0.80	0.68	
	8		0.83	0.67	
	9		0.83	0.66	
CHLA–1, WTEMP, SAL, TIDE,	3	CHLA	0.81	0.80	6
	4		0.82	0.81	
CHLA–1, WTEMP, SAL, TIDE, OP, NH4, NN	3	CHLA	0.82	0.79	7
	4		0.84	0.78	
	5		0.83	0.78	
	6		0.84	0.76	
	7		0.84	0.75	
CHLA–1, WTEMP, SAL, TIDE, OP, NH4, NN, TPF, TNF, ISS	3	CHLA	0.84	0.78	8
	4		0.85	0.77	
	5		0.87	0.75	
	6		0.86	0.77	
	7		0.87	0.75	
	8		0.87	0.77	
	9		0.89	0.74	
	10		0.89	0.76	

Table 3 Pearson correlation coefficients between an input variable and chlorophyll concentration ($n > 2000$ points)

Input	*CHLA*	*CHLA+1*	*CHLA+2*
TNW	0.66	0.59	0.52
TPW	0.71	0.62	0.53
DOC	0.08	0.06	0.01
TSS	0.55	0.45	0.35
OSS	0.47	0.40	0.32
SECCHI	0.59	0.55	0.50
PHAEO	0.74	0.61	0.50
OP	0.37	0.36	0.33
NH4	0.34	**0.36**	**0.35**
NN	0.18	**0.19**	**0.20**
SAL	0.02	**0.03**	**0.05**
TIDE	0.06	0.02	0.04
WTEMP	0.74	0.73	0.72
TNF	0.39	0.38	0.34
TPF	0.38	0.37	0.33
ISS	0.55	0.45	0.36
CHLA-2	0.78	0.69	0.62
CHLA-1	0.79	0.77	0.68
CHLA	–	0.87	0.77
CHLA+1	–	–	0.68
CHLA+2	–	–	–

Note that CHLA+1 denotes the value of CHLA on the following day, CHLA+2 denoted the value of CHLA on the second following day, and CHLA–1 denoted the value of CHLA on previous day.
Bold, denotes improved correlation (see Table 1 for abbreviations/acronyms).

improvement in the prediction of chlorophyll concentration. For example, comparison of trial #4 to trial #1 showed that the highest $R^2_{tt} = 0.75$ occurred in both trials, while the highest $R^2_t = 0.66$ for trial #4 and $R^2_t = 0.68$ for trial #1 (Table 2). These findings suggest that although correlation analysis showed that phytoplankton needed time to respond to ammonia and nitrate/nitrite, the time delay did not significantly improve the prediction of chlorophyll concentration.

It should be noted that water temperature has a higher correlation to chlorophyll concentration ($r = 0.74$) than all the other input variables (Table 3). In the following, sensitivity analysis will show that this variable was an important predictor of chlorophyll concentration.

We tried to improve the response predictions by including three additional input variables that were thought to influence chlorophyll concentration: total filtered nitrogen, total filtered phosphorus, and inorganic suspended sediments. As shown in trials #1 (WTEMP, SAL, TIDE, OP, NH4, and NN) and #5 (WTEMP to NN, TPF, TNF, and ISS) (Table 3), there was a slight increase in R^2_{tt} (from 0.75 to 0.83) but no difference in R^2_t (~0.68). These results suggest that the inclusion of total filtered nitrogen, phosphorus, and inorganic suspended sediments did not significantly improve the prediction of chlorophyll concentration.

9.09.3.1.3 Prediction of chlorophyll concentration using physical and chemical data and the chlorophyll concentration from the previous day

Melesse et al. (2008) suggested that the inclusion of chlorophyll concentration from the previous sampling day would significantly improve the prediction of chlorophyll concentration as an input variable to the ANN model. We tested this hypothesis by including chlorophyll concentration (CHLA–1) from the previous day as an additional input to the ANN model.

Comparison of trial #6 to trial #2 (Table 2) revealed that both R^2_{tt} and R^2_t significantly increased when chlorophyll concentration from the previous sampling day was included as an input variable. Similarly, comparison of trial #7 to trial #1 and trial #8 to trial #5 showed an improvement of more than 10% in the R^2_t of the linear regressions. These findings support the hypothesis proposed by Melesse et al. (2008).

9.09.3.1.4 Optimization of the number of hidden neurons

The effects of hidden neurons on model performance could be gleaned from Table 2. If the difference in the values of R^2_{tt} and R^2_t for the same model was small, then the number of hidden neurons used in the model was optimal. However, if the difference was large, then the ANN model was over- or underfitting the data and the number of hidden neurons in the model was not optimal. As shown in Table 2, the ANN model (trial #8), which had 10 input variables and 10 hidden neurons, yielded $R^2_{tt} = 0.89$ and $R^2_t = 0.76$, a difference of more than 10%. In contrast, the difference between R^2_{tt} and R^2_t for trial #6 was less than 1%, indicating that the number of hidden neurons in that model was optimal. These results suggest that the increased number of hidden neurons (i.e., >5 hidden neurons) in trial #8 resulted in overfitting of the data and that the optimal number of hidden neurons for the ANN model was 3.

In summary, (1) increasing number of input variables or the number of hidden neurons generally increased R^2_{tt} but resulted in overfitting of the data because there was no or little change in the R^2_t of the ANN models; (2) the ANN models that included input variables that were highly correlated to the output variable usually yielded a better prediction of CHLA than poorly correlated variables – hence, the correlation between input variables and the response variable (CHLA) might be useful for the selection of input variables; and (3) the best chlorophyll prediction model consisted of four input variables (e.g., CHLA-1, WTEMP, SAL, and TIDE) and four hidden neurons ($R^2_{tt} = 0.82$ and $R^2_t = 0.81$; trial #6).

9.09.3.2 Sensitivity Analysis

Sensitivity analysis identified which of the input variables most contributed to the prediction of chlorophyll concentration. We reasoned that the most sensitive input variables could be used to develop ANN models with better predictability than the previous ANN models, which used all nine input variables. Assessment of the predictability of a model was determined by the R^2 of the linear regression between actual and predicted chlorophyll concentration. For these experiments, sensitivity analysis was conducted using three hidden neurons and the same variables as those used in trial #8 (Table 2).

This analysis method involved changing one variable at a time to see what effect this has on the predicted output. Typically, the change in the predicted output is determined using the maximum and minimum values of an input variable, while keeping all other input variables constant (i.e., the average). Unlike other studies, we considered the change in the output using the minimum to average value of an input

variable and the maximum to average value of an input variable, in addition to the conventional approach of using the maximum and minimum values of an input variable only. As shown below, setting limits on the value of the input variable of interest (i.e., minimum value to average or maximum value to average) sheds more information on the effects of input values on output predictions than using the maximum and minimum values of an input variable alone (i.e., the conventional approach).

Figure 4 shows three colored bars for each input variable. The blue bars represent the change in the output variable (chlorophyll concentration) relative to minimum values of each input variable, the red bars represent the change in the output variable relative to maximum values of each input variable, and the green bars represent the change in the output variable over the entire range of values for each input variable.

When the minimum to average values of the input variables were considered, the following variables were found to reflect a low chlorophyll concentration: low water temperature, low total nitrogen filtered, low inorganic suspended solids, and high salinity. These results are consistent with environmental conditions in the winter season at NI estuary.

When the maximum to average values of the input variables were considered, the following variables were found to reflect a high chlorophyll concentration: high chlorophyll concentration (from the previous sampling day), high water temperature, high orthophosphate, high nitrate–nitrite, and high inorganic suspended sediments. These conditions are consistent with the late spring–early summer season at NI estuary. Hence, the results obtained from sensitivity analysis are highly dependent on the range of values used as inputs.

The conventional sensitivity analysis (i.e., based on the minimum and maximum values as inputs) revealed much less information than the analysis conducted using either the minimum or the maximum input values. If we had only used conventional sensitivity analysis, we would not have known that low water temperature and low total nitrogen were very predictive of low chlorophyll concentrations, or that high chlorophyll concentration (from the previous sampling day), high water temperature, high orthophosphate, high nitrate–nitrite, and high inorganic suspended sediments were all indicative of high chlorophyll concentration. Nonetheless, the conventional approach to sensitivity analysis did reveal that the chlorophyll concentration collected from the previous day, the water temperature, and the inorganic suspended solids were the main input variables affecting chlorophyll concentration.

Comparison of the rank order obtained from sensitivity analysis (Table 4) with that obtained from the correlation of the input variable and the output variable (from Table 3) revealed that input variables highly correlated to chlorophyll

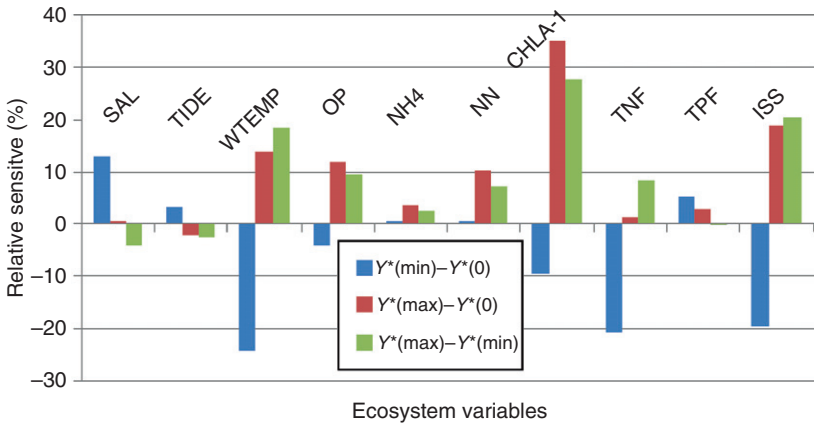

Figure 4 Sensitivity analysis of the variables used to predict chlorophyll concentration. Results indicate that the previous-day chlorophyll concentration (CHLA–1), the amount of inorganic suspended solids (ISS) in the water column, and the water temperature (WTEMP) are the top three variables for predicting chlorophyll concentration (see Table 1 for abbreviations/acronyms).

Table 4 Sensitivity analysis of an ANN model of chlorophyll using 10 inputs and 3 hidden neurons (trial #8, Table 2) (see Table 1 for abbreviations/acronyms)

Rank (high to low)	*Y*(min) – Y*(0)*	*Y*(max) – Y*(0)*	*Y*(max) – Y*(min)*	*Combined rank*
1	WTEMP	CHLA-1	CHLA-1	CHLA-1
2	TNF	ISS	ISS	WTEMP
3	ISS	WTEMP	WTEMP	ISS
4	SAL	OP	OP	TNF
5	CHLA-1	NN	TNF	OP
6	TPF	NH4	NN	SAL
7	OP	TPF	SAL	NN
8	TIDE	TIDE	TIDE	TPF
9	NN	TNF	NH4	NH4
10	NH4	SAL	TPF	TIDE

concentration also ranked high in the sensitivity analysis (e.g., CHLA–1, WTEMP, ISS, and TNF). However, this consistency was not apparent for the lower ranked variables, such as OP, SAL, NN, TPF, NH4, and TIDE.

9.09.3.2.1 ANN models based on input variables with high sensitivity

Input variables for constructing new ANN models were selected according to their sensitivity ranking. The four highest-ranked input variables are shown in the top three rows of Table 4: CHLA–1, WTEMP, ISS, and TNF. For OL samples (Table 5), the ANN model for CHLA–1 and WTEMP (trial #9) yielded high R^2_{tt} (0.80) and R^2_t (0.79), as did CHLA-1 and ISS (trial #10): R^2_{tt} (0.79) and R^2_t (0.75).

As more input variables were included in the ANN models (trials #11–#12), there was a slight increase in R^2_{tt} but little change in R^2_t, indicating that the addition of more variables did not significantly improve the ANN model. The best prediction model was obtained using three inputs, CHLA–1, WTEMP, ISS, and three hidden neurons (trial #11).

Increasing the number of hidden neurons in the ANN models also increased R^2_{tt} but not R^2_t (see trials #13 and #14). These findings suggest that increasing the number of hidden neurons does not improve the predictability of the models.

Including the less sensitive input variables (e.g., SAL) with the most sensitive inputs CHLA–1 and WTEMP, slightly increased R^2_{tt} (from 0.80 to 0.81) but there was little difference in the observed R^2_t (both 0.79) (i.e., compare trial #9 with #15). Hence, in some models, less sensitive inputs did not significantly affect the predictions of the output variable. However, in other models, fewer input variables improved the R^2_t as shown when trials #8 and #6 are compared ($R^2_t = 0.78$ to $R^2_t = 0.81$, respectively). This finding supports the idea that input variables with low sensitivity can add noise to the ANN model and therefore affect model performance.

The results shown in Table 5 were based on using CHLA–1 as an input variable. It is possible that having an input variable that is highly correlated to the output variable could potentially bias our interpretations. To circumvent this potential bias, we repeated ANN models shown in Table 5 and excluded CHLA–1 as an input variable. This allowed reassessment of the effects of input variables according to their rank in sensitivity analysis and correlation results shown in Table 3.

Comparison of Table 5 with Table 6 revealed that the R^2_t were significantly lower when CHLA–1 was excluded as an

Table 5 R^2 of predicted and actual chlorophyll concentrations for ANN models developed using data from the OL sampling site

Input variable	# Hidden Neuron	Output	OL R_{tt}^2	OL R_t^2	CB R_t^2	TC R_t^2	Trial #
CHLA–1, WTEMP	2	CHLA	0.80	0.79	0.75	0.69	9
CHLA–1, ISS	2	CHLA	0.79	0.75	0.73	0.66	10
CHLA–1, WTEMP, ISS	3	CHLA	0.81	0.80	0.75	0.70	11
CHLA–1, WTEMP, ISS, TNF	3	CHLA	0.83	0.79	0.75	0.69	12
	4		0.84	0.80	0.76	0.68	
CHLA–1, WTEMP, ISS, TNF, OP	3	CHLA	0.83	0.78	0.75	0.66	13
	4		0.84	0.78	0.75	0.67	
	5		0.85	0.78	0.74	0.66	
CHLA–1, WTEMP, ISS, TNF, OP, SAL	3	CHLA	0.83	0.79	0.75	0.67	14
	4		0.84	0.79	0.74	0.67	
	5		0.85	0.78	0.74	0.62	
	6		0.86	0.74	0.74	0.63	
CHLA–1, WTEMP, SAL	3	CHLA	0.81	0.79	0.77	0.67	15

The input variables were selected by sensitivity analyses.
Also shown is R^2 for the other sampling sites (see Table 1 for abbreviations/acronyms).

Table 6 R^2 of predicted and actual chlorophyll concentrations for ANN models developed using data from the OL sampling site

Input variables	# Hidden Neuron	Output	OL R_{tt}^2	OL R_t^2	CB R_t^2	TC R_t^2	Trial #
WTEMP, ISS	2	CHLA	0.66	0.63	0.52	0.46	16
WTEMP, ISS, TNF	3	CHLA	0.72	0.68	0.56	0.41	17
WTEMP, ISS, TNF, OP	3	CHLA	0.72	0.69	0.56	0.40	18
	4		0.73	0.62	0.52	0.35	
WTEMP, ISS, TNF, OP, SAL	3	CHLA	0.73	0.68	0.55	0.40	19
	4		0.74	0.69	0.55	0.42	
	5		0.74	0.69	0.16	0.14	

The input variables were selected by sensitivity analyses.
Also shown are R^2 for the other sampling sites (see Table 1 for abbreviations/acronyms).

input (i.e., from 0.80 to 0.73 for three input variables). Nonetheless, the interpretation of the results was consistent with our previous interpretation of models that included CHLA–1 as an input.

In summary, based on the results of sensitivity analysis: (1) the two most sensitive input variables, CHLA–1 and WTEMP, were able to effectively model CHLA, yielding $R_t^2 = 0.79$ (trial #9, Table 5); (2) the highest R_t^2 for predicting chlorophyll concentration was 0.80, and was obtained using the three most sensitive input variables, CHLA–1, WTEMP, and ISS (#11) with three hidden neurons, or four input variables, CHLA–1, WTEMP, ISS, and TNF (#12) with four hidden neurons; (3) more input variables or more hidden neurons resulted in higher R_{tt}^2 but lower R_t^2, which suggests overfitting of the ANN model; (4) in some models, adding less sensitive input variables such as salinity did not significantly affect the prediction of chlorophyll concentration – however, fewer input variables (with high sensitivity) did improve ANN model predictions; and (5) sensitivity analysis provided an effective method for choosing input variables for modeling ANNs.

9.09.3.3 Applying ANN Models to Other Estuary Sites

All ANN models were developed using the OL data set. We further investigated the robustness of the ANN models by using data collected from different sampling sites within the same estuary: CB and TC. The results are shown in Tables 5 and 6.

In general, the R_t^2's of the linear regression between predicted and actual chlorophyll concentration using data from the CB and TC sampling sites were lower than those obtained using data from the OL site. The highest R_t^2 for CB was obtained by using three input variables (CHLA–1, WTEMP, and SAL) with three hidden neurons ($R_t^2 = 0.77$; trial #15), which is 3% lower than that obtained for OL ($R_t^2 = 0.80$; trial #12) (Table 5). The highest R_t^2 for TC was obtained using three inputs (CHLA–1, WTEMP, and ISS) with three hidden neurons ($R_t^2 = 0.70$; trial #11), which was 10% lower than that for OL ($R_t^2 = 0.80$; same trial) (Table 5).

We repeated the above analysis using input data sets that excluded the chlorophyll concentration from the previous day (Table 6). The predictions of the ANN model were much lower, with the highest $R_t^2 = 0.56$ for CB (WTEMP, ISS, and TNF as inputs; trials #17 and #18) and the highest $R_t^2 = 0.46$ for TC (WTEMP and ISS as inputs; trial #16). Hence, the R_t^2's were 13% and 23% lower (respectively) than those obtained using the OL data set.

The presumed reason for the lower R_t^2 is subtle differences in the range and value of input variables among the sampling sites. These subtle differences affect the accuracy of chlorophyll concentration prediction because the ANN models were originally developed using sampling data from OL only. The presumed reason that chlorophyll concentrations were more accurately predicted using sampling data from the CB site than data from the TC site is because the CB site is more similar to the OL site in terms of range and values of environmental parameters than the TC site (data not shown). Moreover, both the CB and OL sampling sites are affected by terrestrial runoff, while the TC site is not. This line of reasoning is supported by another study that included the same sampling sites (i.e., OL, CB, and TC; Noble et al., 2003). The Noble et al. (2003) study showed statistically significant differences in the concentration of silicate and ammonia among the sites (see Table 1 in Noble et al. (2003)). In addition, the TC site is more affected by coastal ocean processes than either the OL or the CB site.

The results suggest that the prediction performance of an ANN model developed from one sampling site does not yield the same prediction performance when applied to other sampling sites within the same estuary. This finding is consistent with the hypothesis that subtle differences in the values of ecological variables at the other sampling sites affected ANN model predictions and that ANN models developed using data from one sampling site should not be expected to have the same prediction potential when applied to other sampling sites – even when the sites are within the same estuary.

9.09.3.4 Comparison of ANN Models with Linear Fitting Models

We compared ANN models with linear regression models to determine if ANN models were more accurate in terms of chlorophyll concentration predictions than linear regression-based models. For the linear models, we selected the same input variables that were used for the ANN models (Tables 5 and 6) and fitted the output chlorophyll data (CHLA) by using the linear equation

$$y = a + b_i x_i \quad [4]$$

where a and b_i are constants and x_i is the input variable. The OL data were used to determine the values of a and b_i and then the same model was applied to data collected from the CB and TC sampling sites.

Trials #20–#25 (Table 7) revealed the R_t^2 of linear regression models developed using the same data as those used for trials #9–#14 (Table 5). In general, the linear regression models yielded lower R_t^2 than those obtained using ANN models. The highest fitting $R_t^2 = 0.78$ for the linear model with two inputs (CHLA-1 and WTEMP; trial #20), which was lower than that obtained using ANNs ($R_t^2 = 0.80$). Similar results were obtained for other model comparisons shown in Tables 6 (trials #16–#19) and 7 (trials #26–#29). Hence, ANN models yielded better CHLA predictions than linear-based models.

9.09.4 Discussion

9.09.4.1 Input Data and Prediction Potential

Ecological modeling studies often use all possible variables as inputs to an ANN model (e.g., Karul et al., 2000; Maier and Dandy, 2000). As mentioned by Lee et al. (2003) and demonstrated in this study, input variables not contributing to the prediction (i.e., variables with low sensitivity) often provide noise rather than useful information to the model. As a consequence, these input variables compromise model performance.

One approach to minimize this problem is to select input variables based on the correlation of the input variable to the response variable (chlorophyll concentration). Melesse et al. (2008) selected input variables based on the correlations between chlorophyll concentration and total phosphate, nitrite and ammonium, temperature, turbidity, and dissolved oxygen. Consistent with that study, we found that the correlation of an

Table 7 Linear fit models of chlorophyll concentration based on data from the OL sampling site and the application of the model to other sampling sites (see **Table 1** for abbreviations/acronyms).

	OL	*OL*	*CB*	*TC*	
Input variable	R_{tt}^2	R_t^2	R_t^2	R_t^2	*Trial #*
CHLA–1, WTEMP	0.79	0.78	0.74	0.68	20
CHLA–1, ISS	0.78	0.73	0.73	0.67	21
CHLA–1, WTEMP, ISS	0.79	0.76	0.74	0.70	22
CHLA–1, WTEMP, ISS, TNF	0.80	0.76	0.75	0.69	23
CHLA–1, WTEMP, ISS, TNF, OP	0.80	0.76	0.74	0.69	24
CHLA–1, WTEMP, ISS, TNF, OP, SAL	0.80	0.76	0.74	0.69	25
WTEMP, ISS	0.58	0.48	0.46	0.43	26
WTEMP, ISS, TNF	0.61	0.52	0.47	0.40	27
WTEMP, ISS, TNF, OP	0.62	0.51	0.47	0.38	28
WTEMP, ISS, TNF, OP, SAL	0.62	0.51	0.47	0.37	29

input variable to chlorophyll concentration provided a reasonable estimate of the prediction potential for an input variable.

Another approach to determine the contribution of an input variable to the output prediction is to perform sensitivity analysis (Noble et al., 2000). Comparison of the rank order of input variables determined by correlation analysis (i.e., the method mentioned above) with those determined by sensitivity analysis revealed a high congruency for moderate to highly correlated input variables (see **Tables 3** and **4**). However, poorly correlated input variables differed in their ranking.

Given these findings, we recommend selecting input variables rather than using all input variables for building an ANN model because highly sensitive input variables offered better prediction of the response variable (chlorophyll concentration) (e.g., compare trial #6 with trial #8; **Table 2**). Second, we recommend that sensitivity analysis be used as the method of choice for selecting input variables for ANN models rather than the correlation analysis proposed by Melesse et al. (2008).

It should be noted that the inclusion of the chlorophyll concentration from the previous sampling day as an input variable significantly improved the prediction of chlorophyll concentration (in our study by more than 10%). This finding is consistent with the studies of Melesse et al. (2008) and Lee et al. (2003). Hence, antecedent algal concentrations offer good prediction of future phytoplankton chlorophyll concentrations.

9.09.4.2 Complexity of the ANN Model

It has been well established that a complex ANN model (i.e., having many hidden neurons and inputs) has the potential to overfit the data, whereas a simple ANN model (i.e., having few hidden neurons and inputs) has the potential to underfit the data (Bishop, 1995), which will affect prediction performance. Clearly, it is desirable to have an ANN model that is generalizable (Ozesmi and Ozesmi, 1999; Ozesmi et al., 2006) – that is, the model neither overfits nor underfits the data. Our study showed that less complex ANN models (i.e., those having few hidden neurons) yielded better predictions than more complex ANN models (e.g., see trial #5). Also with respect to the complexity of the ANN model, we found that a model with more input variables than necessary did not improve ANN prediction performance (e.g., compare trial #8 with trial #6; **Table 2**).

The reason for emphasizing the relationship between complexity and ANN model performance is that one can be misled to believe that increasing the number of inputs and/or number of hidden neurons improves ANN performance, as demonstrated by the increased R_{tt}^2 in **Table 2**. However, although the R_{tt}^2 increased with increasing hidden neurons and input variables, the R_t^2 of the same relationships did not change. Presumably, the reason that $R_{tt}^2 > R_t^2$ was because the ANN models were overfitting the training/testing data set. In an optimally fitted ANN model, $R_{tt}^2 \approx R_t^2$ (e.g., see **Table 2**, trial #6).

Another important finding of this study is that the second data set was key to determining whether or not the ANN model was properly optimized. Without this second data set, we would have not known that high R_{tt}^2 values were due to overfitting of the data as demonstrated in trial #8 (**Table 2**). Unfortunately, many ANN studies do not utilize a second data set to validate the performance of their ANN model. As a consequence, the reported ANN model performance values are often overestimated.

9.09.4.3 Sensitivity Analysis

Conventional sensitivity analysis involves determining the change in output prediction as a function of the minimum and maximum values of an input variable while keeping all other variables constant. We found that considering the minimum to average (min-to-average) value of an input variable or the maximum to average (max-to-average) value provided more information on the contribution of an input variable to the response variable (chlorophyll concentration) than the conventional approach (i.e., minimum to maximum values of an input variable). As demonstrated in the study, the conventional sensitivity analysis approach masks the sensitivity of some input variables, which is important for understanding ecological relationships. It should be noted that before we performed sensitivity analysis using the min-to-average/max-to-average approach, we did not know that low water temperature, low total nitrogen (filtered), low inorganic suspended solids, and high salinity were indicative of low

chlorophyll concentration; nor did we know that high water temperature, high orthophosphate, high nitrate–nitrite, and high inorganic suspended sediment were indicative of high chlorophyll concentration. That is, none of this new information would have been revealed had we only performed conventional sensitivity analysis. Hence, we recommend that, in addition to conventional sensitivity analysis, the min-to-average/max-to-average sensitivity analysis approach be used because it reveals the influence of low and high input values on the response variable, which, in our study, provided additional information of ecosystem functioning.

9.09.4.4 Application of Optimized ANN Models to Adjacent Sampling Sites

In addition to the depth and breadth of the LTER data sets from the Baruch Institute, which provided extensive data for us to test our ANN models, the other reason we chose the Baruch LTER data sets was because the sampling sites were in close proximity of one another. This offered the opportunity to apply ANN models (developed on one sampling site) to adjacent sites in order to assess the robustness of the model. We initially hypothesized that having a large number of records to train and test the ANN model and conducting validation tests using a second data set from the same site would produce a robust model that could be applied to all sites in the estuary. However, to our surprise, subtle differences in the ecological variables used as input variables to the ANN did affect the accuracy of chlorophyll concentration prediction. That is, the application of the developed models to adjacent sampling sites yielded lower R^2 values between actual and predicted chlorophyll concentrations than data from the OL sampling site. These findings are consistent with the hypothesis that subtle differences in the values of ecological variables at the different sampling sites affect ANN model predictions.

9.09.4.5 Relevance of This Study to Other Phytoplankton Modeling Studies

It is important to recognize that ANN models have been previously used to model the succession, timing, and magnitudes of planktonic species (Recknagel et al., 1997; Aoki et al., 1999; Barciela et al., 1999; Karul et al., 2000; Scardi, 2001; Lee et al., 2003; Velo-Suarez and Gutierrez-Estrada, 2007; Melesse et al., 2008; Talib et al., 2008). Direct comparisons of these studies to our study were not possible due to differences in the input variables, sampling frequencies, and response variables used for ANN modeling.

9.09.5 Conclusions

ANN models were developed for predicting phytoplankton chlorophyll concentration from biological, physical, and chemical data. Sensitivity analysis successfully identified the input variables most contributing to chlorophyll concentration predictions and these variables were then used to make accurate ANN models. In general, ANN models with low complexity in terms of their architecture (i.e., number of input variables and number of hidden neurons) provided more accurate predictions than more complex ones. Input variables with high sensitivity and/or those that were partially correlated to the response variable (i.e., chlorophyll concentration) provided better predictions than input variables that had low sensitivity and that were not correlated to the response variable. Comparison of the optimized ANN models to other sites within the same estuary provided reasonable-to-moderate predictions of chlorophyll concentration. The lower prediction potential for chlorophyll concentration at other sites (rather than the site used to develop the ANN model) suggests that subtle differences in the composition of environmental variables might be responsible for these differences (e.g., proximity of the sites to land and coastal water). ANN models were found to be superior to linear-based models in terms of the accuracy of output predictions. Sensitivity analysis played a useful role in determining which of the input variables contributed most to the accuracy of model predictions.

Acknowledgment

This work was supported (in part) by the NSF-CREST (HRD#0734232), the University of Washington (UW) Royalty Research Fund, and the UW Provost Bridge Funding Program to P.A.N.

References

Almeida, J.S., Noble, P.A., 2000. Neural computing in microbiology. Journal of Microbiological Methods 43, 1–2.

Aoki, I., Komatsu, T., Hwang, K., 1999. Prediction of response of zooplankton biomass to climatic and oceanic changes. Ecological Modelling 120, 261–270.

Baird, M.B., Emsley, S.M., McGlade, J.M., 2001. Modelling the interacting effects of nutrient uptake, light capture and temperature on phytoplankton growth. Journal of Phytoplankton Research 23, 829–840.

Barciela, R.M., Garcia, E., Fernandez, E., 1999. Modelling primary production in a coastal embayment affected by upwelling using dynamic ecosystem models and artificial neural networks. Ecological Modelling 120, 199–211.

Basheer, I.A., Hajmeer, M., 2000. Artificial neural networks: fundamentals, computing, design, and application. Journal of Microbiological Methods 43, 3–31.

Bishop, C.M., 1995. Neural Networks for Pattern Recognition. Oxford University Press, London.

Gardner, L., Thombs, L., Edwards, D., Nelson, D., 1989. Time series analyses of suspended sediment concentrations at North Inlet, South Carolina. Estuaries and Coasts 12, 211–221.

Goethals, P., Dedecker, A., Gabriels, W., Lek, S., De Pauw, N., 2007. Applications of artificial neural networks predicting macroinvertebrates in freshwaters. Aquatic Ecology 41, 491–508.

Hagan, M.T., 1995. Neural Network Design. PWS Publishing Company, Boston, MA.

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

Lee, J.H.W., Huang, Y., Dickman, M., Jayawardena, A.W., 2003. Neural network modelling of coastal algal blooms. Ecological Modelling 159, 179–201.

Maier, H.R., Dandy, G.C., 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.

Melesse, A.M., Krishnaswamy, J., Keqi, Z., 2008. Modeling coastal eutrophication at Florida Bay using neural networks. Journal of Coastal Research 24, 190–196.

Noble, P.A., Almeida, J.S., Lovell, C.R., 2000. Application of neural computing methods for interpreting phospholipid fatty acid profiles from natural microbial communities. Applied and Environmental Microbiology 66, 694–699.

Noble, P.A., Tribou, E.H., 2007. Neuroet: an easy-to-use artificial neural network for ecological and biological modeling. Ecological Modelling 203, 87–98.

Noble, P.A., Tymowski, R.G., Morris, J.T., Fletcher, M., Lewitus, A.J., 2003. Contrasting patterns of phytoplankton community pigment composition in two salt marsh estuaries in Southeastern United States. Applied and Environmental Microbiology 69, 4129–4143.

Ozesmi, S., Tan, C., Ozesmi, U., 2006. Methodological issues in building, training, and testing artificial neural networks in ecological applications. Ecological Modelling 195, 83–93.

Ozesmi, S.L., Ozesmi, U., 1999. An artificial neural network approach to spatial habitat modelling with interspecific interaction. Ecological Modelling 116, 15–31.

Rao, V.B., Rao, H.V., 1993. C^{++} Neural Networks and Fuzzy Logic. MIS Press, New York, NY.

Rabouille, C., Mackenzie, F.T., Ver, L.M., 2001. Influence of the human perturbation on carbon, nitrogen, and oxygen biogeochemical cycles in the global coastal ocean. Geochimica et Cosmochimica Acta 65, 3615–3641.

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

Rumelhart, D.E., Hutton, G.E., Williams, R.J., 1986. Learning representations by back-propagation errors. Nature 323, 533–536.

Sarkar, R.R., Chattopadhayay, J., 2003. Occurrence of planktonic blooms under environmental fluctuations and its possible control mechanism – mathematical models and experimental observations. Journal of Theoretical Biology 224, 501–516.

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

Smayda, T.J., 1997. Harmful algal blooms: their ecophysiology and general relevance to phytoplankton blooms in the sea. Limnology and Oceanography 42, 1137–1153.

Smith, M., 1996. Neural Networks for Statistical Modeling. International Thomson Computer Press, Boston, MA.

Talib, A., Hasan, Y.A., Varis, O., 2008. Application of machine learning techniques in data mining of ecological datasets. International Conference on Environmental Research and Technology (ICERT 2008), 677–681. 28–30 May 2008, Penang, Malaysia.

Velo-Suarez, L., Gutierrez-Estrada, J.C., 2007. Artificial neural network approaches to one-step weekly prediction of *Dinophysis acuminata* blooms in Huelva (Western Andalucia, Spain). Harmful Algae 6, 361–371.

Relevant Websites

http://links.baruch.sc.edu – Belle W. Baruch Institute for Marine & Coastal Sciences; Daily Water Sample Data.

http://tidesandcurrents.noaa.gov – NOAA: Tides & Currents.