Eutrophication forecasting and management by artificial neural network: a case study at Yuqiao Reservoir in North China

Ya Zhang, Jinhui Jeanne Huang, Liang Chen and Lan Qi

ABSTRACT

Yuqiao Reservoir is the potable water supply source for a city with a population of more than 14 million. Eutrophication has threatened the reliability of drinking water supplies and, therefore, the forecasting systems for eutrophication and sound management become urgent needs. Water temperature and total phosphorus have long been considered as the major influencing factors to eutrophication. This study used the artificial neural network (ANN) model to forecast three water quality variables including water temperature, total phosphorus, and chlorophyll-a in Yuqiao Reservoir. Two weeks in advance for forecasting was chosen to ensure a sufficient preparation response time for algae outbreak. The Nash–Sutcliffe coefficient of efficiency ($R^2$) was between 0.84 and 0.99 for the training and over-fitting test data sets, while it was between 0.59 and 0.99 for the validation data set. To better respond to the algae outbreak, a number of management scenarios formed by orthogonal experimental design were modeled to assess the responses of chlorophyll-a and an optimal management scenario was identified, which can reduce chlorophyll-a by 23.8%. This study demonstrates that ANN model is potentially useful for forecasting eutrophication up to 2 weeks in advance. It also provides valuable information for the sound management of nutrient loads to reservoirs.

Key words | artificial neural network, data-driven technique, orthogonal experimental design, prediction and forecast, reservoir eutrophication, sensitivity analysis

INTRODUCTION

Eutrophication is the most common and severe environmental hazard in lake and reservoir ecosystems. Eutrophication has seriously threatened the reliability of drinking water supplies and has caused serious ecological damage to the environment. The development of cyanobacteria and green algae blooms is common in eutrophicated waters (Smith 1990). The vast proliferation of cyanobacteria in lakes and reservoirs has become a growing aquatic environmental problem. Thus, a number of researchers have used various models to study eutrophication in lakes and reservoirs based on mass balance since the 1970s (Chen 1970). Recently, improved knowledge regarding the eutrophication processes and advanced computing capabilities, which include multi-dimensional lake and reservoir hydrodynamic and water quality models, have been developed and used to study water quality problems (Orlob 1983; Martin & McCutcheon 1999; Cioffi & Gallerano 2001). In addition, researchers are searching for new models to solve water and environmental issues. Recently, the artificial neural network (ANN) model has been widely used. This model includes highly flexible function approximations that are useful for multiple water resources applications. The ANN model has been tested in limnological (Recknagel 1997; Karul et al. 2000; Wilson & Recknagel 2001), river (Wen & Lee 1998; Huang & Foo 2002) and coastal (Aguilera et al. 2001; Lee et al. 2005; Palani et al. 2008) systems. Most of these studies have shown that the ANN model performs better than classical modeling.
methods. Currently, the back-propagation (BP) learning technique has been successfully used in nonlinear complex systems because it can approximate functions arbitrarily by using the gradient descent algorithm or the faster algorithm (Hagan et al. 1996). Recknagel et al. (1997) used the ANN model to simulate and predict algal blooms in four freshwater systems with different trophic levels. In addition, Clair & Ehrman (1998) used network technology to simulate the effects of changing climate on discharge and on two water quality parameters. Similarly, Karul et al. (2000) used the three-layer Levenberg–Marquardt feed forward learning algorithm to model the eutrophication process in three water bodies in Turkey. Wei et al. (2001) used ANN technology to predict the timing and magnitude of Microcystis, Phormidium and Synedra algal blooms, and also to quantify the interactions between abiotic factors and algal genera in Lake Kasumigaura (Japan). Kuo & Wang (2006) used a combined neural network and genetic algorithm that was developed for the water quality management of the Feitsui Reservoir in Taiwan. This study indicated that the ANN model could effectively simulate the dynamics of reservoir water quality. Kim et al. (2012) used the back-propagation algorithm of the feed forward neural networks model to predict algae blooms over short periods in the Daecheong Reservoir of the Geum River.

Many studies on prediction of algal growth focused on environmental factors such as total nitrogen (TN), total phosphorus, pH, biological oxygen demand (BOD), and dissolved oxygen (DO) without the consideration of meteorological factors, e.g. air temperature (AT), precipitation (Prep), light, etc. There is also a lack of consideration on identifying key factors for the selection of input variables or the optimal combination of factors that would effectively control algae bloom. In addition, many forecast systems predict a very short time in advance (e.g. 3 or 7 days); there is not sufficient time for taking appropriate actions to combat eutrophication; for example, shutting down certain point sources and harvesting emerged/submerged/floating plants in the reservoir, etc. All these mentioned actions may take a relatively longer time to be effective. In this study, the feasibility of the ANN model for predicting eutrophication 2 weeks in advance in Yuqiao Reservoir has been investigated. As water temperature (WT) and total phosphorus have long been considered to have significant influence on eutrophication and its indicator of chlorophyll-a (Chl-a), the ANN model was used to predict the three water quality variables (WT, total phosphorus, and Chl-a). Based on the selected water quality variables that were predicted by the ANN model, the relative importance of the factors that affect algal growth were evaluated by sensitivity analysis. In addition, optimal combinations of the four most sensitive parameters were determined by the orthogonal experimental design. This will provide valuable information for sound management of nutrient loads to reservoirs and consequently control the algae bloom.

**MATERIALS AND METHODS**

**The study area and water quality data**

Yuqiao Reservoir was completed in 1965 with an initial storage capacity of $1.56 \times 10^9$ m$^3$. The main dam is located at the outlet of the reservoir (Figure 1). The reservoir has a surface area of 433 km$^2$, a mean depth of 4.74 m and a maximum depth of 12.16 m (near the dam). Yuqiao Reservoir serves as an important drinking water source for the city of Tianjin, which has a population of over 15 million. However, several algal blooms have occurred in Yuqiao Reservoir in the past 10 years. The algal blooms have adversely affected the safety of potable water supply and the reservoir management.

The water quality data used in this study were collected from four field-monitoring sites, namely, Center (117$°30'15.3''$E, 40$°02'32.0''$N), North (117$°30'15.3''$E, 40$°03'26.2''$N), East (117$°32'28.7''$, 40$°02'32.0''$N) and West (117$°29'29.6''$, 40$°02'32.0''$) of the reservoir as shown in Figure 1. The frequency of sampling was twice a month from 2003 to 2010. Owing to the occurrence of algal blooms, the data were collected weekly for the summer and autumn seasons between 2008 and 2010. Samples were immediately preserved in 1 L polypropylene sampling bottles in darkness at 4 °C and analyzed within 24 h. Water quality parameters were measured at all sampling stations using State Environmental Protection Administration standard methods (Jin & Tu 1990). Chl-a was measured after extraction in 90% acetone by a freeze-thaw method (Lewitus et al. 1998). The following parameters were measured: WT, pH, electrical conductivity, TN,
ammonium nitrogen (NH$_4^+$-N), nitrate nitrogen (NO$_3^-$-N), nitrite nitrogen (NO$_2^-$-N), DO, chemical oxygen demand (COD), BOD, total phosphate (TP), phosphate (PO$_4^{3-}$-P), suspended solids (SS), total dissolved solids (TDS), Secchi disk depth (SD) and Chl-a. To study the effects of meteorological parameters on algae blooms, meteorological data including AT, Prep and sunshine hours (SH) from a nearby station Baodi (No. 54525; Figure 1) was obtained from China Meteorological Data Sharing Service System (http://cdc.cma.gov.cn). Table 1 presents the characteristics of these water quality data and meteorological data.

**Artificial neural networks**

**Architecture**

An ANN is an information processing system that replicates the behavior of the human brain roughly by emulating the operations and connectivity of biological neurons. McCulloch & Pitts (1943) first introduced the concept of artificial neurons. Rumelhart et al. (1985) first applied ANNs in research, which began with the introduction of the BP training algorithm for feed forward ANN. The ANN technique has recently been proposed as an efficient tool for modeling and forecasting, mainly due to its wide range of applications and its usefulness for solving complicated and nonlinear problems. ANNs represent complex, nonlinear functions that have many selected parameters and are trained or calibrated so that the ANN output corresponds to a known set of data. The multilayer perceptron (MLP) neural network model was designed to function well for modeling nonlinear phenomena. A feed forward MLP network consists of an input layer and an output layer with one or more hidden layers between them. In addition, each layer contains a certain number of artificial neurons. The main differences between the various types of ANNs involve network architecture and the method that is used for determining the weights and functions for the input and neurodes (training). For example, an artificial neuron in a typical ANN architecture (Figure 2) receives the input (signals ($x$) with weight ($w$)), calculates their weighted average ($z$), using the summation function) and uses an activation function to produce an output ($o = f(z)$, where $z = \sum_{i=1}^{n} x_i w_i$). In time series prediction, supervised training
is used to train the ANN in such a way as to minimize the difference between the network output and the measured target. Therefore, training is a process of weight adjustment that attempts to obtain a desirable outcome with least squares residuals. The most common training algorithm used in the ANN literature is called BP which is a multilayer feed-forward BP network trained by error back propagation algorithm (Rumelhart et al. 1988).

Specht (1991) invented the general regression neural network (GRNN) model, which predicts continuous outputs.

The typical GRNN architecture is shown in Figure 3. To calculate the differences between all paired input pattern vectors and to estimate the probability density functions of the input variables, the GRNN nodes require two main functions. The differences between the input vectors are calculated from the simple Euclidean distance between the data points in space. The predicted output value is determined by weighting the calculated distance of any point with the probability of other points that occur in the area yields. This process is illustrated in Equations (1) and (2) (Masters 1995)

$$E_{Y/X}(X) = \frac{\int_{-\infty}^{\infty} y f_{XY}(x,y)dy}{\int f_{XY}(x,y)dy}$$

$$f = \sum_{i=1}^{n} \frac{y_j \exp(-D(x,x_i))}{\sum_{i=1}^{n} \exp(-D(x,x_i))}$$

$$D(x,x_i) = \sum_{j=1}^{p} (x_j - x_{ij}/\sigma_j)^2$$

Table 1 | Basic statistics for the water quality variables and meteorological data that were measured between 2003 and 2010 in the Yuqiao Reservoir

<table>
<thead>
<tr>
<th>Variable</th>
<th>Unit</th>
<th>Time interval of data</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Mean</th>
<th>Std. deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>WT</td>
<td>°C</td>
<td>Semimonthly</td>
<td>0.0</td>
<td>32.0</td>
<td>14.1</td>
<td>10.1</td>
</tr>
<tr>
<td>pH</td>
<td>µS/cm</td>
<td>Semimonthly</td>
<td>7.4</td>
<td>9.9</td>
<td>8.4</td>
<td>0.4</td>
</tr>
<tr>
<td>EC</td>
<td>µS/cm</td>
<td>Semimonthly</td>
<td>180</td>
<td>643</td>
<td>450</td>
<td>63</td>
</tr>
<tr>
<td>TN</td>
<td>mg/L</td>
<td>Semimonthly</td>
<td>0.460</td>
<td>4.080</td>
<td>1.691</td>
<td>0.665</td>
</tr>
<tr>
<td>NH4-N</td>
<td>mg/L</td>
<td>Semimonthly</td>
<td>0.020</td>
<td>0.590</td>
<td>0.200</td>
<td>0.099</td>
</tr>
<tr>
<td>NO3-N</td>
<td>mg/L</td>
<td>Semimonthly</td>
<td>0.040</td>
<td>2.540</td>
<td>1.111</td>
<td>0.670</td>
</tr>
<tr>
<td>NO2-N</td>
<td>mg/L</td>
<td>Semimonthly</td>
<td>0.004</td>
<td>0.132</td>
<td>0.032</td>
<td>0.022</td>
</tr>
<tr>
<td>DO</td>
<td>mg/L</td>
<td>Semimonthly</td>
<td>6.0</td>
<td>19.9</td>
<td>10.5</td>
<td>2.3</td>
</tr>
<tr>
<td>COD</td>
<td>mg/L</td>
<td>Semimonthly</td>
<td>1.8</td>
<td>6.3</td>
<td>3.8</td>
<td>0.8</td>
</tr>
<tr>
<td>BOD</td>
<td>mg/L</td>
<td>Semimonthly</td>
<td>0.7</td>
<td>5.3</td>
<td>2.1</td>
<td>0.8</td>
</tr>
<tr>
<td>TP</td>
<td>mg/L</td>
<td>Semimonthly</td>
<td>0.005</td>
<td>0.072</td>
<td>0.032</td>
<td>0.013</td>
</tr>
<tr>
<td>PO4-P</td>
<td>mg/L</td>
<td>Semimonthly</td>
<td>0.005</td>
<td>0.226</td>
<td>0.033</td>
<td>0.033</td>
</tr>
<tr>
<td>SS</td>
<td>mg/L</td>
<td>Semimonthly</td>
<td>1</td>
<td>33</td>
<td>5.2</td>
<td>4.2</td>
</tr>
<tr>
<td>TDS</td>
<td>mg/L</td>
<td>Semimonthly</td>
<td>89</td>
<td>414</td>
<td>298.4</td>
<td>35.6</td>
</tr>
<tr>
<td>SD</td>
<td>mg/L</td>
<td>Semimonthly</td>
<td>60</td>
<td>390</td>
<td>143.9</td>
<td>72.6</td>
</tr>
<tr>
<td>Chl-a</td>
<td>mg/L</td>
<td>Semimonthly</td>
<td>0.001</td>
<td>0.032</td>
<td>0.007</td>
<td>0.005</td>
</tr>
<tr>
<td>AT</td>
<td>°C</td>
<td>Daily</td>
<td>-10.50</td>
<td>30.30</td>
<td>13.24</td>
<td>11.21</td>
</tr>
<tr>
<td>Prep</td>
<td>mm</td>
<td>Daily</td>
<td>0</td>
<td>71.50</td>
<td>1.34</td>
<td>5.91</td>
</tr>
<tr>
<td>SH</td>
<td>hour</td>
<td>Daily</td>
<td>0</td>
<td>12.60</td>
<td>5.84</td>
<td>3.77</td>
</tr>
</tbody>
</table>
where \( n \) is the total number of observations in the data set, \( x \) is the input vector, \( x_i \) is the \( i \)th case vector, \( x_j \) is the \( j \)th data value in the input vector, \( x_{ij} \) is the \( j \)th data value in the \( i \)th case vector, \( y_i \) is the \( i \)th case actual output value and \( \sigma_j \) is the smoothing factor (Parzen’s window) for the \( j \)th variable. The error represents the mean of the mean square errors, which is the averaged square of the difference between the estimated and actual values. After determining the error and based on the optimization technique, the above calculation is run numerous times with different smoothing factors. The training process stops when the threshold minimum square error value is reached or when the test set square error begins to increase.

**Parameter selection**

*Hidden layers and nodes.* In the ANN model, hidden layers and nodes are usually determined by trial-and-error. A rule of thumb for selecting the number of hidden nodes relies on the fact that the number of samples in the training set should at least be greater than the number of synaptic weights. Hecht-Nielsen (1987) indicated that between \( I \) and \( 2I + 1 \) hidden nodes \( M \) occur in this model, where \( I \) represents the number of input nodes. In addition, \( M \) should not be less than \( I/3 \), and the number of output nodes and the optimum value of \( M \) should be determined by trial and error. If the number of nodes is too small to capture the underlying behavior of the data, the performance of the network may be impaired. Owing to their better generalization capabilities and fewer over-fitting problems, networks with fewer hidden nodes are generally preferred. In addition, a trial and error procedure for selecting hidden nodes was conducted in this study by gradually varying the number of nodes in the hidden layer.

*Learning rate and momentum.* Learning rate \((\eta)\) and momentum \((\alpha)\) are used to expedite the training process while reducing error. No specific rule exists for selecting values for these parameters. The learning rate ranges from 0 to 1 and the momentum ranges from 0 to 0.9. However, the training process is started by adopting one set of values (i.e. \( \eta = 0.6, \alpha = 0.9 \)) and then adjusting these values as necessary. The ANN model with the best performance (based on validation data set) is selected.

*Initial weights.* The weights of a network trained by BP must be initialized to small, non-zero, and normally distributed random number in the range of \(-1\) to \(1\). The initial weight is set a reasonable value first and then is continuously adjusted by trials. When a network has identified a local minimum and training is terminated before reaching an acceptable solution, the number of hidden nodes or the

![General GRNN architecture](image)
learning rate may be changed to solve the problem. Furthermore, a different set of initial weights can be used or the network can be restarted. However, a network can reach an acceptable solution that may not correspond with the global minimum of the problem.

Stopping criteria. Two stopping criteria are frequently used, including stopping after a certain number of epochs and stopping when the total sum-squared error, which depends on the actual problem, reaches a defined minimum. By considering the computational time, the training process is generally terminated when the iteration reaches a maximum number or when the number of good patterns is large enough.

Data partitioning

Data in neural networks are categorized into training sets, over-fitting test sets, and production sets. The training set is used to determine the adjusted weights and the biases of a network. The test set is used for calibration, which prevents networks overtraining. To select a good training set from the available data series, all extreme events (i.e. all possible minimum and maximum values in the training set) are generally considered. The test sets should contain a representative data set. The order in which training samples are presented to the network is randomized between iterations, which improve the performance of the BP algorithm. The Ward Systems Group (NeuroShell 2™ 2007) suggested that both training and test data sets are statistically comparable and that the test data set should contain approximately 10–40% of the data in the training data set.

In this study, 6 months/12 data sets from May to October each year were used for each station. Seventy-two water quality data sets for the north, west and east stations were obtained. In the current case, the water quality data from the north and center stations (total of 172 data points for each variable) were divided into training (124 data points) and test sets (48 data points). The training set contained 72% of the records and the test set contained 28%. The data for Stations West (92 data points) and East (92 data points), which were not used for training and testing in the previous step, were used as the validation set. A vector pair with missing measurements cannot be used for training a model, and an input vector with a missing measurement cannot be evaluated by a developed model. So the efficacy of implementing a neural network model is highly dependent on the quantity, historical range and quality of the data used for its development.

Input variable selection

Determining the input variables in an ANN model is one of main tasks that affect the output variables. For this purpose, an extensive field-monitoring program has been conducted for nearly 8 years. Among the three meteorological variables and the 16 water quality data factors that were listed before, several are believed to control eutrophication and include chemical and physical parameters that were regularly measured in the Yuqiao Reservoir between 2003 and 2010. Statistical analysis, a priori knowledge of causal variables, and time series plot inspections of potential input and output data are important for choosing input reference variables. For this neural network model, the input variables were chosen based on the statistical correlation analysis of the field-monitoring data, the prediction accuracy of the water quality variables and the domain knowledge. Pearson indices calculated for Chl-a is correlated with COD ($r^2 = 0.69$), SD ($r^2 = 0.54$), SS ($r^2 = 0.61$), NO3-N ($r^2 = 0.51$) and TP ($r^2 = 0.54$) at time $t$ in the reservoir field-monitoring domain and all the $P$-values less than 0.05.

Because the growth of algae is an ongoing process, the use of a lag time with input variables may be more efficient for predicting Chl-a. Statistical analyses of the spatiotemporal water quality data were used in combination with a factor analysis data mining technique to identify important parameters and locations in the selected study area. The factor analysis is the statistical methods that use several factors to describe the connections of many factors or to reflect most information of the original data (Akaike 1987). We selected water quality variables in the ANN model to identify the optimal predictive model.

After choosing appropriate input variables, the next step involved determining the appropriate lag times for each of these variables. This step was particularly important for complex problems with a large number of potential input data or when no prior knowledge is available, which suggests possible lag times when strong relationships exist between the output and the input time series. We retained
the best performance network for a single variable. In addition, the effects of adding each of the remaining inputs were assessed. This process was repeated for each combination of input variables until the addition of extra variables does not significantly improve the model’s performance (Masters 1995).

Model performance evaluation

The results of model training (with the training set) can be evaluated by comparing the model predictions with the measured values in the over-fitting test set that uses the scatter and time series plots. Model performances are assessed based on the root mean square error (RMSE) (Equation (3)), the mean absolute error (MAE) (Equation (4)), the Nash–Sutcliffe coefficient of efficiency ($R^2$) (Equation (5)) (Nash & Sutcliffe 1970) and the correlation coefficient ($r$).

\[
RMSE = \sqrt{\frac{1}{N} \sum (X_{\text{observed}} - X_{\text{predicted}})^2}
\] (3)

\[
MAE = \frac{1}{N} \sum |X_{\text{predicted}} - X_{\text{observed}}|
\] (4)

\[
R^2 = 1 - \left(\frac{F}{F_0}\right)
\]

\[
F = \sum (X_{\text{observed}} - X_{\text{predicted}})^2
\]

\[
F_0 = \sum (X_{\text{observed}} - X_{\text{mean observed}})^2
\]

where $N$ denotes the total number of observations in the data set.

Model description

A commercial neural net software package (NeuroShell 2™ (2007) Release 4.2, Ward Systems Group) was used to develop the neural network prediction and forecast model. A set of input and output was defined and a suitable training set was selected. These following steps were conducted during the development of these models: choosing the performance criteria, the division and pre-processing of the available data, determining the appropriate model inputs and network architecture, optimizing the connection weights, and validation. Figure 4 illustrates the diagram of modeling steps involved in this study and explained above under artificial neural networks.

The ANN models were based on Ward net (BP). The GRNN training algorithm was used to determine the non-linear relationships between the water quality and eutrophication indicators. The Ward net ANN architecture consisted of BP with three hidden layers and with different Gaussian, hyperbolic-tangent and Gaussian complement activation functions. BP is especially useful for fitting high dimensional and continuously valued functional approximations with data. However, the model data must be arranged into input/output vector pairs that represent the variables of interest for training and evaluating neural networks.

The modeled WT, TP, and Chl-a values at times $t$ and $t+1$ were the only output variables that were considered for the ANN prediction and forecast model of the water quality variables. The prediction model was based on other variables at the same time $t$. The forecast model was
used to predict the variables at time $t + 1$, based on the same variables used in the predicted model and the modeled variables itself at time $t$, $t - 1$, and $t - 2$. Figure 5 contains the input variables, acceptable lag times, output variables and ANN architecture for the WT, TP and Chl-a prediction and forecast model.

**RESULTS AND DISCUSSION**

**Water temperature modeling results**

Water temperature is an extremely important variable that should be considered when evaluating water quality. Water temperature can dramatically affect chemical and biological processes, such as the distribution and abundance of organisms, the solubility of chemical compounds in water, the growth rate of biological organisms, water density, mixing of different water densities, and current movements. However, the WT scenario predicted by the model in this study was related to the AT and SH input parameters at time $t$ from the north and center stations and to the best net WT ($t$) prediction. In addition, the forecast model scenario was related to AT, SH and WT at a lag time of $t$, $t - 1$ and $t - 2$ for forecast WT ($t + 1$). Semi-monthly WTs in the Yuqiao Reservoir were simulated (with respect to time and space) in the ANN model, which contained BP architecture, three hidden layers with different activation functions (Ward net was selected for the WT model) and an initial weight of 0.3. In addition, an optimum learning rate of 0.6 and a momentum of 0.9 were selected. The parameters that produced the best net fit for the two validation data sets were retained for the final prediction. The WT prediction and forecast model results of the training and over-fitting test sets and the validation sets at the east and west stations are shown in Figure 6(a) and 6(b), respectively. The results indicate that an adequate temperature prediction and forecast was obtained by using a half and full month of WT data. The neural network was capable of simulating WT with a high accuracy when using the prediction ($R^2 > 0.99$, $MSE < 0.02\text{ °C}$) and forecast ($R^2 > 0.75$, $MSE < 0.3\text{ °C}$) model. For the forecast model, the accuracy of training, over-fitting test and validation were not very good with respect to the prediction model because the monitoring interval was semimonthly and the variation of WT was relatively large during this time. So the monitoring time interval had to be reduced if we wanted to improve the forecast model’s performance.

**Total phosphorus model results**

The ANN model was developed to simulate TP concentrations semimonthly in the Yuqiao Reservoir relative to time and space. In this model, the GRNN architecture was used with genetic adaptive calibration. To predict or forecast the TP concentrations, the ANN model was trained with four selected input variables from 16 water quality variables at different lag times. Several model scenarios with different input parameters and ANN architectures were tested. The input variables were optimized by removing or adding parameters one by one. However, in this study, the TP predicted by the model was related to the WT, COD, SS, SD and Prep at time $t$ for the input parameters from the north and center stations and the net GRNN. In addition, the forecast model scenario, which used the same variables and TP itself at lag times of $t$, $t - 1$ and $t - 2$, was used to model TP ($t + 1$).

The ANN TP prediction model results are presented as a scatter diagram of the modeled and field-monitoring TP concentrations for the training and over-fitting test data sets ($R^2 > 0.99$, $MSE < 0.001$) from the range of TP validation values at the east ($R^2 = 0.93$, $MSE = 0.01$) and west ($R^2 = 0.96$, $MSE = 0.02$) stations in Figure 7(a) and 7(b), which contains the best TP forecast model for the training ($R^2 = 0.99$, $MSE = 0.001$), over-fitting ($R^2 = 0.99$, $MSE = 0.001$) and validation data sets ($R^2 > 0.90$, $MSE < 0.005$), respectively. The developed ANN models accurately simulated the TP concentrations in the reservoir. The results reveal that an acceptable agreement between the ANN modeled and field-monitoring data can be achieved. The predicted and forecasted ANN model can accurately simulate the measured TP concentrations at any location in the model domain. This TP results from non-point source pollution due to Prep and from point source pollution due to sewage and periodic phytoplankton blooms. In tropical reservoirs, where frequent phytoplankton blooms (in terms of Chl-a) occur, TP concentrations fluctuate and become supersaturated during the
The selected water quality variables were used for the prediction and forecast model of the Yuqiao Reservoir. (a) Water temperature; (b) total phosphorus; (c) chlorophyll-a (t−1 and t−2 indicate lag times of 2 weeks and 1 month, respectively). (a) Water temperature prediction model. (b) Water temperature forecast model.
blooms. This super-saturation occurred at the center and east stations in 2006 (Figure 8). The Pearson correlation analysis indicated that the measured TP is positively correlated with the measured Chl-a concentration ($R^2 = 0.81$, $P < 0.05$). The TP concentrations in reservoirs can be predicted within an acceptable accuracy relative to measured data. Next, the verified TP ANN forecast model can be used as a guide for water quality management and for controlling eutrophication.

**Chlorophyll-a model results**

Regarding the existence and degree of eutrophication in the reservoir, Chl-a is one of the most important indicators. The growth rate of algae is influenced by sunlight, WT, and nutrient concentrations. The eutrophication dynamics in the Yuqiao Reservoir (regarding time and space) were modeled in terms of the Chl-a concentration. These models were trained to predict and forecast the algal biomass from 16
water quality variable inputs, including Chl-a at different lag times. Various model scenarios with different input parameters and different ANN architectures were tested regarding their prediction and forecast abilities for Chl-a(t) and Chl-a(t+1), respectively. The input variables were optimized by removing or adding one parameter at a time. However, in this case, the predicted model scenario was related to WT, COD, SS, SD, TP, NO₃-N, pH, and DO at time t from the north and center stations and the best performed net GRNN for the prediction of Chl-a(t). In addition, the forecast model scenario that used these variables and the Chl-a itself at lag times of t, t – 1 and t – 2 were used for forecast Chl-a(t + 1).

The typical ANN Chl-a prediction model results are shown in a scatter diagram (Figure 8) that contains the modeled and field-monitoring Chl-a concentrations for the training and over-fitting test sets (R² > 0.99, MSE < 0.001). In addition, Figure 9(a) and 9(b) contains the range of Chl-a validation values from the east (R² = 0.98, MSE = 0.01) and west (R² = 0.98, MSE = 0.02) stations, the best Chl-a forecast model for the training and over-fitting tests (R² > 0.99, MSE < 0.001) and the validation data are from the east (R² = 0.90, MSE =...
0.01) and west ($R^2 = 0.59, MSE = 0.03$) stations, respectively. The results indicated that the developed ANN models accurately simulated the Chl-a concentrations in the reservoir. The model can be used for predicting and forecasting purposes. The predicted and forecasted model can simulate the phytoplankton concentration dynamics (in terms of Chl-a) in the Yuqiao Reservoir with limited input parameters. The uncertainty in the results may be due to the time of the samples being collected, the dynamic characteristics of water quality in the reservoir, hydrodynamic forcing and unknown pollution sources, or due to the addition of freshwater in the region of interest over a short period. The ANN model also demonstrates that only data from two monitoring stations are necessary for predicting/forecasting water quality variables within the study domain. This approach could therefore reduce monitoring costs by allowing researchers to choose appropriate numbers of sampling stations and variables for field monitoring.

In addition, the model could be used to simulate other complex behaviors during the eutrophication process by using the ANN technique. After the model is built, further analyses are required to evaluate the effects of each input variable in the model and to determine the optimal combination for controlling the occurrence and development of harmful algal bloom.

**Sensitivity analysis**

Sensitivity analysis is the most common method for evaluating the effects of each input variable impact to the output variable in a model. An itinerating method was used as a trained forecaster by varying each of the input variables individually and at a constant rate. Various constant rates (-5, 5, 10 and 20%) were selected. For every input variable, the percentage of change in the output, which resulted from the changes in the input variables, was observed. The sensitivity level of each input variable was calculated by using Equation (6):

$$Sensitivity\ Level\ of\ X_i(\%) = \frac{1}{N} \sum_{j=1}^{N} \left( \frac{\% \ change \ in \ output}{\% \ change \ in \ input} \right)_j \times 100 \quad (6)$$

where $N$ represents the number of data sets that were used in the study.

The sensitivities of each of the input variables are provided in Figure 10. These results indicate that the output Chl-a is very sensitive to changing NO$_3$-N, DO, TP, COD and pH (in decreasing order). It was observed that the diatoms and *potamogeton crispus* are the dominant species in spring (March–May) and cyanobacteria is the dominant species in the summer and autumn in Yuqiao Reservoir. The phytoplankton biomass in the reservoir is expressed by Chl-a concentration and therefore does not consider the proportion of each type of algae here.

Dissolved inorganic nitrogen (DIN = NH$_4^+$-N + NO$_3$-N + NO$_2$-N) was the major source of algae in the reservoir. The relationship between Chl-a and the ratio of NH$_4^+$-N: (NO$_3$-N + NO$_2$-N), based on the monthly average monitoring data during the period of 2003–2010, was significant (Figure 11). The initial surge in Chl-a from April to June was preceded by an increase in the ratio of NH$_4^+$-N: (NO$_3$-N + NO$_2$-N) from 0.14 in April to 0.17 in May to 0.19 in June (Figure 12). The ratio of NH$_4^+$-N: (NO$_3$-N + NO$_2$-N) increase from April to June was due to a small decrease in NH$_4^+$-N concentration and a larger decrease in NO$_3$-N concentration (Table 2). Given the affinity of diatoms for NO$_3$-N (Heil et al. 2007; Hyenstrand et al. 1998a, 1998b) and the dominance of diatoms from January through May, NO$_3$-N may have been almost depleted and led to a higher ratio of NH$_4^+$-N: (NO$_3$-N + NO$_2$-N). A substantial increase in Chl-a from June to September was preceded by a sharp increase in NH$_4^+$-N: (NO$_3$-N + NO$_2$-N) from 0.19...
in June to 0.57. The ratio of NH$_4^+$-N:(NO$_3^-$-N + NO$_2^-$-N) increase from June to August was due to a small increase in NH$_4^+$-N concentration and a larger decrease in NO$_3^-$-N concentration.

Similar to the idea above that diatom may have depleted NO$_3^-$-N and NO$_2^-$-N prior to the summer cyanobacteria maximum, the observed decrease in NH$_4^+$-N:(NO$_3^-$-N + NO$_2^-$-N) from 0.57 in August to 0.42 in September may have been due to rapid NH$_4^+$-N uptake by the dominant cyanobacteria combined with NO$_3^-$-N and NO$_2^-$-N accumulation from runoff during the summer wet season and/or nitrification in the water column. As a reference, NH$_4^+$-N uptake accounted for 53% of total N uptake, with NO$_3^-$-N uptake accounting for only 19% during the summer in Lake Okeechobee (Gu et al. 1997).

During the proliferation process of Chl-a, the concentration of NH$_4^+$-N and NO$_2^-$-N have a little variation (Table 2), but the NO$_3^-$-N has undergone great changes.
Based on these results, algal growth may be more dependent on the NO$_3$-N supply in the Yuqiao Reservoir. It is not difficult to understand why Chl-a is more sensitive to NO$_3$-N than other parameters.

As described above, the sensitivity analyses demonstrate how the trained network reacts to changes of each input. Although NO$_3$-N is the most sensitive variable for Chl-a, the sensitivity of NO$_3$-N is similar to that of DO, TP and COD. Thus, further analysis is required to understand the impacts of input variables and to determine the optimal combination that is necessary for controlling algal bloom.
Environmental control of algal bloom

The control of nutrient loadings could be done in order to reduce the probability of algal bloom. However, due to the complicated chemical and biological processes involving the algal bloom as well as due to the changing environment, it is uncertain what the most effective way of reducing nutrient loading is. In this study, an orthogonal experimental design approach, which is a design method about multi-factor and multi-level research and which selects representative portion of the test points based on orthogonality for experiment from a comprehensive point (Taguchi 1987), was used with four selected most sensitive parameters NO$_3$-N, DO, TP and COD (Figure 10) with semimonthly changes of concentrations. These concentrations were sensitive to anthropogenic activities and could be managed; however, even WT has substantial influence on algal bloom, and it is not selected because it cannot be artificially regulated. Next, a Latin table with four factors and three empirically changed levels ($\frac{1}{10}$, 0 and 10%) was designed ($L_9(4^{3})$). From the nine trials that are presented in Table 3, the optimal combination of the four selected most sensitive parameters maintain a constant TP concentration while reducing the COD by 10% and increasing the NO$_3$-N and DO concentrations by 10% (No. 9), which resulted in a Chl-a decrease of 23.8%. However, if reducing the NO$_3$-N by 10%, and while improving the other three by 10% (No. 4) resulted in a Chl-a increase of 27.4%. These results provide valuable references for the sound environmental control of algal bloom in the Yuqiao Reservoir through human intervention, e.g. shutting off certain point sources and increasing aeration, etc.

CONCLUSIONS

The ANN model was applied to dynamic water quality modeling (regarding WT, TP and Chl-a) by using semimonthly measurements at four monitoring stations. Although multiple unknown factors controlling water quality variations still exist and the complex eutrophication processes are still not fully understood, the model provides a good fit between the measured and predicted values. This study has provided the following findings:

1. Based on the modeling approach described in this study for the analysis of the eutrophication problem in Yuqiao Reservoir, the Ward net architecture was more appropriate for modeling of WT, while the GRNN architecture was more appropriate for the modeling of total phosphorus and Chl-a.

2. Both water quality factors and meteorological factors were considered in the development of the ANN model. As well, the data back to $t$, $t-1$ and $t-2$ were all employed for the prediction of $t+1$ data. These approaches ensured a better coverage and better

### Table 2

<table>
<thead>
<tr>
<th>Month</th>
<th>NH$_4$-N (mg/L)</th>
<th>NO$_3$-N (mg/L)</th>
<th>NO$_2$-N (mg/L)</th>
<th>NH$_4$-N(NO$_3$-N-NO$_2$-N)</th>
<th>Chl-a (mg/L)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.18</td>
<td>1.74</td>
<td>0.04</td>
<td>0.10</td>
<td>0.0037</td>
</tr>
<tr>
<td>2</td>
<td>0.18</td>
<td>1.49</td>
<td>0.03</td>
<td>0.12</td>
<td>0.0035</td>
</tr>
<tr>
<td>3</td>
<td>0.20</td>
<td>1.46</td>
<td>0.02</td>
<td>0.14</td>
<td>0.0036</td>
</tr>
<tr>
<td>4</td>
<td>0.18</td>
<td>1.28</td>
<td>0.01</td>
<td>0.14</td>
<td>0.0042</td>
</tr>
<tr>
<td>5</td>
<td>0.16</td>
<td>0.90</td>
<td>0.02</td>
<td>0.17</td>
<td>0.0052</td>
</tr>
<tr>
<td>6</td>
<td>0.15</td>
<td>0.75</td>
<td>0.04</td>
<td>0.19</td>
<td>0.0071</td>
</tr>
<tr>
<td>7</td>
<td>0.19</td>
<td>0.72</td>
<td>0.04</td>
<td>0.25</td>
<td>0.0084</td>
</tr>
<tr>
<td>8</td>
<td>0.20</td>
<td>0.33</td>
<td>0.02</td>
<td>0.57</td>
<td>0.0122</td>
</tr>
<tr>
<td>9</td>
<td>0.23</td>
<td>0.51</td>
<td>0.03</td>
<td>0.42</td>
<td>0.0125</td>
</tr>
<tr>
<td>10</td>
<td>0.24</td>
<td>0.84</td>
<td>0.06</td>
<td>0.26</td>
<td>0.0085</td>
</tr>
<tr>
<td>11</td>
<td>0.27</td>
<td>1.17</td>
<td>0.05</td>
<td>0.22</td>
<td>0.0056</td>
</tr>
<tr>
<td>12</td>
<td>0.24</td>
<td>1.69</td>
<td>0.04</td>
<td>0.14</td>
<td>0.0048</td>
</tr>
</tbody>
</table>

### Table 3

<table>
<thead>
<tr>
<th>Changes in environmental factors (%)</th>
<th>Responses of Chl-a (%)</th>
<th>Mean±SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. NO$_3$-N/D/O/TP/COD</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1  -10 -10 -10 -10</td>
<td>-7.7±12.1</td>
<td></td>
</tr>
<tr>
<td>2  -10 0 0 0</td>
<td>-4.5±18.6</td>
<td></td>
</tr>
<tr>
<td>3  -10 10 10 10</td>
<td>27.4±29.5</td>
<td></td>
</tr>
<tr>
<td>4  0 -10 0 10</td>
<td>20.2±9.1</td>
<td></td>
</tr>
<tr>
<td>5  0 10 10 -10</td>
<td>6.7±17.8</td>
<td></td>
</tr>
<tr>
<td>6  0 10 10 -10 0</td>
<td>-10.1±18.1</td>
<td></td>
</tr>
<tr>
<td>7  10 -10 10 10</td>
<td>15.5±8.6</td>
<td></td>
</tr>
<tr>
<td>8  10 0 -10 10 10</td>
<td>-7.9±16.8</td>
<td></td>
</tr>
<tr>
<td>9  10 10 0 -10 10</td>
<td>-23.8±28.7</td>
<td></td>
</tr>
</tbody>
</table>

(No. 4) resulted in a Chl-a increase of 27.4%. These results provide valuable references for the sound environmental control of algal bloom in the Yuqiao Reservoir through human intervention, e.g. shutting off certain point sources and increasing aeration, etc.
temporal characteristics of data being applied in the forecasting and consequently ensure more accurate forecasting results.

3. For the performance of the ANN model, the Nash-Sutcliffe coefficient of efficiency ($R^2$) of the predicted and measured data was between 0.84 and 0.99 for the training and over-fitting test data sets, while it was between 0.59 and 0.99 for the validation data set.

4. With careful construction of the ANN model, the WT, the concentrations of total phosphorus and Chl-a can be forecasted for up to 2 weeks in advance with reasonable accuracy. The longer time in advance for forecasting, the better management opportunities for algae outbreak.

5. By the assessment through a sensitivity analysis, it is found that Chl-a is very sensitive to NO$_3$-N, DO, TP, COD and WT (in decreasing order).

6. The ANN model is also used to identify the optimal measure to control the algae bloom by an orthogonal experimental design approach. It provides valuable information for the decision-maker in order to take appropriate action when the forecasting shows a high probability of algae bloom. The case study in Yuqiao reservoir demonstrates the wide range of capabilities of the ANN model.

ACKNOWLEDGEMENTS

This research was financially supported by Ministry of Science and Technology (International Science & Technology Cooperation Program of China, 2013DFA71340), Ministry of Water Resources (201201114) and Ministry of Education (NCET-09-0586). We would like to express our thanks to the staff working in Tianjin Institute of Water Resources and Hydropower Research for their great efforts in collecting data to support this research. We also would like to thank Prof. Shie-Yui Liong at National University of Singapore for his valuable advice.

REFERENCES


First received 21 October 2013; accepted in revised form 27 November 2014. Available online 12 January 2015