Ecological Modelling 220 (2009) 888–895

Contents lists available at ScienceDirect

Ecological Modelling

journal homepage: www.elsevier.com/locate/ecolmodel

Artificial neural network modeling of the river water quality—A case study

Kunwar P. Singh*, Ankita Basant[1], Amrita Malik, Gunja Jain

Environmental Chemistry Division, Indian Institute of Toxicology Research, Post Box 80, MG Marg, Lucknow 226 001, India

ARTICLE INFO

Article history:
Received 16 June 2008
Received in revised form 9 January 2009
Accepted 15 January 2009

Keywords:
Artificial neural network
Feed-forward
Back propagation
Modeling
Water quality

ABSTRACT

The paper describes the training, validation and application of artificial neural network (ANN) models for computing the dissolved oxygen (DO) and biochemical oxygen demand (BOD) levels in the Gomti river (India). Two ANN models were identified, validated and tested for the computation of DO and BOD concentrations in the Gomti river water. Both the models employed eleven input water quality variables measured in river water over a period of 10 years each month at eight different sites. The performance of the ANN models was assessed through the coefficient of determination (R^2) (square of the correlation coefficient), root mean square error (RMSE) and bias computed from the measured and model computed values of the dependent variables. Goodness of the model fit to the data was also evaluated through the relationship between the residuals and model computed values of DO and BOD. The model computed values of DO and BOD by both the ANN models were in close agreement with their respective measured values in the river water. Relative importance and contribution of the input variables to the model output was evaluated through the partitioning approach. The identified ANN models can be used as tools for the computation of water quality parameters.

1. Introduction

The surface water quality in a region largely depends on the nature and extent of the industrial, agricultural and other anthropogenic activities in the catchments. Escalated anthropogenic activities in the basins and reduced river discharges registered during the last few decades have caused many-fold increase in the organic pollution load of the surface water bodies in India (Singh et al., 2004). The river systems are most adversely affected due to their dynamic nature and an easy accessibility for the waste disposal directly or indirectly through drains/tributaries. Since, the rivers and streams are among most important sources of water for irrigation, industrial and other uses, these serve as the lifelines of the population staying in the basins. In general, the organic pollution in an aquatic system is measured and expressed in terms of the biochemical oxygen demand (BOD) and declined dissolved oxygen (DO) level. The BOD measures an approximate amount of bio-degradable organic matter present in water and serves as an indicator parameter for the extent of water pollution. The BOD of any aquatic system is the foremost parameter needed for assessment of the water quality as well as development of management strategies for the protection of water resources. This warrants for a foolproof method for its determination. Currently available method for BOD determination is very tedious and prone to measurement errors (Beltran et al., 1998; Einax et al., 1998, 1999). The method is subject to various complicating factors such as the oxygen demand resulting from the respiration of algae in the sample and the possible oxidation of ammonia. Presence of toxic substances in sample may also affect microbial activity leading to a reduction in the measured BOD value. The laboratory conditions for BOD determination usually differ from those in aquatic systems. Therefore, interpretation of BOD results and their implications may be associated with large variations. Moreover, it is a time taking parameter determined over a period of 5 days under constant conditions of temperature (20 °C) through out, which is difficult to maintain in developing countries. Overall, the BOD measurement results are associated with large uncertainties, thus making its estimate more unreliable. Since, BOD is inversely related to the dissolved oxygen in water, the high values of the earlier indicate for a low level of the dissolved oxygen (DO) or even anoxic conditions in water. The DO level is measure of the health of the aquatic system and a certain level is essentially required for the aquatic life to survive. Moreover, for determination of BOD, pre-knowledge of DO concentration in water is essentially required. Therefore, both these parameters (DO-BOD) are generally needed to be determined simultaneously and there is a need to devise some suitable secondary (indirect) method for predicting these variables in a large number of samples for water quality assessment.

Although, parametric statistical and deterministic models have been the traditional approaches for modeling the water quality but these require vast information on various hydrological subprocesses in order to arrive the end results. In recent years, several

* Corresponding author. Tel.: +91 522 2436077; fax: +91 522 2628227.
E-mail addresses: kpsingh_52@yahoo.com, kunwarpsingh@gmail.com (K.P. Singh).
[1] Current address: Department of Molecular Biosciences, UMDNJ, USA.

0304-3800/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.ecolmodel.2009.01.004

researches have been conducted on water quality forecast models (Chen et al., 2003; Kurunc et al., 2005; Li, 2006). However, since a large number of factors affecting the water quality have a complicated non-linear relation with the variables; traditional data processing methods are no longer good enough for solving the problem (Wu et al., 2000; Xiang et al., 2006). On the other hand, the artificial neural networks (ANNs) capable of imitating the basic characteristics of the human brain such as self-adaptability, self-organization and error tolerant and have been widely adopted for model identification, analysis and forecast, system recognition and design optimization (Niu et al., 2006; Shu, 2006). Unlike many statistically based water quality models, which assume a linear relationship between response and prediction variables and their normal distribution, ANNs are able to map the non-linear relationships that are characteristics of aquatic eco-systems (Lek et al., 1996). During last about two decades, ANNs have undergone an explosive development in application in almost all the areas of research (Rahim et al., 1993; Chu and Bose, 1998; Kung and Taur, 1995; Smits et al., 1992; Lerner et al., 1994; Lo et al., 1995; Messikh et al., 2007; Cabreta-Mercader and Staelin, 1995; Hanbay et al., 2008). The ANN approach has several advantages over traditional phenomenological or semi-empirical models, since they require known input data set without any assumptions (Gardner and Dorling, 1998). The ANN develops a mapping of the input and output variables, which can subsequently be used to predict desired output as a function of suitable inputs (Schalkoff, 1992). A multi-layer neural network can approximate any smooth, measurable function between input and output vectors by selecting a suitable set of connecting weights and transfer functions (Gardner and Dorling, 1998). ANN models have been widely applied to the water quality problems (Rogers and Dowla, 1994; Raman and Chandramouli, 1996; Wen and Lee, 1998; Lek and Guegan, 1999; Bowers and Shedrow, 2000; Kuo et al., 2004, 2007).

The main aim of the present work is to construct an artificial neural network (ANN) model of the Gomti river water quality (DO-BOD) and demonstrate its application to complex water quality data as how it can improve the interpretation of the results. Here, we have investigated the possibility of training ANN models correlating the primary water quality variables (independent) with their secondary attribute (dependent variable). The DO and BOD of the river water were taken as the dependent variables here and set of other parameters constituted the independent variables. In this study, ANN models have been identified for computing the DO and BOD of the river water.

2. Methodology

2.1. Water quality data set

The data set used in this study was generated through continuous monitoring of the water quality of Gomti, a polluted river flowing through the northern alluvial Gangetic plains in India (Singh et al., 2007). The river during its course receives low to very high pollution load from various diffuse and point sources in its different stretches while flowing through urban townships, thus exhibiting very large variations in water quality variables. The sampling sites are selected such that these are in low, moderate and high polluted regions (Singh et al., 2004). The first three sites are located in the area of relatively low river pollution and are upstream of the Lucknow city. Other three sites are located in the region of high river pollution as there are a number of wastewater drains (26 nos.) and two highly polluted tributaries emptying into the river in this stretch. The last two sites are in the downstream region of moderate pollution as the river considerably recovers in the course. The river water quality was monitored regularly each month at eight different sites over a period of 10 years (January 1994–December 1999 and January 2002–December 2005). The sampling sites are spread over a distance of about 500 km and the river drains an area of about 25,000 km^2. Grab water samples were collected from a depth of 15 cm below the surface at eight different sampling sites on the river. All the water samples collected during the study period were analyzed for 19 different parameters. Details on sampling network and analytical procedures are available elsewhere (Singh et al., 2004). Here for ANN modeling, the following 13 parameters have been included: water pH, total alkalinity (T-Alk, mg L^{-1}), total hardness (T-Hard, mg L^{-1}), total solids (TS, mg L^{-1}), chemical oxygen demand (COD, mg L^{-1}), ammonical nitrogen (NH_4-N, mg L^{-1}), nitrate nitrogen (NO_3-N, mg L^{-1}), chloride (Cl, mg L^{-1}), phosphate (PO_4, mg L^{-1}), potassium (K, mg L^{-1}), sodium (Na, mg L^{-1}), dissolved oxygen (DO, mg L^{-1}), and 5-day biochemical oxygen demand (BOD, mg L^{-1}). The analytical data quality was ensured through careful standardization, procedural blank measurements, spiked and duplicate samples. The laboratory also participated in regular national program on analytical quality control (AQC). Here, the river water DO and BOD constituted the dependent data array, while all the remaining (eleven) variables as the independent data array sets.

2.2. Artificial neural networks modeling

The artificial neural network, as the name implies, employs the model structure of a neural network which is very powerful computational technique for modeling complex non-linear relationships particularly in situations where the explicit form of the relation between the variables involved is unknown (Gallant, 1993; Smith, 1994). The basic structure of an ANN model is usually comprised of three distinctive layers, the input layer, where the data are introduced to the model and computation of the weighted sum of the input is performed, the hidden layer or layers, where data are processed, and the output layer, where the results of ANN are produced. Each layer consists of one or more basic element(s) called a neuron or a node. A neuron is a non-linear algebraic function, parameterized with boundary values (Dreyfus et al., 2002). The signal passing through the neuron is modified by weights and transfer functions. This process is repeated until the output layer is reached (Govindaraju, 2000). The number of neurons in the input, hidden and output layers depends on the problem. If the number of hidden neurons is small, the network may not have sufficient degrees of freedom to learn the process correctly. On the other hand, if the number is too high, the training will take a longer time and the network may over-fit the data (Karunanithi et al., 1994).

In this study, three-layer feed-forward neural networks with back propagation (BP) learning were constructed for computation of the river water DO and BOD with eleven input variables, $p(x_{pi}$, $i = 1, \ldots, 11)$, as shown in Fig. 1. A feed-forward neural network (FFNN) is very powerful in function optimization modeling and has extensively been used for the prediction of water resources variables (Palani et al., 2008). A single hidden layer was used in both the networks.

All the computations were performed using the EXCEL 97 and MATLAB (MathWorks, Inc., Natwick, MA).

2.2.1. Back propagation neural network and learning algorithm

The back propagation (BP) is a commonly used learning algorithm in ANN application. It uses the back propagation (BP) of the error gradient. This training algorithm is a technique that helps distribute the error in order to arrive at a best fit or minimum error. After the information has gone through the network in a forward direction and the network has predicted an output, the back propagation algorithm redistributes the error associated with this output back through the model, and weights are adjusted accordingly. Minimization of the error is achieved through sev-

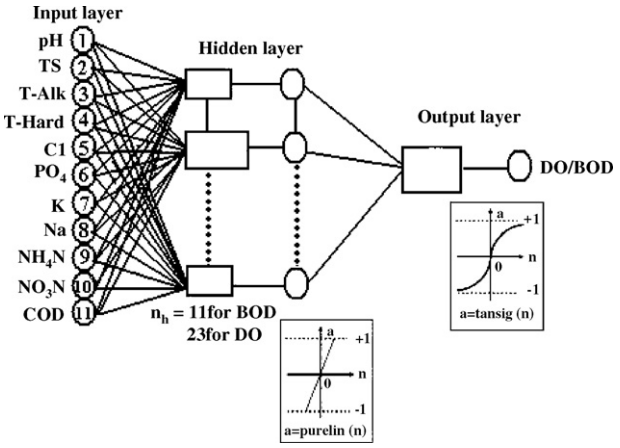

Fig. 1. General conceptual neural network for the DO and BOD computation in the Gomti river water.

eral iterations. One complete cycle is known as the 'epoch'. Each neuron in a layer is connected to every neuron in the next layer. These links are given a synaptic weight that represents its connection strength (Govindaraju, 2000). Although, traditional BP uses a gradient descent algorithm to determine the weights in the network, it computes rather slowly due to linear convergence. Hence, Levenberg–Marquardt algorithm (LMA), which is much faster as it adopts the method of approximate second derivative (Wang, 2004) was used here. The LMA is similar to the quasi-Newton method in which a simplified form of the Hessian matrix (second derivative) is used. The Hessian matrix can be approximated as;

$$H = J^{\mathrm{T}}J \tag{1}$$

and the gradient can be computed as

$$g = J^{\mathrm{T}}e \tag{2}$$

in which J is the Jacobian matrix which contains first derivatives of the network errors with respect to the weights and biases, and e is a vector of network errors. One iteration of this algorithm can be written as

$$x_{k+1} = x_k - [J^{\mathrm{T}}J + \mu I]^{-1}J^{T}e \tag{3}$$

where μ is the learning rate and I is the identity matrix (Dedecker et al., 2004). During training the learning rate μ is incremented or decremented by a scale at weight updates. When μ is zero, this is just Newton's method, using the approximate Hessain matrix. When μ is large, this becomes gradient descent with a small step size. The LMA is reported to have the fastest convergence for the neural networks that contain up to few hundred neurons (Karul et al., 2000).

During the training process, there are three factors that are associated with the weight optimization algorithms. These are: (1) initial weight matrix, (2) learning rate, and (3) stopping criteria such as (3a) fixing of the number of epoch size, (3b) setting a target error goal and (3c) fixing minimum performance gradient. The initial weights are randomly generated between −1 and +1 with a random number generator. The learning rate is an indicator of the rate of convergence. If it is too small, the rate of convergence will be slow due to the large number of steps needed to reach the minimum error. If it is too large, the convergence initially will be fast, but will produce undue oscillations, and may not reach the minimum error. The value of the learning parameter is not fixed as the optimization of learning parameter is highly problem dependent and should be selected so that oscillations in error surface can be avoided (Maier and Dandy, 1998, 2000). Hagan et al. (1996) demonstrated that the learning becomes unstable for higher values (>0.035). Thus, the learning rate was set to 0.001.

The mean square error (MSE), used as the target error goal, is defined as (Hagan et al., 1996; Karul et al., 2000);

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_{pi} = O_i)^2 \tag{4}$$

where y_{pi} and O_i represent the model computed and measured values of the variable, and N represents the number of observations. The maximum number of epochs, target error goal MSE and the minimum performance gradient were set to 300, 10^{-10}, and 10^{-10}, respectively. Training stops when any of these conditions occur.

2.2.2. Optimization of the ANN structure

The optimal architecture of the ANN models and its parameter variation were determined based on the minimum value of the mean squared error (MSE) of the training and validation sets. In optimization of the networks, four neurons were used in the hidden layer as an initial guess. With increase in number of neurons, the networks yielded several local minimum values with different MSE values for the training set. Through trial and error approach, various combinations of the number of neurons in hidden layer, back propagation algorithms, and transfer functions (linear and sigmoid) were used. Lowest MSE for the training and the validation sets was the criteria for selecting the best case (Karul et al., 2000).

2.2.3. Modeling performance criteria

To determine the performance of each of the selected network model, three different criteria were used: the root mean square error (RMSE), the bias, and the coefficient of determination (R^2) (Chenard and Caissie, 2008). The RMSE represents the error associated with the model and can be computed as:

$$\mathrm{RMSE} = \sqrt{\sum_{i=1}^{N}\frac{(y_{pi} - O_i)^2}{N}} \tag{5}$$

where y_{pi} and O_i represent the model computed and measured values of the variable, and N represents the number of observations. The RMSE, a measure of the goodness-of-fit, best describes an average measure of the error in predicting the dependent variable. However, it does not provide any information on phase differences. The bias represents the mean of all the individual errors and indicates whether the model overestimates or underestimates the dependent variable. It is calculated as:

$$\mathrm{Bias} = \frac{1}{N}\sum_{i=1}^{N}(y_{pi} - O_i) \tag{6}$$

The coefficient of determination (R^2) represents the percentage of variability that can be explained by the model and is calculated as:

$$R^2 = \left[\frac{N\sum_{i=1}^{N}O_i y_{pi} - \left(\sum_{i=1}^{N}O_i\right)\left(\sum_{i=1}^{N}y_{pi}\right)}{\sqrt{\left[N\sum_{i=1}^{N}O_i^2 - \left(\sum_{i=1}^{N}O_i\right)^2\right] \times \left[N\sum_{i=1}^{N}y_{pi}^2 - \left(\sum_{i=1}^{N}y_{pi}\right)^2\right]}}\right]^2 \tag{7}$$

Adequacy of the created models was evaluated through enumerating the relative extent of participation of various input variables in the model performance. Fitness of the created ANN models was checked through analysis of the residuals.

In a dependent–independent variable modeling approach, it is desirable to assess the relative importance and contribution of each of the independent variables in a model and in subsequent computation prediction of the dependent variable. Here, we used

partitioning approach to express the importance of independent variables for the output layer rather than a single output node, functioning as to partition the sum of effects on the output layer (Garson, 1998). Here, the general network consists of eleven environmental variables. The importance of each variable can be expressed as (Lee et al., 2003):

$$I = \frac{\sum_{j=1}^{n_h} \mathrm{ABS}(w_{ji})}{\sum_{k=1}^{n_v} (\sum_{j=1}^{n_h} \mathrm{ABS}(w_{ji}))_k} \quad (8)$$

where n_h is the number of hidden nodes, n_v is the number of input variables, w_{ji} is the connection weight from the ith input node to jth hidden node, and ABS denotes the absolute value of the function. Eq. (8) reflects the relative weights of each input variable to the whole hidden layer.

2.2.4. Input variables and data processing

The monthly data of thirteen water quality parameters measured over a period of 10 years at all the eight sampling sites were selected for this analysis. The basic statistics of the aforesaid variables is presented in Table 1. The DO and BOD are two major parameters in water quality assessment. Based on existing measured values of different variables and their correlative analysis, total 11 factors (variables) including pH, T-Alk, T-Hard, TS, COD, NH_4-N, NO_3-N, Cl, PO_4, K and Na were identified which affect the water quality (DO and BOD) to certain degree and finally selected for the model development. The spatial (site-wise) data sets were statistically compared for any significant difference among them using F-test ($p = 0.05$) for each of the measured variables. The calculated values of the test statistics ($F_{calculated}$) were higher than the respective critical value at 95% probability level ($F_{calculated} > F_{critical}$) in case of all the variables indicating that there is a significant difference among the sites. The difference among the sites may be attributed to the fact that in the upper stretches (sites 1–3), the river water quality is mainly dominated by the variables of geogenic origin, as there are no major pollution sources in the region. In the mid-course (sites 4–6), the river receives heavy load of untreated wastewater through twenty-six drains in Lucknow city and two major tributaries emptying into the river carrying mixed untreated domestic and industrial wastewater from nearby towns. Therefore, the water quality in this stretch is largely dominated by the variables of anthropogenic origin. However, in the lower stretch (sites 7 and 8), the river through natural processes and by virtue of its recharging nature recovers considerably, exhibiting moderate level of various water quality variables (Singh et al., 2004, 2007).

Concentration of both the dependent variables showed large variations between the samples, with a high coefficient of variation (47.5% for DO and 82.6% for BOD). The coefficient of variation (CV), a measure of statistical dispersion of data, is the mean normalized standard deviation of the given data set. It (CV%) is computed as (standard deviation/mean) × 100. The large variation in concentration of the dependent variables corresponds to the nature and types (point and non-point) of sources distributed in the large geographical area of the river basin. The river during its course passes through several townships and a number of wastewater drains and tributaries pour huge quantities of untreated wastewater into the main channel of the river. The independent variables also showed a coefficient of variation between 3.7% and 120%. Such variability among the samples may be attributed to the large geographical variations in climate and seasonal influences in the study region. pH showed lowest variation and it may be due to the buffering capacity of the river. Variables of anthropogenic origin showed larger variations as compared to those of the natural origin variables (Table 1). It may be attributed to the fact that the geogenic processes are almost in equilibrium state, whereas, the anthropogenic processes are time dependent in nature.

In this study, ANNs were identified to predict the water quality (DO and BOD) of the Gomti river (India). For ANN identification, the complete river water quality data (independent variables) set of 10 years (960 samples × 11 variables) was divided in to three sub-sets. The calibration (or training), validation and test data subsets comprised of 576 (60%), 192 (20%) and 192 (20%) samples each, respectively. Thus, the data of first 6 years were taken as the training data, another 2 years as the validation and remaining last 2 years data as the test set. Finally, for the model input (independent variables), the training, validation and test data sets have dimensions of 576 samples × 11 variables, 192 samples × 11 variables, and 192 samples × 11 variables, respectively. The output variables (DO, BOD) corresponding to the input variables belong to the same water sample, thus measured in same time and space.

In view of the requirements of the neural computation algorithm, the raw data of both the independent and dependent variables were normalized to an interval by transformation. The transformation modifies the distribution of the input variables so that it matches the distribution of the estimated outputs. Here, all the variables are transformed to the same ground-uniform distributions on −1, +1. The ANNs were applied to provide a non-linear relationship between sets of inputs comprised of some selected characteristic variables of river water quality and the network output (DO and BOD of the river water).

3. Results and discussion

3.1. Artificial neural network (ANN)

Different ANN models were constructed and tested in order to determine the optimum number of nodes in the hidden layer and transfer functions. Selection of an appropriate number of nodes in

Table 1
Basic statistics of the measured water quality variables in Gomti river water, India ($n = 960$).

Variable	Unit	Min	Max	Median	Mean	SD	CV%
pH	–	6.02	9.03	8.34	8.29	0.31	3.72
TS	mg L^{-1}	60.00	586.00	290.70	301.21	70.26	23.33
T-Alk	mg L^{-1}	75.33	346.70	215.15	204.69	52.20	25.50
T-Hard	mg L^{-1}	28.00	324.00	188.33	178.86	48.04	26.86
Cl	mg L^{-1}	0.21	34.33	7.33	8.73	5.61	64.29
PO_4	mg L^{-1}	0.01	2.07	0.19	0.31	0.34	109.23
K	mg L^{-1}	1.78	13.65	4.70	5.36	2.19	40.85
Na	mg L^{-1}	4.67	83.10	33.20	34.26	14.19	41.41
NH_4-N	mg L^{-1}	0.01	2.92	0.25	0.45	0.54	120.09
NO_3-N	mg L^{-1}	0.01	2.65	0.46	0.69	0.62	90.05
COD	mg L^{-1}	1.20	57.39	14.70	17.46	10.14	58.06
BOD	mg L^{-1}	0.12	31.67	4.23	6.18	5.10	82.59
DO	mg L^{-1}	0.00	9.97	6.00	5.61	2.66	47.47

SD, Standard deviation; CV, coefficient of variation.

Table 2
Performance parameters of the artificial neural network models for computation of the DO and BOD in Gomti river water (India).

Model	ANN-structure		RMSE	Bias	R^2
DO	11 → 23 → 1	Training	1.5	−0.05	0.70
		Validation	1.44	0.49	0.74
		Test	1.23	−0.43	0.76
BOD	11 → 11 → 1	Training	2.25	0.14	0.85
		Validation	1.84	0.87	0.85
		Test	1.38	−0.22	0.77

the hidden layer is very important aspect as a larger number of these may result in over-fitting, while a smaller number of nodes may not capture the information adequately. Fletcher and Goss (1993) suggested that the appropriate number of nodes in a hidden layer ranges from $(2n^{1/2}+m)$ to $(2n+1)$, where n is the number of input nodes and m is the number of output nodes. Subsequently, two different ANN models were constructed for the computation of DO and BOD in the river water. The network was trained using the training data set, and then it was validated with the validation data set. The optimal network size was selected from the one which resulted in minimum mean square error (MSE) in training and validation data sets. The model features for both the ANNs are given in Table 2.

3.1.1. DO and BOD models

The architecture of the best ANN models for the dissolved oxygen (DO) and biochemical oxygen demand (BOD) in the river water is shown in Fig. 1. The selected ANN for the DO model is composed of one input layer with eleven input variables, one hidden layer with twenty three nodes and one output layer with one output variable, whereas, the BOD model differed in number of nodes in the hidden layer, as it optimized with eleven nodes in this layer. The constructed ANN models (DO and BOD) were trained using the Levenberg–Marquardt algorithm (LMA). The LMA is much faster than other algorithms used in BP (Ying et al., 2007). In both of these ANNs, linear transfer function (purelin) was used in the hidden layer and a non-linear transfer function (tansig) in the outer one. The coefficient of determination (R^2), RMSE, and the bias as computed for the training, validation and test data sets used for the two models (DO and BOD) are presented in Table 2. Fig. 2a–c shows the plots between measured and model computed values of DO in training, validation and testing sets. The selected ANN (11 nodes in input layer, 23 nodes in hidden layer, and single node in output layer) provided a best fit model for all the three data sets. The coefficient of determination (R^2) values ($p < 0.001$) for the training, validation and test sets were 0.70, 0.74, and 0.76, respectively. The respective values of RMSE and bias for the three data sets are 1.50 and −0.05 for training, 1.44 and 0.49 for validation, and 1.23 and −0.43 for testing. A closely followed pattern of variation by the measured and model computed DO concentrations in river water (Fig. 2a–c), R^2, RMSE and bias values suggest for a good-fit of the DO model to the data set.

Further, the model computed DO values (DO_{pred}) and residuals corresponding to the training, validation and testing sets are plotted in Fig. 3a–c. The observed relationship between residuals and model computed DO values for all the three sets shows complete independence and random distribution. It is further supported by the respective correlations ($n = 576$, $R^2 = 0.001$ for training; $n = 192$, $R^2 = 0.011$ for validation; and $n = 192$, $R^2 = 0.006$ for testing) which are negligible small. Fig. 3a–c shows that the points are well distributed on both sides of the horizontal line of zero ordinate representing the average of the residuals. Plots of the residuals versus model computed values can be more informative regarding model fitting to a data set. If the residuals appear to behave randomly it suggests that the model fits the data well. On the other hand, if non-random distribution is evident in the residuals, the model does not fit the data adequately (NIST/SEMATECH, 2006; Singh et al., 2007).

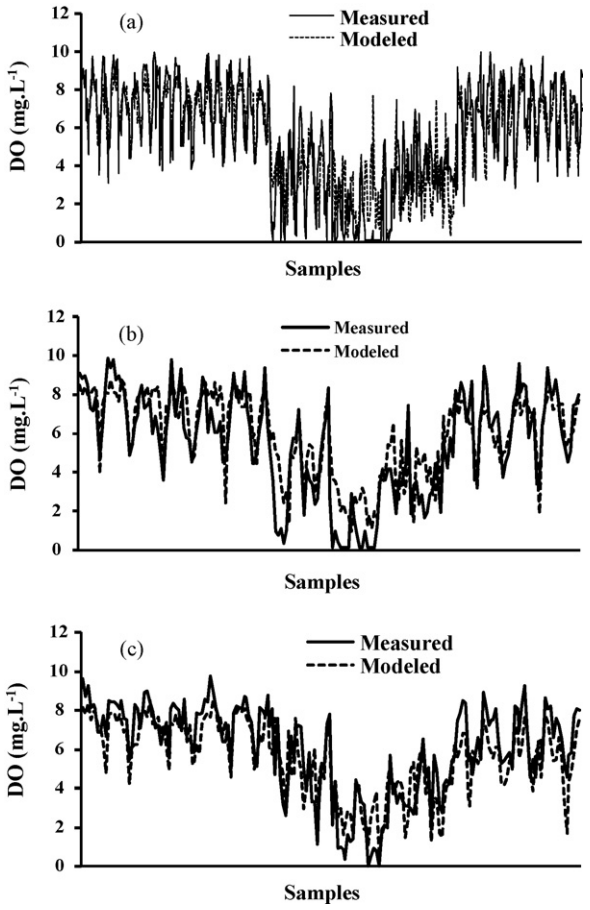

Fig. 2. Comparison of the model computed and measured DO levels in the river water (a) training, (b) validation, and (c) testing sets using DO-ANN model.

In case of the BOD, the selected ANN (11 nodes each in input and hidden layers and single node in output layer) provided a best fit model for all the three (training, validation and test) sets. Fig. 4a–c shows the plots between the measured and model computed values of BOD in training, validation and testing sets. The coefficient of determination (R^2) values ($p < 0.001$) for the training, validation and test sets were 0.85, 0.85, and 0.77, respectively. The respective values of RMSE and bias for the three data sets are 2.25 and 0.14 for training, 1.84 and 0.87 for validation, and 1.38 and −0.22 for testing (Table 2). A closely followed pattern of variation by the measured and model computed BOD values (Fig. 4a–c), R^2 and RMSE values suggest for a good-fit of the selected BOD model to the data set. Further, the model computed BOD values (BOD_{pred}) and residuals corresponding to the training, validation and testing sets are plotted in Fig. 5a–c. The observed relationship between residuals and model computed BOD values for all the three sets shows complete independence and random distribution. It is further supported by the negligible small correlations ($n = 576$, $R^2 = 0.000$ for training; $n = 192$, $R^2 = 0.002$ for validation; and $n = 192$, $R^2 = 0.032$ for testing). Fig. 5a–c shows that the points are well distributed on both sides of the horizontal line of zero ordinate representing the average of the residuals suggesting that the model fits the data well.

In recent years, ANNs have been used for prediction and forecasting of the water quality variables (Palani et al., 2008). Sengorur et al. (2006) employed the feed-forward back propagation ANN approach to estimate the monthly concentration of DO in Melen river water (Turkey). Although, a limited number of input variables (NO_2-N,

NO_3-N, BOD, flow, temperature) was used, the coefficient of determination between the measured and model computed values of DO was reasonably high (0.92). Soyupak et al. (2003) applied the FF-BP (LMA) NN approach to compute the pseudo steady state time and space dependent concentrations of DO in three different reservoirs (Turkey) using a limited number of input variables. The correlation coefficient values of >0.95 between the measured and computed DO concentrations in all the three cases suggested for the adequacy of the constructed models. Kuo et al. (2007) computed DO levels in the Te-Chi reservoir (Taiwan) water using a three-layered BP NN model. The input variables were the month, chlorophyll-a, pH, NH_4-N, NO_3-N, and water temperature. Correlation coefficients of 0.75 for the training and 0.72 for the test set were reported between the measured and computed values of DO. Ying et al. (2007) constructed a three-layer BP-LMA neural network for simultaneous computation of COD and DO levels in the Yuqiao reservoir (China) water using eight input variables (water temperature, turbidity, pH, alkalinity, chloride, NH_4-N, NO_2-N and hardness) measured over a short time period of 3 months. A correlation coefficient of 0.94 was reported between the measured and model computed DO values. Dogan et al. (2009) designed three-layered feed-forward ANN models to estimate the BOD in the Melen river water (Turkey) using eight parameters (COD, NH_3-N, chlorophyll-a, NO_2-N, NO_3-N, DO, flow, and water temperature) as input variables. Separate models were constructed for each of the eight input variables as well as in combinations. A correlation coefficient of 0.87 was reported between the measured and model computed BOD values. The sensitivity analysis results revealed that COD was the most effective input variables in the model.

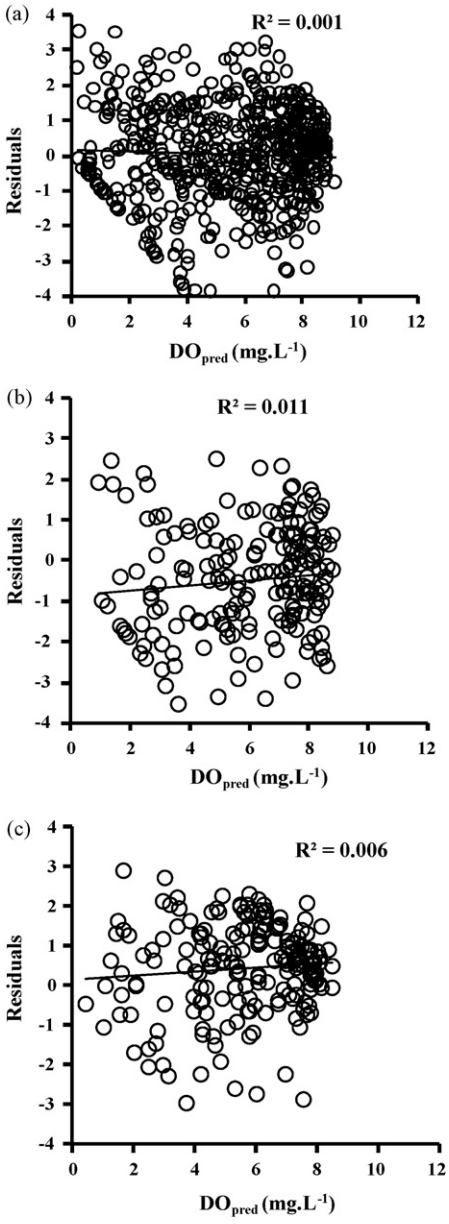

Fig. 3. Plot of the residuals versus model computed values of DO in river water (a) training, (b) validation, and (c) testing sets.

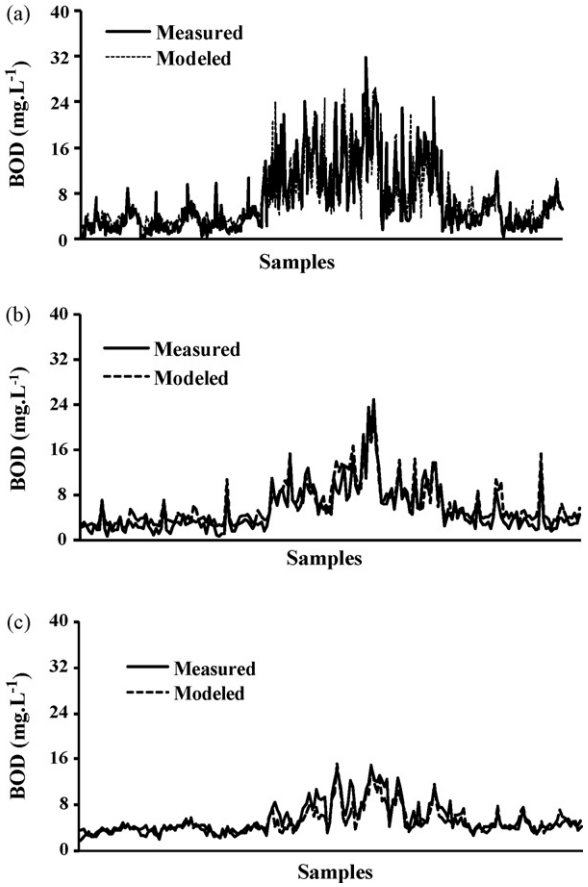

Fig. 4. Comparison of the model computed and measured BOD values in the river water (a) training, (b) validation, and (c) testing sets using BOD-ANN model.

A relatively low correlation (0.70–0.85) between the measured and model computed output variables (DO and BOD) in the present study may be due to the non-homogenous nature of the water quality (input and output) variables as these were measured over a long period of 10 years and at sampling sites distributed over a large geographical area. Moreover, relatively higher correlations between measured and model (NN) computed values of DO and BOD in various aquatic systems (Sengorur et al., 2006; Soyupak et al., 2003; Ying et al., 2007; Dogan et al., 2009) may be attributed to the relatively smaller data sets (samples) as well as limited number of the input variables used.

A relatively better performance (R^2 between measured and computed values) of the BOD model as compared to that of the DO model suggests that the selected affecting factors (input variables) have relatively greater impact on BOD than on DO. Also, selection of the affecting factors might affect the model output remarkably (Ying et al., 2007). Several computed values of DO are more deviated from actual measured values in river water (Fig. 2) due to the fact that the computed values may be affected by many factors during the

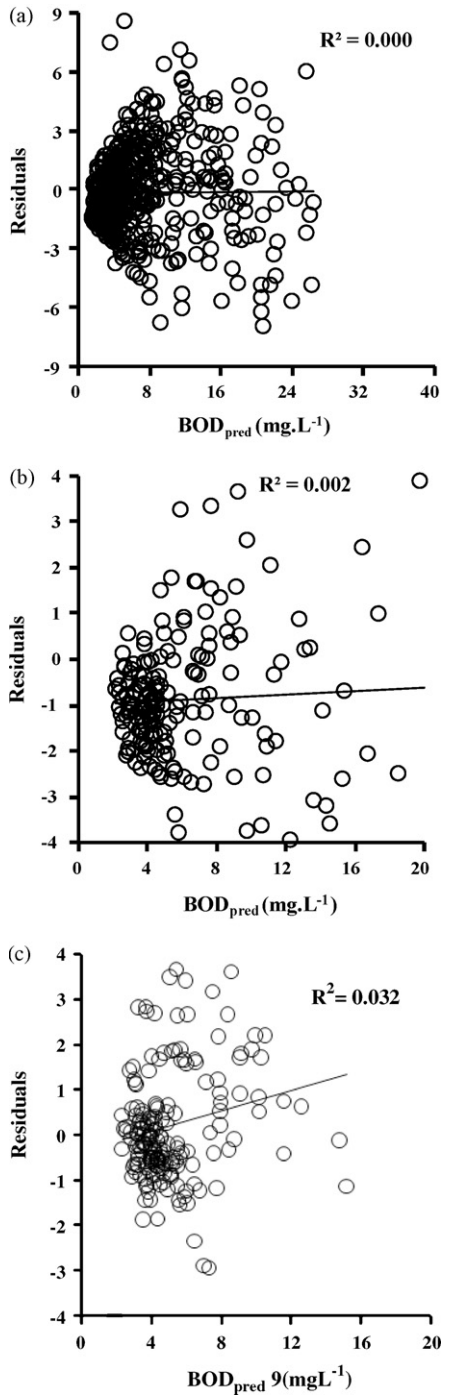

Fig. 5. Plot of the residuals versus model computed values of BOD in river water (a) training, (b) validation, and (c) testing sets.

study. The sources of DO in a water body include re-aeration from the atmosphere, photosynthetic oxygen production and DO loading. The sinks include oxidation of carbonaceous and nitrogenous materials, sediment oxygen demand, and respiration by aquatic plants (Kuo et al., 2007).

3.1.2. Relative importance of input variables in ANN models

Relative importance of each of the input variable was computed using Eq. (8). The results are presented in Fig. 6. It is evident that in DO model, although, all the input variables participated (7.4–10.5%) significantly, PO_4, NH_4-N, NO_3-N and COD provided relatively higher contributions to the network. It may be noted

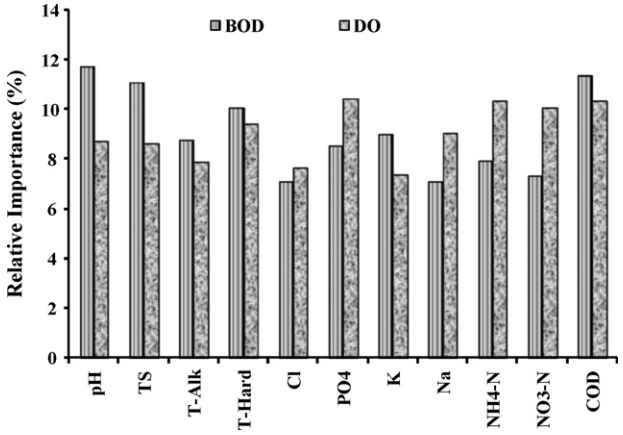

Fig. 6. Plot showing the relative importance of the input variables to the DO and BOD ANN models for the Gomti river.

that these include the oxygen containing (PO_4, NO_3-N) or oxygen demanding (NH_4-N, COD) variables. Although, the network does not necessarily represent physical meaning through the weights, it suggests that all the four variables have direct relevance with the dependent variable in water and play significant role in determining the DO level. In case of the BOD-model, the contribution of different input variables was between 7.1% and 11.7% and pH, TS and COD had relatively higher contributions (Fig. 6) suggesting their direct influence on BOD level in water. These relationships are evident as high levels of the organic matter consume large amounts of oxygen, which undergoes anaerobic fermentation processes leading to formation of ammonia and organic acids. Hydrolysis of these acidic materials causes a decrease of water pH values (Singh et al., 2004).

4. Conclusions

In this paper, two models based on artificial neural networks were identified for computation of the DO and BOD concentrations in water of the Gomti river (India). The identified models were trained, validated and tested on monthly data of DO and BOD measured over a period of 10 years. The feed-forward network with back propagation learning algorithm was employed. The present study shows that the optimal networks are capable to capture long-term trends observed for the tedious water quality variables (DO and BOD), both in time and space. We propose the neural networks as effective tool for the computation of river water quality and it could also be used in other areas to improve the understanding of river pollution trends. The ANN can be seen to be a powerful predictive alternative to traditional modeling techniques.

Acknowledgements

The authors thank the Director, Indian Institute of Toxicology Research, Lucknow (India) for his keen interest in the work. Financial assistance from CSIR, New Delhi is thankfully acknowledged.

References

Beltran, J.L., Ferrer, R., Guiteras, J., 1998. Multivariate calibration of polycyclic aromatic hydrocarbon mixtures from excitation–emission fluorescence spectra. Anal. Chim. Acta 373, 311–319.

Bowers, J.A., Shedrow, C.B., 2000. Predicting stream water quality using artificial neural networks. WSRC-MS-2000-00112. http://www.osti.gov/bridge/.

Cabreta-Mercader, C.R., Staelin, D.H., 1995. Passive microwave relative humidity retrievals using feedforward neural network. IEEE Trans. Geo. Remote Sens. 33, 842–852.

Chen, J.C., Chang, N.B., Shieh, W.K., 2003. Assessing wastewater reclamation potential by neural network model. Eng. Appl. Artif. Intell. 16, 149–157.

Chenard, J.F., Caissie, D., 2008. Stream temperature modelling using neural networks: application on Catamaran Brook, New Brunswick, Canada. Hydrol. Process. (DOI: 1002/hyp.6928).
Chu, W.C., Bose, N.K., 1998. Speech signal prediction using feedforward neural network. Electro. Lett. 34, 999–1001.
Dedecker, A.P., Goethals, P.L.M., Gabriels, W., De Pauw, N., 2004. Optimization of artificial neural network (ANN) model design for prediction of macroinvertebrates in the Awalm river basin (Flanders Belgium). Ecol. Model. 174, 161–173.
Dogan, E., Sengorur, B., Koklu, R., 2009. Modeling biochemical oxygen demand of the Melen River in Turkey using an artificial neural network technique. J. Environ. Manag. 90, 1229–1235.
Dreyfus, G., Martinez, J.-M., Samuelides, M., Gordon, M.B., Badran, F., Thiria, S., Herault, L., 2002. Reseaux de Neurones: Methodologie et Applications. Editions Eyrolles, Paris, France.
Einax, J.W., Kampe, O., Truckenbrodt, D., 1998. Assessing the deposition and remobilisation behavior of metals between river water and river sediment using partial least squares regression. Fres. J. Anal. Chem. 361, 149–154.
Einax, J.W., Aulinger, A., Tumpling, W.V., Prange, A., 1999. Quantitative description of element concentrations in longitudinal river profiles by multiway PLS models. Fres. J. Anal. Chem. 363, 655–661.
Fletcher, D., Goss, E., 1993. Forecasting with neural networks: an application using bankruptcy data. Inform. Manage. 24, 159–167.
Gallant, S.I., 1993. Neural Network Learning and Expert Systems. The MIT press, Massachusetts, USA.
Gardner, M.W., Dorling, S.R., 1998. Artificial neural network: the multilayer perceptron: a review of applications in atmospheric sciences. Atmos. Environ. 32, 2627–2636.
Garson, G.D., 1998. Neural Networks an Introductory Guide for Social Scientists. Sage Publications, California.
Govindaraju, R.S., 2000. Artificial neural network in hydrology. II: hydrologic application, ASCE task committee application of artificial neural networks in hydrology. J. Hydrol. Eng. 5, 124–137.
Hagan, M.T., Demuth, H.P., Beale, M., 1996. Neural Networks Design. PWS Publishing, Boston, MA, USA.
Hanbay, D., Turkoglu, I., Demir, Y., 2008. Prediction of wastewater treatment plant performance based on wavelet packet decomposition and neural networks. Expert Syst. Appl. 34, 1038–1043.
Karul, C., Soyupak, S., Cilesiz, A.F., Akbay, N., German, E., 2000. Case studies on the use of neural networks in eutriphication modeling. Ecol. Model. 134, 145–152.
Karunanithi, N., Grenney, W.J., Whitley, D., Bovee, K., 1994. Neural networks for river flow prediction. ASCE J. Comput. Civil Eng. 8, 210–220.
Kung, S.Y., Taur, J.S., 1995. Decision-based neural networks with signal image classification applications. IEEE Trans. Neural Netw. 6, 170–181.
Kuo, Y., Liu, C., Lin, K.H., 2004. Evaluation of the ability of an artificial neural network model to assess the variation of groundwater quality in an area of blackfoot disease in Taiwan. Water Res. 38, 148–158.
Kuo, J., Hsieh, M., Lung, W., She, N., 2007. Using artificial neural network for reservoir eutriphication prediction. Ecol. Model. 200, 171–177.
Kurunc, A., Yurekli, K., Cevik, O., 2005. Performance of two stochastic approaches for forecasting water quality and streamflow data from Yesilirmak River, Turkey. Environ. Model. Soft. 20, 1195–1200.
Lee, J.H.W., Huang, Y., Dickman, M., Jayawardena, A.W., 2003. Neural network modeling of coastal algal bloom. Ecol. Model. 159, 179–201.
Lek, S., Delacoste, M., Baran, P., Dimopoulos, I., Lauga, J., Aulagnier, S., 1996. Application of neural networks to modelling nonlinear relationships in ecology. Ecol. Model. 90, 39–52.
Lek, S., Guegan, J.F., 1999. Artificial neural networks as a tool in ecological modelling, an introduction. Ecol. Model. 120, 65–73.
Lerner, B., Levinstein, M., Rosenberg, B., Guterman, H., Dinsteun, I., Romen, Y., 1994. Feature selection and chromosomes classification using a multilayer perceptron neural network. In: IEEE International Conference on Neural Networks, Orlando, Florida, pp. 3540–3545.
Li, R.Z., 2006. Advanced and trend analysis of theoretical methodology for water quality forecast. J. Hefei Univ. Technol. 29, 26–30.
Lo, J.Y., Baker, J.A., Kornguth, P.J., Floyd, C.E., 1995. Application of artificial neural networks to the interpretation of mammograms on the basis of the radiologists impressions and optimized BI-RADS™ image features. Radiology 197 (P), 242–1242.
Maier, H.R., Dandy, G.C., 1998. The effect of internal parameters and geometry on the performance of back propagation neural networks: an empirical study. Environ. Model. Softw. 13, 193–209.
Maier, H.R., Dandy, G.C., 2000. Neural networks for the prediction and forecasting of water resources variables: a review of modeling issues and applications. Environ. Model. Softw. 15, 101–124.
Messikh, N., Samar, M.H., Messikh, L., 2007. Neural network analysis of liquid–liquid extraction of phenol from wastewater using TBP solvent. Desalination 208, 42–48.
NIST/SEMATECH e-Handbook of Statistical Methods, 2006. http://www.itl.nist.gov/div898/handbook.
Niu, Z.G., Zhang, H.W., Liu, H.B., 2006. Application of neural network to prediction of coastal water quality. J. Tianjin Polytechnic Univ. 25, 89–92.
Palani, S., Liong, S., Tkalich, P., 2008. An ANN application for water quality forecasting. Mar. Pollut. Bull. 56, 1586–1597.
Rahim, M.G., Goodyear, C.C., Kleijn, W.B., Schroeter, J., Sondhi, M.M., 1993. On the use of neural networks in articulatory speech synthesis. J. Acous. Soc. Am. 93, 1109–1121.
Raman, H., Chandramouli, V., 1996. Deriving a general operating policy for reservoirs using neural networks. J. Water Resour. Plann. Manage. 122, 342–347.
Rogers, L.L., Dowla, F.U., 1994. Optimization of groundwater remediation using artificial neural networks with parallel solute transport modeling. Water Resour. Res. 30, 457–481.
Sengorur, B., Dogan, E., Koklu, R., Samandar, A., 2006. Dissolved oxygen estimation using artificial neural network for water quality control. Fresen. Environ. Bull. 15, 1064–1067.
Schalkoff, R., 1992. Pattern Recognition: Statistical, Structural and Neural Approaches. Wiley, NY.
Shu, J., 2006. Using neural network model to predict water quality. North Environ. 31, 44–46.
Singh, K.P., Malik, A., Mohan, D., Sinha, S., 2004. Multivariate statistical techniques for the evaluation of spatial and temporal variations in water quality of Gomti River (India)—a case study. Water Res. 38, 3980–3992.
Singh, K.P., Malik, A., Basant, N., Saxena, P., 2007. Multi-way partial least squares modeling of water quality data. Anal. Chim. Acta 584, 385–396.
Smith, M., 1994. Neural Networks for Statistical Modelling. Van Nostrand Reinhold, NY, p. 235.
Smits, J.R.M., Breedveld, L.W., Derksen, M.W.J., Katerman, G., Balfoort, H.W., Snoek, J., Hofstraat, J.W., 1992. Pattern classification with artificial neural networks: classification of algae, based upon flow cytometer data. Anal. Chim. Acta 258, 11–25.
Soyupak, S., Karaer, F., Gurbuz, H., 2003. A neural network based approach for calculating dissolved oxygen profiles in reservoirs. Neural Comput. Appl. 12, 166–172.
Wang, Q.H., 2004. Improvement on BP algorithm in artificial neural network. J. Qinghai Univ. 22, 82–84.
Wen, C.W., Lee, C.S., 1998. A neural network approach to multiobjective optimization for water quality management in a river basin. Water Resour. Res. 34, 427–436.
Wu, H.J., Lin, Z.Y., Guo, S.L., 2000. The application of artificial neural networks in the resources and environment. Resour. Environ. Yangtze Basin 9, 237–241 (in Chinese).
Xiang, S.L., Liu, Z.M., Ma, L.P., 2006. Study of multivariate linear regression analysis model for ground water quality prediction. Guizhou Sci. 24, 60–62.
Ying, Z., Jun, N., Fuyi, C., Liang, G., 2007. Water quality forecast through application of BP neural network at Yuquio reservoir. J. Zhejiang Univ. Sci. A 8, 1482–1487.