Environment Protection Engineering

Vol. 36 2010 No. 2

HASAN MERDUN*, ÖZER ÇINAR**

ARTIFICIAL NEURAL NETWORK AND REGRESSION TECHNIQUES IN MODELLING SURFACE WATER QUALITY

Surface water quality variables with the common nature of complex, nonlinear, multivariable and high variability need the application of alternative techniques to define optimum input variables and to develop effective models. The present paper describes the systematic or hierarchical development and validation of artificial neural network (ANN) and multiple linear regression (MLR) models for the purpose of predicting chlorophyll *a* (Chl-*a*) concentration from nine surface water quality variables in order to investigate the significant input variables and their contribution order. A multilayer feed-forward network (FFN) trained by back-propagation algorithm was used as ANN approach. Both FFN and MLR techniques were first calibrated with the same three-fourth of the water quality data (304 samples) and then tested with the remaining one-fourth of the data. The performances of hierarchical models of both techniques were evaluated by using two statistical parameters such as coefficient of determination (R^2) and root mean square error (RMSE). The systematic or hierarchical analysis results showed that only four input variables such as biological oxygen demand (BOD), water temperature (T), dissolved oxygen (DO), and phosphorus (P) among nine variables explained 75 and 88% of the variability in Chl-*a* compared to the full models of FFN and MLR techniques, respectively. Using these four substantial input variables and adding their powers and possible interaction terms as inputs increased the prediction ability of MLR technique by 13%. The best model of MLR was better than the full model of FFN for predicting Chl-*a* based on R^2 of 0.69 and 0.61%, and RMSE of 14.99 and 16.35, respectively. The study results show that both FFN and MLR techniques are capable of simulating Chl-*a* with acceptable accuracy and suggest that the abilities of them should be further investigated in different environments and under different conditions with hierarchical model development.

1. INTRODUCTION

Chlorophyll *a* (Chl-*a*) concentration is an indicator of the abundance and variety of phytoplankton (microscopic plants or algae) or algal biomass is represented by Chl-*a*

* Faculty of Engineering, Department of Environmental Engineering, Akdeniz University, Antalya, Turkey. Phone: +90(242) 3106326. Fax: +90(242) 3106306. E-mail: merdun@alumni.clemson.edu

** Faculty of Engineering and Architecture, Department of Environmental Engineering, Kahramanmaras Sutcu Imam University, Kahramanmaras, 46100, Turkey.

in surface waters. Chl-*a* is a pigment which allows plants (including algae) to use the sunlight as energy and some nutrients (mainly nitrogen and phosphorus) as food to produce organic compounds through the process of photosynthesis. Chl-*a* concentration can be used as an effective measure of trophic status, hence, the quality of water resources [1]. The presence of an excess amount of these nutrients can stimulate algal blooms, hence, Chl-*a* concentration, resulting in reduced water clarity, reduced amount of good quality food, and depleted oxygen levels in deeper water. This process, called eutrophication, causes not only the degradation of water quality but also of aquatic life in ecosystem. Chl-*a* concentrations are often higher after rainfall due to the flow of rain water with nutrients such as nitrogen and phosphorus into the water. Slow moving or stagnant waters are more susceptible to eutrophication compared to running waters because they allow nutrients to increase and cell numbers to grow [2].

Water quality variables, especially Chl-*a* concentration, generally exhibit a complex pattern with time and in space. Modelling water quality has a considerable degree of complexity due to a great number of variables used, nonlinearity of the processes, and scarcity of data [3]. Many statistically based models such as multiple linear regression (MLR) are commonly used to solve many problems, but sometimes MLR has some shortcomings due to the assumption of linear principles even though water quality variables have nonlinear relationships [4]. Unlike many statistically-based water quality models, artificial neural networks (ANNs) are very powerful computational technique for modelling complex nonlinear relationships which are the nature of water quality variables [5]. One of the main advantages of ANNs over other modelling techniques, including the statistically-based models, is that they have the ability to model complex, nonlinear processes without using a priori knowledge of the relationships between input and output variables [6].

Recently, ANNs have been used in different ecological studies such as groundwater [7] and surface water ([4] and [8]) quality modelling and the determination of eutrophication levels of lakes or reservoirs ([9]–[12]). In these studies, the most commonly used ANN structure, a multi-layer feed-forward network (FFN) trained by back-propagation algorithm, has been applied alone to a variety of ecological modelling. However, the application of MLR to ecological or water quality problems is scarce and its comparison with FFN is absent or very limited, even if being available in the literature. Researchers might underestimate its prediction capacity for nonlinear water quality variables without investigating its capacity thoroughly. In addition, LEE et al. [11] observed noted that an optimal choice of input variables was not built based on ecological considerations and most of researchers used almost all possible environmental parameters as inputs. The use of all possible input variables in the model may cause a noise instead of being useful information because some of the input variables may have double effects. The systematic or hierarchical development of FNN and MLR models with various number of input variables is very limited [11]. Investigation of an optimal choice of input variables among several water quality variables using FFN and MLR techniques may offer the better understanding of the characteristics of the relationships between input and output variables and improve the predictive capacities of the models. In addition, such an investigation may prevent unnecessary parameter measurements if the addition of more variables does not affect the model performance.

Therefore, the objective of this study is to simulate Chl-*a* concentration by FFN and MLR techniques using nine water quality variables such as water temperature (T), dissolved oxygen (DO), biological oxygen demand (BOD), electrical conductivity (EC), turbidity (TU), organic nitrogen (ON), nitrite+nitrate nitrogen (N), ammonium nitrogen (NH_4), and phosphorus (P) as the input variables of the models. Systematic and hierarchical models of FFN and MLR are developed by trimming the most complicated network and using backward selection procedure in order to investigate the significant input variables and their contribution order. After the data (304 samples) was normalized as preprocessing, FFN was then trained and tested by using three-fourth and one-fourth of the data set, respectively. The same data sets of FFN for training and testing were used for derivation and validation of MLR models. Finally, the performances of FFN and MLR hierarchical models for prediction of Chl-*a* concentration were evaluated by using two statistical parameters, the coefficient of determination (R^2) and the root mean square error (RMSE).

2. MATERIALS AND METHODS

2.1. SITE CHARACTERISTICS AND WATER DATA

The Saginaw Bay watershed is located in the east-central lower part of Michigan (figure 1) and is the largest watershed in Michigan (approximately 22261 km^2), including 9 sub-watersheds and 22 counties. It has the largest contiguous freshwater coastal wetland system in the USA, extending along the shores of Saginaw Bay and providing habitat for millions of migrating waterfowls and songbirds, and over 90 fish species. Most of the watershed area drains into Saginaw Bay through the Saginaw River system. The watershed also includes 175 lakes, around 11265 km of rivers and streams, and drains approximately 15% of total land area of Michigan. Twenty-eight rivers, creeks, and agricultural drains flow directly into Saginaw Bay. Mean annual precipitation is between 68.58 and 78.74 cm. Mean annual runoff is between 20.066 and 40.132 cm, depending on precipitation, slope, land use and land cover [14].

The Saginaw Bay watershed is one of the ecologically most diverse areas in Michigan, supporting agriculture, industry, tourism, outdoor recreation, forestry, and a variety of wildlife. Land use is very diverse in the Saginaw Bay watershed and consists of: agriculture (46%), forest (29%), open lands (11%), urban (8%), wetlands

(4%), and water (2%). Agricultural activities include sugar beet, corn, dry bean, barley, wheat, and potato production. Extensive land use alterations have significantly changed the quantity, diversity, and quality of habitat for wildlife. Poultry, dairy and beef cattle are also raised in the watershed. Automobile manufacturing and suppliers are main industrial activity in the watershed [14].

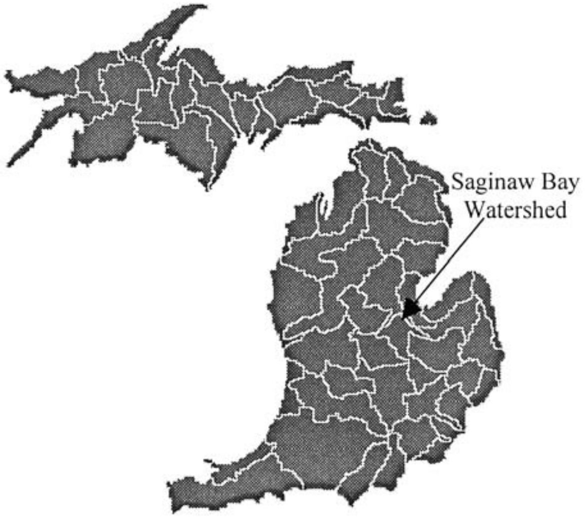

Fig. 1. The map of Saginaw Bay watershed in Michigan (adopted from: http://www.gis.iwr.msu.edu/storet/Form_Lev1.asp?Mode=Watershed&Water=Surface)

The main water quality problems lie in eutrophication (excessive nutrients which accelerate growth of aquatic plants and reduce oxygen levels), sedimentation, and toxic chemical contamination. Various point and nonpoint sources continue to contribute to the contamination of the Saginaw River and Saginaw Bay with industrial and municipal discharges, storm water overflows, contaminated sediments in the river and bay bottom, agricultural runoff, urban storm water runoff, landfill leaching, and atmospheric deposition [14].

Water quality in Saginaw Bay was described as mesotrophic to eutrophic based on the productivity level of monitoring data for nutrients from 1993 to 1999. The amounts of especially phosphorus and nitrogen are the limiting factors for the productivity level, hence, eutrophication level of a water body. Nearly two tons of total phosphorus per day were introduced into Saginaw Bay by the Saginaw River during the 1970s and 1980s, resulting in the growth of blue-green algae. In addition, an increased biological productivity in the bay caused an increase in the organic debris such as decomposing algae, aquatic plants, and small invertebrate animals. Benthic species in the bay are characterized as pollution-tolerant (bottom-dwelling worms and midges), a characteristic of a eutrophic system. Improved agricultural management practices to control fertilizer runoff in the watershed reduced the total phosphorus loads entering Saginaw Bay from 1700 MT/yr in 1973 to 665 MT/yr in 1982, leading to its concentration greater than 0.015 mg/dm^3 which exceeds targets (0.015 mg/dm^3) for the bay. Approximately 80–90% of phosphorus loads entering Saginaw Bay come from non-point sources (agricultural area). The concentration of Chl-*a* often exceeds 10 μg/dm^3, a threshold for eutrophic waters. The nitrogen/phosphorus ratios in the bay are approaching a level which no longer favours the production of blue-green algae [14].

Sediments carrying contaminants into the Saginaw River can cause significant water quality problems. Sediments in the part of the Saginaw River extending into Saginaw Bay contain a wide range of pollutants such as residue from oil and grease discharges, different forms of phosphorus and nitrogen, heavy metals (lead, zinc, nickel, arsenic, cadmium, chromium, copper, and mercury detected above acceptable levels), and organic chemicals (pesticide residues such as DDT derivatives, dieldrin, and chlordane and industrial compounds such as polychlorinated biphenyls (PCBs), polybrominated biphenyls (PBBs), phenolic compounds, and different chlorobenzenes). Therefore, knowing sediment yield is a critical factor in identifying the extent of water quality problems [14]. OUYANG and BARTHOLIC [13] estimated the sediment yield in the Saginaw Bay watershed through three different methods by using a sediment delivery ratio (SDR), expressed as the per cent of gross soil erosion by water that is delivered to a particular point in the drainage system. SDR ranged from 17.1% to 21.6% in the Saginaw Bay watershed.

These heavy metals and organic compounds attached to sediments are toxic at low concentrations and persist for a long time in the environment. Besides, there are a number of potential sources of bacteria from human waste, including sanitary sewer overflows and treatment plant discharges. Sewer discharges of communities in the Saginaw Bay watershed were approximately 27 million gallons in 1996, 12 million gallons in 1997, and only 800 000 gallons in 2001 [14].

The data used in this study was taken from the database http://www.gis.iwr.msu.edu/storet/default.htm for station of 090162 located in Bay County of Saginaw watershed in Michigan, USA. The database included several surface water quality variables measured at different times, starting from 1973 to 1994 for this station. When the database was screened for such water quality variables as *T*, DO, BOD, EC, TU, ON, N, NH_4, P, and Chl-*a*, which were input and output variables used in this study, a number of 304 samples resulted in, because some of the parameter measurements were missing for a given time.

2.2. ARTIFICIAL NEURAL NETWORKS

Artificial neural networks (ANNs) have a computing system similar to human nervous system where several neurons are connected and communicate with each other to arrive at a decision. The most important advantage of ANNs over other modelling techniques is their ability to model complex and nonlinear processes without any a priori knowledge of input and output relationships. A multi-layer feed-forward network (FFN) trained by back-propagation algorithm is ANN type often used in the modelling of processes in a variety of fields, including ecology, water quality, and hydrology. Figure 2 illustrates the architecture of a typical three-layered FFN (input, hidden, and output layers). FFN is briefly introduced as the following and extensive information on it and can be obtained from [15] which includes fundamental concepts and mathematical expressions. In developing the model for the estimation of Chl-*a* (output parameter) from input parameters (*T*, DO, BOD, EC, TU, ON, N, NH_4, P), a relationship between input and output parameters is developed by a training process, which is explained in the next paragraph. The information flows from the input layer towards the output layer through the hidden layer where the input layer receives information from the input data, processes it in hidden layer(s), and produces the output. A total effect, i.e. the sum of all effects of all the inputs on a given node, is determined by adding the product of each input and the corresponding weight, and then the total effect is transformed to the output using an activation function, and finally the output is transferred to each neuron of the next layer through a connection weight. A nonlinear transfer function as an activation function is used to process the sum and to generate the results. The S-shaped sigmoid function is one of the transfer functions most commonly used.

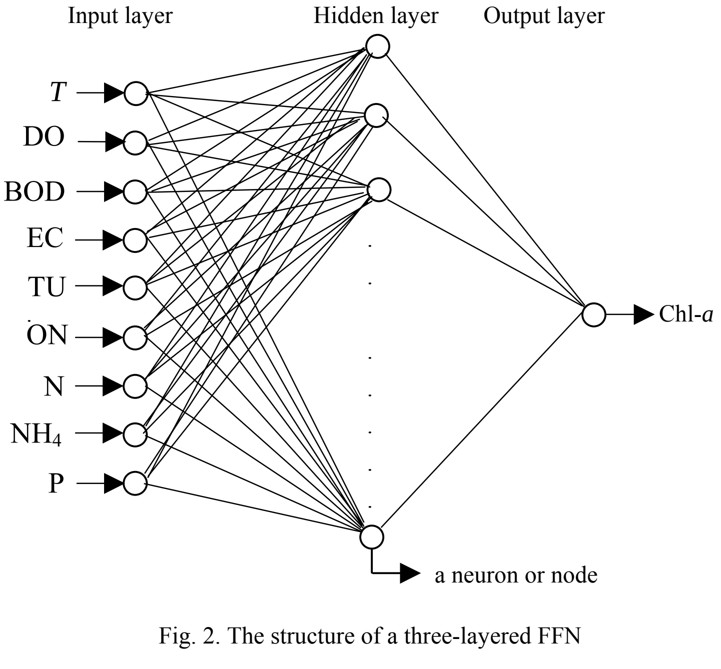

Fig. 2. The structure of a three-layered FFN

An FFN must be trained before testing its ability to predict the response variable. The aim of learning is to determine the set of weights that minimize the error function. A back-propagation method is used in the training of FFN to adjust the connection weights and to obtain the best match between the FFN's response and the expected response. In a training procedure, firstly, the weights are initialized with a set of random values. Then the weights are systematically updated based on the training rule. After several attempts, the training process is terminated when the difference between the measured and predicted values is less than the specified criteria. One of the common problems arising from the training process is the over-training or over-fitting which might produce incorrect results. The over-training of FFN depends on the use of the optimal number of nodes in the hidden layer. If the number of hidden nodes are too small, then the network may have insufficient degrees of freedom to learn the process adequately. If the number is too high, then the training takes a long time and the network may over-fit the data. The number of nodes in the hidden layers are determined by a trial-and-error procedure in this study.

In the application of FFN technique, after the complete data set (304 samples) was normalized as preprocessing, three-fourth of the data (228 samples) was used to train FFN. The trained networks were then tested using one-fourth of the data (76 samples). Systematic models were essentially developed with a network trimming process, starting with the most complicated network, removing one parameter at a time. A total of nine networks were compared in order to predict Chl-*a* using testing data set. All simulations were carried out by the neural network toolbox of the MATLAB$^{\text{TM}}$ package (version 6.5, MathWorks, Inc., USA).

2.3. MULTIPLE LINEAR REGRESSIONS

Multiple regression is used to predict a dependent variable from several independent variables. A set of independent variables explains the proportion of the variance in a dependent variable at a significant level. Hierarchical regression allows someone to see how variance in the dependent variable can be explained by adding or removing one or more independent variables. In MLR model development, it is assumed that the variables have normal distribution, the relationship between dependent and independent variables is linear, and the values of variables are measured reliably or accurately. The significance of the difference between two cases (before and after removing an independent variable from the model) can be tested by the coefficient of determination (R^2) and the root mean square error (RMSE). Power terms and interaction or cross-product terms can be added as independent variables to explore curvilinear and interaction effects, respectively. Hierarchical multiple regression is similar to stepwise regression, but the user (not the computer) determines the order of entry/removal of the variables to/from the model. *F*-test is used to compute the significance of each removed variable (or set of variables) to the explanation reflected in R^2. In this study, firstly, all possible regression models were developed using all independent variables. Secondly, the four most important input variables were selected and then their main effects, powers, and possible interaction terms were included in the model using the Statistical Analysis System software (SAS Institute Inc., Cary, NC). In this study, the independent variables were chosen based on the literature ([10]–[12]) and the availability of the data in the database. The general form of multiple linear regression can be expressed as:

$$Y = b_0 + b_1 X_1 + \ldots + b_4 X_4 + b_5 X_1^2 + \ldots + b_8 X_4^2 + b_9 X_1 X_2 + \ldots + b_{14} X_3 X_4 , \tag{1}$$

where Y is the dependent variable representing Chl-*a*, b_0 is the intercept, $b_1, \ldots, b_n$ refer to the regression coefficients, and X_1 to X_4 are independent variables referring to the four important water quality variables in the model. The same FFN training and testing data set was used for the derivation and validation of MLR models.

2.4. PERFORMANCE EVALUATION

The performances of FFN and MLR methods in predicting chlorophyll *a* concentration from different water quality parameters were evaluated using two statistical parameters: the coefficient of determination (R^2) and the root mean square error (*RMSE*). Each of these criteria is based on the differences between the measured and predicted values and defined as follows:

$$R = \frac{\sum (M_i - \overline{M})(P_i - \overline{P})}{\sqrt{\sum (M_i - \overline{M})^2 \sum (P_i - \overline{P})^2}} , \tag{2}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (M_i - P_i)^2}{n}} \tag{3}$$

where R is the correlation coefficient and its square provides the coefficient of determination (R^2); n is the total number of the set of measurements; M_i and P_i are respectively the measured and predicted data, i = 1, 2, 3, ..., n; $\overline{M}$ is the average of the measured values, and $\overline{P}$ is the average of the predicted values. 1 and 0 refer to the ideal values of R^2 and *RMSE*, respectively, corresponding to a perfect matching between the measured and predicted data. In addition, the analysis of variance (ANOVA) test was performed using SAS (SAS Institute Inc., Cary, NC) to determine whether there were differences between the means of Chl-*a* obtained by using FFN and MLR techniques.

3. RESULTS AND DISCUSSION

The descriptive statistics of surface water quality data used in the training or derivation of FFN and MLR models are given in table 1. They testify to the large variations in Chl-*a* observed between the samples with a coefficient of variation (CV = standard deviation/mean) significantly exceeding 100% (163.11%). The relatively large variations in TU (91.88%), NH_4 (90.24%), *T* (76.04%), and N (71.43%) correspond to the large variations in Chl-*a*. Relatively similar trends are observed in the testing or validation data set. The large variations in these water quality variables may be because of the substantial geographical differences in climate, water resources, ecology, soil characteristics, and land use in the Saginaw Bay watershed. In addition, the variety of activities such as agriculture, industry, tourism, outdoor recreation, and forestry in the watershed may cause such a large variation in the variables.

Table 1

Descriptive statistics of water quality parameters used in the calibration and validation of FFN and MLR

Water quality variables	Training or derivation of data set				Testing or validation of data set			
	Min.	Max.	Mean	SD*	Min.	Max.	Mean	SD
Temperature (°C)	0.00	27.50	12.02	9.14	0.00	26.00	12.16	9.57
Dissolved oxygen (mg/dm^3)	2.70	16.80	8.70	2.66	3.50	13.10	8.41	2.44
Biological oxygen demand (mg/dm^3)	1.00	16.00	4.04	1.90	1.40	8.50	4.00	1.63
Electrical conductivity (mmhos)	240.00	1140.00	690.31	138.76	318.00	1016.00	724.25	117.97
Turbidity (Ftu)	1.80	160.00	19.95	18.33	2.80	170.00	17.27	19.84
Organic nitrogen (mg/dm^3)	0.52	2.80	1.13	0.34	0.50	2.00	1.15	0.33
Nitrogen (NO_2+NO_3) (mg/dm^3)	0.14	5.20	1.40	1.00	0.28	3.70	1.19	0.78
Ammonium (NH_4) (mg/dm^3)	0.01	2.80	0.41	0.37	0.06	2.30	0.48	0.42
Phosphorus (mg/dm^3)	0.04	0.57	0.18	0.09	0.06	0.48	0.17	0.08
Chlrophyll *a* (μg/dm^3)	0.20	520.10	25.13	40.99	0.20	94.00	25.63	25.77

*SD: standard deviation.

The FFN parameters are the number of input, output, and hidden nodes and iteration or epoch number. The number of input and output nodes were 9 and 1, referring to the input and output variables, respectively. Several iterations were made until the optimum number of hidden layers and numbers of neurons in each hidden layer were determined as 3 hidden layers and 10 neurons in each hidden layer. Therefore, the architecture of FFN was determined by a trial-and-error approach. The over-training or over-fitting problem was controlled by using the optimum number of hidden layers and neurons, and epoch number of 1000 determined by trial-and-error procedure to minimize the MSE values.

Table 2

Different neural networks tested for the prediction of Chl-*a*

Model	Input variables*	R^2	*RMSE*
9	BOD	0.13	32.22
8	BOD, *T*	0.34	28.23
7	BOD, *T*, DO	0.41	26.73
6	BOD, *T*, DO, P	0.46	25.34
5	BOD, *T*, DO, P, TU	0.34	27.98
4	BOD, *T*, DO, P, TU, ON	0.37	27.62
3	BOD, *T*, DO, P, TU, ON, NH_4	0.34	28.33
2	BOD, *T*, DO, P, TU, ON, NH_4, EC	0.40	26.95
1	BOD, *T*, DO, P, TU, ON, NH_4, EC, N	0.61	16.35

* BOD: biological oxygen demand (mg/dm^3), *T*: water temperature (°C), DO: dissolved oxygen concentration (mg/dm^3), P: phosphate (mg/dm^3), TU: turbidity (Ftu), ON: organic nitrogen (mg/dm^3), NH_4: ammonium nitrogen (mg/dm^3), electrical conductivity (mmhos), N: nitrite+nitrate nitrogen (mg/dm^3), R^2: coefficient of determination, and *RMSE*: root mean square error.

In the systematic analysis of FFN performance, each of nine FFN models was evaluated by R^2 and *RMSE* and the results are displayed in table 2. It can be seen from the table that the models from 2 to 5 show similar performance with approximately similar R^2 (0.40–0.34) and *RMSE* (26.95–27.98 μg/dm^3) during testing. However, relatively significant differences between the models from 6 to 9 were observed with R^2 of 0.46–0.13 and *RMSE* of 25.34–32.22. The model 1 (the full model) was the best in the prediction of Chl-*a* with R^2 of 0.61 and *RMSE* of 16.35, but the improvement was only 15%, referring to the model 6 (R^2 from 0.46 to 0.61). The results show that using more environmental variables as the network inputs does not give much advantage. This improvement in the models can be sacrificed to time, labour, and expenses in the measurements of additional five variables if this much accuracy is considered to be unimportant. In the prediction of Chl-*a* from ten input variables such as solar radiation, total inorganic nitrogen, time-lagged Chl-*a*, phosphorus, DO, secchi-disc depth, *T*, rainfall, wind speed, and tidal range with ten different tested systematic neural networks, LEE et al. [11] defined the network that consisted of time-lagged Chl-*a* only as inputs as the optimal network. The testing accuracies of their ten networks were in the ranges of R^2 = 0.83–0.95 and *RMSE* = 3.045–5.653 μg/dm^3. YABUNAKA et al. [9], KARUL et al. [10], and KUO et al. [12] simulated Chl-*a* using similar water quality variables as inputs in lakes or reservoirs YABUNAKA et al. [9] and KUO et al. [12] found that the R^2 values for the prediction of Chl-*a* were 0.64 and 0.74, respectively. The R^2 values between the measured and calculated Chl-*a* values were between 0.36 and 0.56 in Keban Dam Reservoir, whereas R^2 was as high as 0.90 in Mogan and Eymir Lakes in the study of KARUL et al. [10]. Some of these studies have somewhat better simulation abilities than this study, but this may be normal because these studies were conducted in a relatively much smaller geographic area compared to this study which covers large and ecologically diverse the Saginaw Bay watershed. A variety of point and nonpoint contaminant sources resulted from different activities such as agricultural runoff, industrial and municipal discharges, sewer overflows, and contaminated sediments may be responsible for large variability in water quality variables, thereby inversely affecting the performances of the models.

Table 3

Coefficients of variables used in hierarchical MLR models for prediction of Chl-*a*

Model	Input variables*	Coefficients	R^2	*RMSE*
9	BOD	–19.46257, 11.03946	0.23	23.14
8	BOD, *T*	–27.04492, 9.15228, 1.26546	0.46	19.81
7	BOD, *T*, DO	–91.45760, 8.37836, 2.55198, 5.98324	0.45	21.09
6	BOD, *T*, DO, P	–140.66649, 5.13716, 3.11359, 9.09703, 162.32572	0.49	21.16
5	BOD, *T*, DO, P, TU	–142.37270, 4.76658, 3.22499, 9.62101, 9.62101, –0.31724	0.34	40.12
4	BOD, *T*, DO, P, TU, ON	–145.22126, 3.91909, 2.96398, 9.17404, 172.80993, –0.33955, **13.72112**	0.53	19.39
3	BOD, *T*, DO, P, TU, ON, NH_4	–120.40021, 4.31718, 2.31407, 7.48578, 196.59744, –0.46263, **14.89556**, **–16.86663**	0.55	19.06
2	BOD, *T*, DO, P, TU, ON, NH_4, EC	–135.65783, 4.12085, 2.19854, 7.35297, 200.02004, –0.37001, **13.86446**, –22.25491**,** **0.02831**	0.55	19.11
1	BOD, *T*, DO, P, TU, ON, NH_4, EC, N	–127.91475, 3.96400, 2.04117, 7.25249, 194.28967, –0.35338, **15.16170**, –23.85886, **0.02715**, **–2.62499**	0.56	18.46
10	BOD, *T*, DO, P, BOD^2, T^2, DO^2, P^2, BOD.*T*, BOD.DO, BOD.P, *T*.DO, *T*.P, DO.P	138.2, –22.25, **–1.597**, **–14.35**, **–279.4**, **0.0290**, **–0.029**, **0.345**, **60.65**, 0.284, 1.550, 37.27, **0.189**, 7.700, **9.353**	0.69	14.99

*BOD: biological oxygen demand (mg/dm^3), *T*: water temperature (°C), DO: dissolved oxygen (mg/dm^3), P: phosphate (mg/dm^3), TU: turbidity (Ftu), ON: organic nitrogen (mg/dm^3), NH_4: ammonium nitrogen (mg/dm^3), electrical conductivity (mmhos), N: nitrite+nitrate nitrogen (mg/dm^3), BOD^2: BOD.BOD, T^2: *T*.*T*, DO^2: DO.DO, P^2: P.P, R^2: coefficient of determination, and *RMSE*: root mean square error.

Note: The first values of the coefficients of each variable represent the intercept and the others correspond to the variables in respective order. Bold coefficients are statistically not significant at 5% level of probability ($p > 0.05$). All of the ten models are statistically significant at the probability level of $p < 0.00001$.

Hierarchical MLR models developed by the backward selection method to investigate the contribution of each water quality variable to the prediction of Chl-*a* using validation data set are presented in table 3. As can be seen from the table, during validation the models from 1 to 6 show similar performance with approximately similar R^2 (0.56–0.49) and *RMSE* (18.46–21.16 μg/dm^3) except the model 5. However, relatively significant differences between the models from 6 to 9 were observed with R^2 of 0.49–0.23 and *RMSE* of 21.16–23.14. The model 5 had some instability, maybe because of collinearity, which is a near perfect linear relationship between some or all of the independent variables in the regression model. This means that there is some degree of redundancy or overlap among the independent variables. This causes a difficulty in distinguishing the individual influences of the independent variables on the response variable (Chl-*a*). The model 9 relating Chl-*a* to BOD alone explained 23% of the variability (R^2 = 0.23). The addition of the second variable (*T*) significantly improved (23%) the prediction capacity of the model 8. Similar to FFN, the inclusions of the first four variables to MLR models significantly improved their performances for prediction of Chl-*a*, where R^2 increased from 0.23 to 0.49 and the corresponding *RMSE* decreased from 23.14 to 21.16. However, further additions of the remaining variables to the models improved their performances in total by only 7 (= 0.56–0.49)%. Since the further additions of input variables to the models had no significant effect on the prediction capacity of the models, these four essential input variables were selected and then their main effects, powers, and possible interaction terms were included to the model (model 10). The ability of this model with only four variables proved to be by 13% better than that of the full model (model 1). This indicated that the interaction and power terms in the model increased its prediction capacity. However, the contribution of the some of the variables and terms (bold coefficients in table 3) to the model were not statistically significant. This suggests that – as in FFN – the simpler the model, the better the result.

The time-series of measured and predicted values of Chl-*a* for training and testing the full FFN model and for the derivation and validation of the best MLR model are presented in figure 3. The time-series in figure 3 cover the time range between January 1, 1974 and January 22, 1992. As can be seen, FFN was very successful in the training with an R^2 of 0.89. The results of the testing procedures (R^2 = 0.61) were also satisfactory but not as good as those of training. The results of the derivation procedures of MLR were also very successful with an R^2 of 0.81, but the derivation procedures (R^2 = 0.69) were not so successful as opposed to the derivation. In general, both FFN and MLR techniques simulated acceptably Chl-*a* concentrations with time, even though Chl-*a* had large variations with time. The predicted results of the best models of both techniques are exhibited in figure 4. Less scattering of data points around the line passing through the origin and the right corner of the plots indicates that MLR is somewhat better than FFN in predicting Chl-*a*. Although MLR was slightly better than FFN in the prediction of Chl-*a*, ANOVA test results showed that this difference was not statistically significant ($p > 0.05$). Therefore, the results of this study indicated that both FFN and MLR techniques can be applied in the modelling of surface water quality, even though these variables are complex, nonlinear, and show large variations with time.

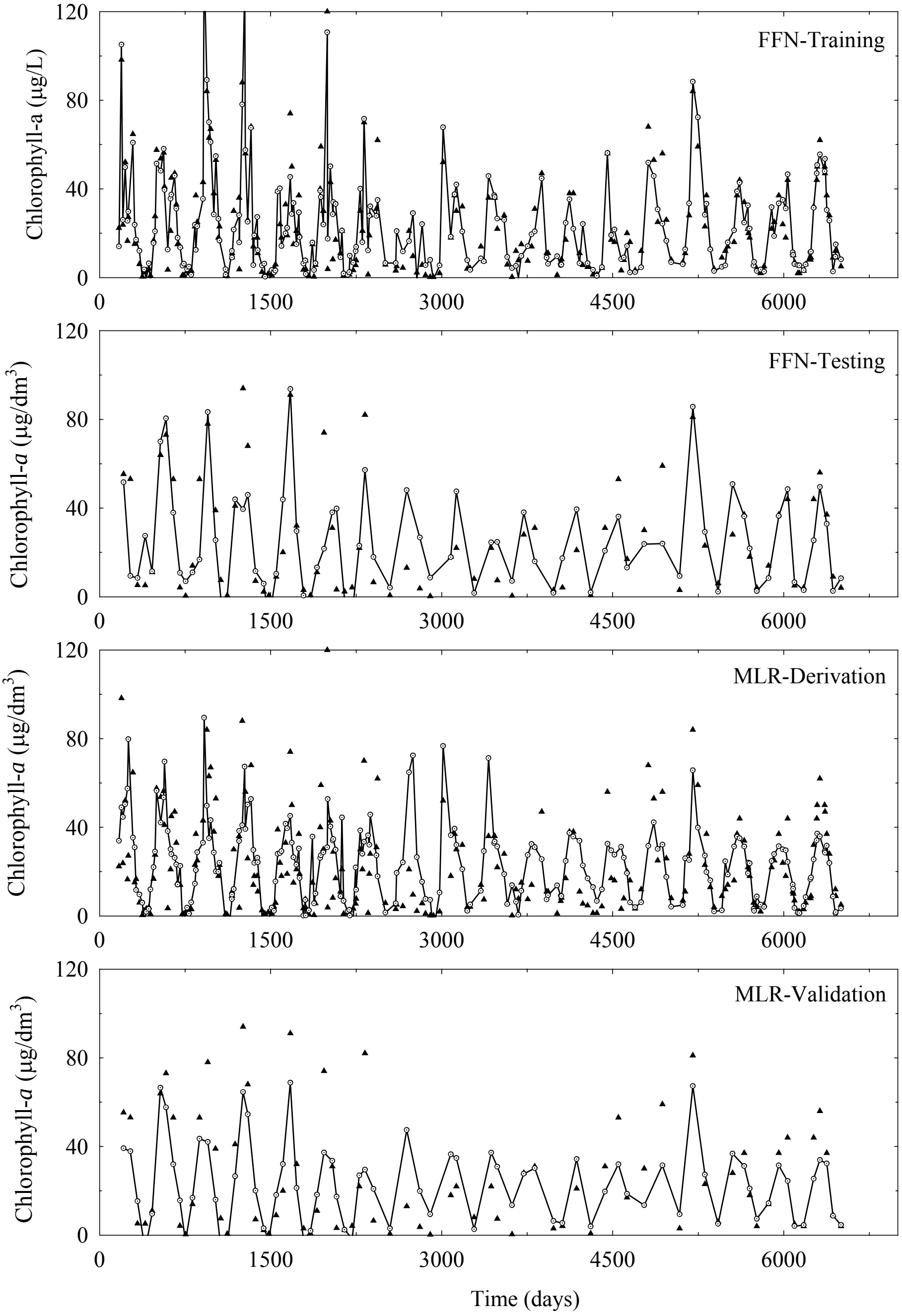

Fig. 3. Time-series of measured (triangle) and predicted (dotted circle line) values of Chl-*a* for the training or derivation (left) and testing or validation (right) of the best FFN and MLR models

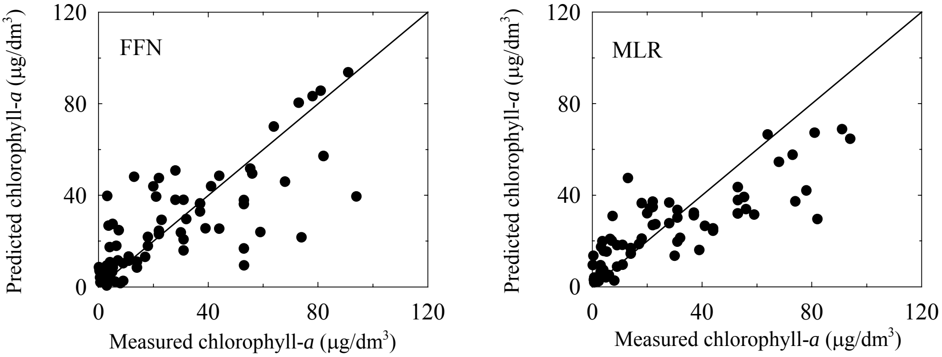

Fig. 4. Measured versus predicted values of Chl-*a* for the best FFN and MLR models

4. CONCLUSION

In this study, Chl-*a* concentration was simulated by FFN and MLR techniques using nine water quality variables as the input to the models. Systematic or hierarchical models of FFN and MLR were developed by trimming the most complicated network and using backward selection procedure in order to investigate the significant input variables and their contribution order. Of nine input variables, four explained a significant variability in Chl-*a* based on systematic or hierarchical model analyses of both techniques. The best model of MLR was somewhat better than the full FFN model in the prediction of Chl-*a*, but this difference was not statistically significant.

In this study, the selection of effective input variables and then further analysis may make MLR technique superior to FFN. This suggests that even though water quality variables are complex and non-linearly related, statistically based models such as MLR which assumes linear relationship between input and output variables [5] should be investigated further for different environmental conditions. In addition, as concluded by LEE et al. [11], the effect of sampling intervals (daily, weekly, or monthly) on the performances of both techniques for real-time predictions of Chl-*a* need to be further evaluated.

REFERENCES

[1] HARDING L.W., PERRY E.S., *Long-term increase of phytoplankton biomass in Chesapeake Bay, 1950–1994*, Marine Ecology Progress Series, 1997, 157, 39–52.

[2] HARDING L.W., *Long-term trends in the distribution of phytoplankton in Chesapeake Bay: Roles of light, nutrients and streamflow*, Marine Ecology Progress Series, 1994, 104, 267–291.

[3] CHAVES P., TSUKATANI T., KOJIRI T., *Operation of storage reservoir for water quality by using optimization and artificial intelligence techniques*, Mathematics and Computers in Simulation, 2004, 67, 419–432.

[4] LEK S., GUIRESSE M., GIRAUDEL J.L., *Predicting stream nitrogen concentration from watershed features using neural networks*, Water Research, 1999, 33, 3469–3478.

[5] LEK S., DELACOSTE M., BARAN P., DIMOPOULOS I., LAUGA J., AULANIER S., *Application of neural networks to modeling nonlinear relationships in ecology*, Ecological Modeling, 1996, 90, 39–52.

[6] CHAU K.W., *A review on integration of artificial intelligence into water quality modeling*, Marine Pollution Bulletin, 2006, 52, 726–733.

[7] SAHOO G.B., RAY C., MEHNERT E., KEEFER D.A., *Application of artificial neural networks to assess pesticide contamination in shallow groundwater*, Science of the Total Environment, 2006, 367, 234–251.

[8] HOLMBERG M., FORSIUS M., STARR M., HUTTUNEN M., *An application of artificial neural networks to carbon, nitrogen and phosphorus concentrations in three boreal streams and impacts of climate change*, Ecological Modeling, 2006, 195, 51–60.

[9] YABUNAKA K., HOSOMI M., MURAKAMI A., *Novel application of a back-propagation artificial neural network model formulated to predict algal bloom*, Water Science and Technology, 1997, 36, 89–97.

[10] KARUL C., SOYUPAK S., ÇILESIZ A.F., AKBAY N., GERMEN E., *Case studies on the use of neural networks in eutraphication modeling*, Ecological Modeling, 2000, 134, 145–152.

[11] LEE J.H.W., HUANG Y., DICKMAN M., JAYAWARDENA A.W., *Neural network modeling of coastal algal blooms*, Ecological Modeling, 2003, 159, 179–201.

[12] KUO J.T., HSIEH M.H., LUNG W.S., SHE N., *Using artificial neural network for reservoir eutrophication prediction*, Ecological Modeling, 2007, 200, 171–177.

[13] OUYANG D., BARTHOLIC J., *Predicting sediment delivery ratio in Saginaw Bay Watershed*, the 22nd National Association of Environmental Professionals Conference Proceedings, May 19–23, Orlando, FL, 659–671.

[14] COSCARELLI M., *Targeting environmental restoration in the Saginaw River/Bay area of concern (AOC)*, 2002, prepared for the Great Lakes Commission on behalf of the Partnership for the Saginaw Bay Watershed.

[15] HAYKIN S., *Neural Networks: A Comprehensive Foundation*, second edition, Upper Saddle River, New Jersey, 1999.