ISEIS

Journal of Environmental Informatics 22(2) 92-101 (2013)

Journal of Environmental Informatics

www.iseis.org/jei

Modeling of Dissolved Oxygen in River Water Using Artificial Intelligence Techniques

O. Kisi[1], N. Akbari[2], M. Sanatipour[2], A. Hashemi[3], K. Teimourzadeh[4], and J. Shiri [5,*]

[1]*Department of Civil Engineering, Architecture and Engineering Faculty, Canik Basari University, Samsun, Turkey*
[2]*Department of Civil Engineering, Faculty of Engineering, University of Tabriz, Tabriz, Iran*
[3]*Water engineering Department, Shahid Abbaspour University, Tehran, Iran*
[4]*Sama Technical and Vocational Training College, Islamic Azad University, Tabriz Branch, Tabriz, Iran*
[5]*Water Engineering Department, Faculty of Agriculture, University of Tabriz, Tabriz, Iran*

Received 14 December 2012; revised 4 April 2013; accepted 15 October 2013; published online 20 December 2013

ABSTRACT. The accuracy of artificial neural networks (ANNs), adaptive neuro-fuzzy inference system (ANFIS) and gene expression programming (GEP) in modeling dissolved oxygen (DO) concentration was investigated in this study. Water temperature, specific conductance, pH, discharge and DO concentration data from South Platte River at Englewood, Colorado were used. Various input combinations of these data were tried as inputs to the ANN and ANFIS methods. The ANN and ANFIS models with the water temperature, specific conductance, pH and discharge input parameters performed the best. The optimal GEP model was obtained for the best input combination and compared with the ANN and ANFIS models with respect to correlation coefficient, root mean square error, mean absolute error and mean absolute relative error criteria. Results revealed that the GEP model performed better than the ANN and ANFIS models in modeling DO concentration.

Keywords: dissolved oxygen, modeling, neural networks, neuro-fuzzy, gene expression programming

1. Introduction

The concentration of dissolved oxygen (DO) is of importance for the health functioning and indicating the state of the aquatic ecosystems and its modeling is crucial for river water quality and wetlands ponds analysis. DO level is the measure of the health of the aquatic system. DO concentration is frequently used to evaluate the water quality in different reservoirs and watersheds (Schmid and Koskiaho 2006; Singh et al., 2009; Rankovic et al., 2010; Ay and Kisi, 2012).

In the recent past, the use of Artificial Intelligences (AI) techniques, e.g. Artificial Neural Networks (ANNs), Adaptive Neuro-Fuzzy Inference System (ANFIS) and Genetic Programming (GP) in water resources engineering has become viable. Notable works have been reported in literature regarding the application of ANNs in modeling rainfall-runoff and other hydrologic factors (ASCE, 2000). The complete review of such applications is beyond the scope of the present paper and only some most relevant papers will be discussed here. Maier and Dany (1996) investigated the ANN capabilities in modeling river water salinity. Schmid and Koskiaho (2006) applied ANNs for modeling near-bottom concentrations of dissolved oxygen in a wetland. Singh et al. (2009) used ANNs for modeling dissolved oxygen and biochemical oxygen demand in Gomti River, India. Faruk (2010) applied a hybrid Auto Regressive Integrated Moving Average (ARIMA) - ANN model for predicting time series of water quality data. Rankovic et al. (2010) applied ANNs for modeling dissolved oxygen in the Gruza Reservoir, Serbia. Ay and Kisi (2012) modeled daily DO concentration in Foundation Creek, El Paso County, Colorado by using multilayer perceptron and radial basis neural network methods. Palani et al. (2008) applied ANN models for the estimation of water variables such as temperature, salinity, DO and Chl-a data from East Johor Strait, Malaysia.

ANFIS is a combination of an adaptive neural network and a fuzzy inference system. The parameters of the fuzzy inference system are determined by the NN learning algorithms. Since this system is based on the fuzzy inference system, an important aspect is that the system should be always interpretable in terms of fuzzy IF-THEN rules. ANFIS is capable of approximating any real continuous function on a compact set to any degree of accuracy (Jang et al., 1997). ANFIS identifies a set of parameters through a hybrid learning rule combining back propagation gradient descent error digestion and a least squared error method. There are two approaches for fuzzy inference systems, namely the approach of Mamdani (Mamdani and Assilian, 1975) and approach of Sugeno (Takagi and Sugeno, 1985). The neuro-fuzzy model used in this study implements the Sugeno's fuzzy approach to obtain the values for the output variable (e.g. dissolved oxygen) from those of input variables

* Corresponding author. Tel.: +0098 4113340081.
E-mail address: j_shiri2005@yahoo.com (J. Shiri).

ISSN: 1726-2135 print/1684-8799 online
 doi: 10.3808/jei.201300248

(e.g. water temperature, specific conductance, pH and discharge parameters).

Chang and Chen (2001) applied counter propagation fuzzy-neural network modeling approach to real-time streamflow prediction. Kisi (2006) investigated the ability of ANFIS technique to improve the accuracy of daily evaporation estimation. Bae et al. (2007) applied weather forecasting information and neuro-fuzzy technique for predicting monthly dam inflow. Kisi et al. (2008) investigated the accuracy of ANFIS and ANN techniques in modeling daily suspended sediment of rivers in Turkey. Ozger and Yildirim (2009) used ANFIS to determine turbulent flow friction coefficient. Shiri and Kisi (2010) introduced a new wavelet-ANFIS model for predicting short term and long term streamflow values. Shiri et al. (2011a) used ANFIS for predicting short term operational water levels. Shiri et al. (2011b) compared ANFIS to ANNs in estimating daily pan evaporation values in local and regional (cross station) scales and found ANFIS better than ANNs. Azamathulla and Ghani (2011) used ANFIS for predicting scour depth at culvert outlets and they found ANFIS to be more effective when compared with the results of regression equations and ANN. Kisi and Shiri (2012a) introduced a wavelet-neuro-fuzzy model for predicting short term groundwater table depth fluctuations.

Genetic Programming (GP), firstly proposed by Koza (1992) as a generalization of Genetic Algorithm (GA) (Goldberg, 1989), employs a "parse tree" structure for the search of solutions. This technique has the capability for deriving a set of explicit formulations that rule the phenomenon, to describe the relationship between the independent and dependent variables using various operators. Harris et al. (2003) used GP to predict velocity in compound channels with vegetated flood plains. Giustolisi (2004) determined the Chezy resistance coefficient using GP. Shiri and Kisi (2011a) compared the GP to ANFIS for predicting short-term groundwater table depth fluctuations. Shiri and Kisi (2011b) applied various AI model for estimating daily pan evaporation values from available and estimated climatic data. Kisi and Shiri (2011) introduced a new wavelet-GEP conjunction model for precipitation forecasting. Shiri et al. (2012) applied GEP for modeling daily evapotranspiration. Kisi and Shiri (2012b) compared GEP to ANFIS and ANNs in modeling river suspended sediment with climatic variables implication and found GEP better than ANFIS and ANNs. Karimi et al. (2012) compared GEP to ANFIS in forecasting daily lake level fluctuations and found GEP as the best model in this field. Kisi et al. (2013) used GEP, ANFIS and ANNs for modeling rainfall-runoff process. Based on their results, GEP surpasses both ANFIS and ANN models in this regard. To the best knowledge of the authors, there is not any published study indicating the input–output mapping capability of GEP technique in modeling of dissolved oxygen concentration.

The main purpose of this study is to investigate the ability of GEP, ANFIS and ANN techniques in modeling DO concentration using mean water temperature, specific conductance, pH and discharge inputs. GP technique is firstly used for DO modeling in this study.

2. Materials and Methods

2.1. Used Data

Daily mean water temperature, specific conductance, pH, discharge and DO concentration data from South Platte River at Englewood, Colorado (USGS Station No: 06711565, latitude 39°39'54" NW, longitude 105°00'13" NE, height 1,600 m above mean sea level) operated United States Geological Survey (USGS), were used in this study. The drainage area at this site is 8,769 km^2. The station is located in Arapahoe County, on right bank, 483 m downstream from Dartmouth Ave Bridge at Englewood, and 2,253 m downstream from Bear Creek. Natural flow of stream affected by transmountain diversions, storage and flood control reservoirs, power developments, diversions for irrigation and municipal use, and return flow from irrigated areas. Flow regulated by Chatfield Dam since May 29, 1975 (station 06709600), and Bear Creek Dam since July 1979 (see http://waterdata.usgs.gov/nwisweb/local/nwis_host/nwisdcolka/local/site_text/site_info/txt06711565.htm).

The data consisted of fourteen years (1996 ~ 2012) of daily records of water temperature (Tmean), specific conductance (SC), pH, discharge (Q) and DO. The number of data is 4,020 because they have lots of missing values. Missing data were removed from the data set and the first 2,010 values (50% of the whole data set) were used to train ANN, ANFIS and GEP models. And, the remaining 1,005 data (25% of the whole data set) were used for testing and 1,005 data were used for validation of the applied models. Table 1 shows the statistical parameters of the daily data. In Table 1, the X_{max}, X_{min}, X_{mean}, SD, C_v, and C_{sx} denote the maximum, minimum, mean, standard deviation, coefficient of variation, and skewness coefficient, respectively. It can be seen from the Table 1 that the *pH* and *Q* data show negative and positive high skewed distribution, respectively.

Table 1. Statistical Parameters of the Applied Data during the Study Period

Parameter	Unit	X_{max}	X_{min}	X_{mean}	SD	C_V	C_{SX}
Tmean	^oC	24.8	0	11.7	6.9	0.59	0.07
SC	-	1940	0	766	298.3	0.38	0.46
pH	-	9	3.9	8.1	0.25	0.03	-3.26
Q	ft^3/s	1970	2	169.4	217.7	1.28	3.17
DO	mg/l	13	2.35	9.4	1.7	0.18	-0.16

2.2. Artificial Neural Networks (ANNs)

The neural network usually has two or more layers of neurons in order to process non-linear signals. Figure 1 illustrates a three layered ANN structure which is composed of layers *i*, *j*, and *k*, with the interconnection weights (W_{ij} and W_{jk}) between input, hidden and output layers. During the learning process, initial assigned weights are progressively corrected in which the predicted outputs are compared with the observed outputs, and the errors are backpropagated (from right to left in Figure 1) to obtain the appropriate weight adjustments necessary to minimize the errors. The input layer admits the incoming information, which is processed by the hidden layer(s), and the out-

put layer presents the network result. In the present study, three-layer feed-forward networks were employed with a sigmoid transfer function in the hidden layer and a linear transfer function in the output layer. As there is not yet a definite theoretical background for determining the interconnections of neurons, the hidden-layer-node numbers of each model were determined through a trial and error process. Further details about ANNs can be found in e.g. Haykin (1999).

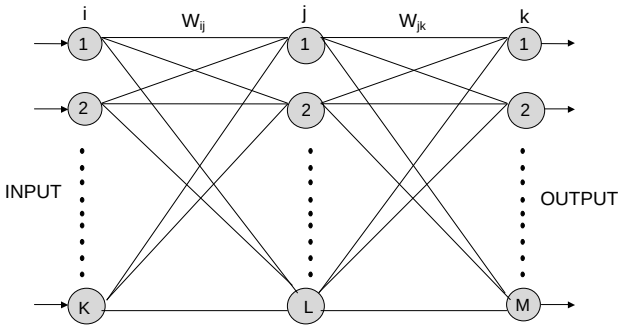

Figure 1. A three-layered ANN structure.

2.3. Adaptive Neuro-Fuzzy Inference System (ANFIS)

As a simple example a fuzzy inference system with two inputs x_1 and x_2 and one output y is assumed. Here, x_1 and x_2 might be considered as water temperature T_{mean} and Specific Conductivity SC, while the output y would represent the dissolved oxygen DO. Suppose that the rule base contains two fuzzy IF-THEN rules:

Rule 1: IF x_1 is A_1 and x_2 is B_1, THEN $y = p_1x_1 + q_1x_2 + r_1$ (1a)

Rule 2: IF x_1 is A_2 and x_2 is B_2, THEN $y = p_2x_1 + q_2x_2 + r_2$ (1b)

in which the IF (antecedent) part is fuzzy in nature, while the THEN (consequent) part is a crisp function of an antecedent variable (as a rule, a linear equation). Applied on the above example for pan evaporation, Equations (1a) and (1b) read:

Rule 1: IF T_{mean} is LOW and *S*C is LOW, THEN $DO = p_1T_{mean} + q_1SC + r_1$ (2a)

Rule 2: IF T_{mean} is HIGH and *S*C is MEDIUM, THEN $DO = p_2T_{mean} + q_2SC + r_2$ (2b)

A common rule set may have n inputs and m IF-THEN rules and can be expressed as:

$$y = k_ix_1 + l_ix_2 + ...p_ix_{n-1} + q_ix_n + r_i \quad (3)$$

where k_i, l_i,……p_i, q_i and r_i are parameters with i = 1, 2, 3, …, m corresponding to Rules 1, 2, 3, …, m. The node function in the same layer of the same function family, is described as follow (Jang, 1993):

Layer 1: Every node i in this layer is an adaptive node with node function O_i^1 given by:

$$O_i^1 = \mu_{A_i}(T_{mean}) \quad (4)$$

where T_{mean} is the input to the i-th node and μ is the membership function of A_i which is a linguistic label (such as HIGH, or LOW) associated with this node function. A similar equation as Equation (4) may be considered for the input *SC*.

The node function O_i^1 is the membership function of A_i and specifies the degree to which the given input T_{mean} (or *S*C) satisfies the quantifier A_i. The membership function for A is usually described by bell-functions, such as:

$$\mu_{A_i}(T_{mean}) = \frac{1}{1+[(T_{mean}-c_i)/a_i]^{2b_i}} \quad (5)$$

or

$$\mu_{A_i}(T_{mean}) = \exp\{-(\frac{T_{mean}-c_i}{a_i})^2\} \quad (6)$$

where $\{a_i, b_i, c_i\}$ is the parameter set and μ is the membership function of A_i. As the values of these parameters change, the bell-shaped function varies accordingly, thus exhibiting various forms of membership functions depending on the linguistic label A_i. In fact, any continuous and piecewise differentiable functions, such as commonly used triangular of trapezoidal membership functions, are also qualified candidates for the node function in this layer. Parameters in this layer are referred to as *premise parameters*.

Layer 2: This layer consists of circle nodes labeled TT which multiply incoming signals and sending the product out. For instance:

$$O_i^2 = w_i = \mu_{A_i}(T_{mean})\mu_{B_i}(SC), \; i = 1, 2. \quad (7)$$

Each node output represents the firing strength of a rule.

Layer 3: In this layer, the circle nodes labeled N, calculate the ratio of the i-th rule firing strength to the sum of all rule firing strengths:

$$O_i^3 = \overline{w}_i = \frac{w_i}{w_1 + w_2}, \text{ for } i = 1, 2 \quad (8)$$

The outputs of this layer are referred to as *normalized firing strengths*.

Layer 4: All of the nodes in this layer are adaptive with a node function:

$$O_i^4 = \overline{w}_i y_i = \overline{w}_i(p_iT_{mean} + q_iSC + r_i) \quad (9)$$

where $\overline{w}_i$ is the output of layer 3, and $\{p_i, q_i, r_i\}$ is the para-

Table 2. Summary of the Training, Testing and Validation Processes of ANN and ANFIS Models

Model	Inputs	Model structure	R	RMSE (mg/l)	MAE (mg/l)	MARE (%)
Training process						
ANN1	Tmean	1,1,1	0.841	0.937	0.687	7.953
ANN2	Tmean, SC	2,2,1	0.845	0.928	0.675	7.824
ANN3	Tmean, SC, pH	3,8,1	0.867	0.802	0.568	6.547
ANN4	Tmean, SC, pH, Q	4,6,1	0.877	0.834	0.593	6.862
ANFIS1	Tmean	3	0.844	1.314	1.365	2.028
ANFIS2	Tmean, SC	3,3	0.862	1.242	1.268	2.026
ANFIS3	Tmean, SC, pH	3,3,2,3	0.865	1.094	1.101	2.020
ANFIS4	Tmean, SC, pH, Q	3,3,3,3	0.865	1.112	1.112	2.020
Testing process						
ANN1	Tmean	1,1,1	0.680	1.321	0.823	10.47
ANN2	Tmean, SC	2,2,1	0.699	1.258	0.785	10.04
ANN3	Tmean, SC, pH	3,8,1	0.738	1.245	0.731	9.29
ANN4	Tmean, SC, pH, Q	4,6,1	0.731	1.239	0.731	9.34
ANFIS1	Tmean	3	0.672	1.343	0.835	10.75
ANFIS2	Tmean, SC	3,3	0.691	1.293	0.781	10.59
ANFIS3	Tmean, SC, pH	3,3,2,3	0.756	1.370	0.804	10.43
ANFIS4	Tmean, SC, pH, Q	3,3,3,3	0.774	1.187	0.755	9.10
Validation process						
ANN1	Tmean	1,1,1	0.893	0.768	0.513	6.11
ANN2	Tmean, SC	2,2,1	0.882	0.801	0.577	6.72
ANN3	Tmean, SC, pH	3,8,1	0.908	0.738	0.484	5.67
ANN4	Tmean, SC, pH, Q	4,6,1	0.916	0.673	0.474	5.54
ANFIS1	Tmean	3	0.891	0.775	0.521	10.32
ANFIS2	Tmean, SC	3,3	0.893	0.768	0.557	10.31
ANFIS3	Tmean, SC, pH	3,3,2,3	0.919	0.720	0.520	10.22
ANFIS4	Tmean, SC, pH, Q	3,3,3,3	0.928	0.705	0.537	5.90

Table 3. Preliminary Selection for the Fitness Function of GEP Model Using SI Index

Fitness function based on the absolute error	SI	Fitness function based on the relative error	SI
Absolute error with selection range	0.14	Relative error with selection range	0.13
Absolute/hits	0.21	Relative/hits	0.20
Mean squared error (MSE)	0.15	r-MSE	0.15
Root mean squared error (RMSE)	0.14	r-RMSE	0.15
Mean absolute error (MAE)	0.13	r-MAE	0.13
Relative squared error (RSE)	0.17	r-RSE	0.16
Root relative squared error (RRSE)	0.13	r-RRSE	0.14
Relative absolute error (RAE)	0.13	r-RAE	0.14

meter set. Parameters in this layer are called *consequence parameters*.

Layer 5: The single circle node of this layer, labeled Σ, computes the overall outputs as the summation of all incoming signals:

$$O_i^5 = \sum_{i=1}^{i} \overline{w}_i y_i = \frac{\sum_{i=1}^{i} w_i y_i}{\sum_{i=1}^{i} w_i} \tag{10}$$

Thus, an adaptive network which is functionally equivalent to a Type 3 fuzzy inference system has been constructed. Further details about ANFIS can be found in Jang (1993).

2.4. Genetic Programming

The advantages of a system like Gene Expression Programming (GEP) are clear from nature, but the most important are (Ferreira, 2001a): (1) the chromosomes are simple entities: linear, compact, relatively small, easy to manipulate genetically (replicate, mutate, recombine, etc); (2) the expression trees are exclusively the expression of their respective chromosomes; they are entities upon which selection acts, and according to fitness, they are selected to reproduce with modification. In the present work the GeneXpro program was used for modeling dissolved oxygen. There are also some problems regarding the GP (GEP) application. For instance, in some cases, it is usually observed that the program size (depth of parse tree) starts growing which leads to producing nested functions (i.e., the Bloat

Phenomena) and is not accompanied by any corresponding increase in model fitness. It has some practical effects, because the large programs are computationally expensive to evolve and later use can be hard to interpret. The nested functions give no sense about the physical basis of studied phenomena (Ploi and McPhee, 2008). To overcome this weakness, one should employ some penalization of complex models (limitation of the depth of the parse tree), from which, the Parsimony Pressure tool may be considered as a powerful way for removing un-necessary nesting in the programs (Shiri and Kisi, 2011a).

The procedure to model dissolved oxygen is as follows. The first step is the fitness function. For this problem, the Root Mean Squared Error (RMSE) fitness function, E_i, of an individual program, i, was applied (Ferreira, 2006):

$$E_i = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(P_{ij} - T_j)^2} \tag{11}$$

where P_{ij} is the value predicted by individual program i for fitness case j (out of sample case), and T_j is the target value for fitness case j. For a perfect fit, $P_{ij} = T_j$ and $E_i = 0$. For evaluating the fitness f_i of an individual program i, the following equation should be applied:

$$f_i = 1000\frac{1}{1+E_i} \tag{12}$$

which obviously ranges between 0 and 1,000 with 1,000 corresponded to the ideal (Ferreira, 2006). In case of the application of Parsimony Pressure, which uses the fitness measure as raw fitness, the raw maximum fitness is rf_{max} = 1000, and the overall fitness f_{ppi} (with parsimony pressure) is evaluated by:

$$f_{ppi} = rf_i\cdot\left[1 + \frac{1}{5000}\cdot\frac{S_{\max} - S_i}{S_{\max} - S_{\min}}\right] \tag{13}$$

where the S_i is the program size, S_{max} and S_{min} are the maximum and minimum program sizes, respectively and are evaluated by:

$$S_{max} = G(h + t) \tag{14}$$

$$S_{min} = G \tag{15}$$

where G is the number of genes and h and t are, respectively, the head and tail sizes. In this case the f_{ppmax} is evaluated by the following formula:

$$f_{ppmax} = 1.0002rf_{max} \tag{16}$$

The second step consists of choosing the set of terminals T and the set of functions F, to create the chromosomes. In the current problem, the terminal set includes recorded river water quality data: Tmean, SC, pH, Q and DO. The study examined the various combinations of these parameters as inputs to the GEP models to evaluate the degree of effect of each of these variables on DO at specified time step. The choice of the appropriate function is not so obvious and depends on the viewpoint and guess of user. In this study, different mathematical functions were utilized, including basic arithmetic operators (+, -, *, /) as well as some of the other basic mathematical functions ($\sqrt{}$, $\sqrt[3]{}$, ln(x), e^x, x^2, x^3, sin, cos, arctg). Length of head, h = 8, and three genes per chromosomes were employed, which are commonly used values in literature (e.g., Ferreira, 2001a, 2001b). The fourth step is to choose the linking function. The linking function must be chosen as "addition" or "multiplication" for algebraic sub trees (Ferreira, 2001a). In general, the choice of linking function depends on the problem and there is not any basic rule to identify which of these functions is preferred to another. Here, various linking functions will be examined to choose the best one. The fifth and final step is to choose the genetic operators. The parameters used per run are summarized as follows: number of chromosomes: 30, head size: 8, number of genes: 3, linking function: addition, fitness function error type: root mean squared error, mutation rate: 0.044, inversion rate: 0.1, one point recombination rate: 0.3, two point recombination rate: 0.3, gene recombination rate: 0.1, gene transposition rate: 0.1, insertion sequence transposition rate: 0.1, root insertion sequence transposition: 0.1. It is noted that these parameters are default values of GeneXpro program and can be used in various modeling applications (e.g. Shiri and Kisi, 2011a; Shiri et al., 2012).

2.5. Models' Assessment Parameters

Four statistical evaluation parameters were used to assess the models' performances:

(1) The Correlation coefficient (R):

$$R = \frac{\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \overline{x})^2\sum_{i=1}^{N}(y_i - \overline{y})^2}} \tag{17}$$

(2) The root mean square error (*RMSE*) defined as:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(y_i - x_i)^2}{N}} \tag{18}$$

(3) The mean absolute relative error (*MARE)* defined as:

$$MARE = \frac{1}{n}\left(\sum_{i=1}^{n} abs\left(\frac{y_i - x_i}{x_i}\right)\right) \tag{19}$$

(4) Mean absolute error (*MAE*):

$$MAE = \frac{\sum_{i=1}^{n} abs(y_i - x_i)}{n} \tag{20}$$

(5) Scatter index (*SI*):

$$SI = \frac{RMSE}{\overline{x}} \tag{21}$$

where, x_i is the value observed at the ith time step, y_i is the corresponding simulated value, n is number of time steps, $\overline{x}$ is mean of observational values and $\overline{y}$ is mean value of the simulations. The perfect value of *R* index is 1, representing the best fit of observed versus simulated values, but it cannot be alone used as the unique performance evaluation index since it is sensitive to outliers (Legates and McCabe, 1999). Therefore, it is better to apply other indexes along with R^2, e.g. *RMSE*, *MARE* and *MAE*. *RMSE* describe the average magnitude of the errors between the observational values and model results. *MARE* and *MAE* are linear scouring rule and describes only the average magnitude of the errors, ignoring their direction. The combined use of these parameters can provide a sufficient insight about the performance of the applied methodologies.

3. Application and Results

Various input combinations of water temperature, specific conductance, pH and discharge parameters were tried as inputs to the ANN and ANFIS methods in this study to evaluate the degree of effect of each of variables on dissolved oxygen concentration. The input combinations used in the current study are: (1) Tmean; (2) Tmean and SC; (3) Tmean, SC and pH; (4) Tmean, SC, pH and Q.

Before applying the ANN to the data, the training input and output values were normalized using the equation:

$$c\frac{x_i - x_{\min}}{x_{\max} - x_{\min}} + d \tag{22}$$

where x_{min} and x_{max} indicates the minimum and maximum of the training data. Various values can be taken for the *c* and *d* scaling factors. There is no fixed rule for standardization (Dawson and Wilby 1998). In this study, the *c* and *d* were assigned as 0.6 and 0.2, respectively. Thus, the data were normalized to fall in the range (0.2, 0.8). The training, testing and validation data were scaled between 0.2 and 0.8 following the suggestion of Cigizoglu (2003) who showed that scaling input data between 0.2 and 0.8 gives the ANNs the flexibility to predict beyond the training range. Different ANN structures were tried to find optimal model for each input combination. The optimal ANN structures are given in Table 2. In this table, (4, 6, 1) indicates an ANN model comprising four inputs corresponding to Tmean, SC, pH and Q inputs, 6 hidden and 1 output nodes. As can be clearly seen from the table that six hidden nodes are enough for modeling DO concentration. The testing and validation results of the ANN models are shown in Table 2. It is clear from the table that the ANN4 model whose inputs are Tmean, SC, pH and Q has the lowest RMSE (1.239 mg/l), MAE (0.731 mg/l) and MARE (9.34%) and highest R (0.731) in testing phase.

Different ANFIS structures were also employed to find optimal model in modeling DO concentration. The optimal ANFIS structures are provided in Table 2. In this table, (3, 3, 3, 3) reveals an ANFIS model comprising three membership functions for each input. The RMSE, MAE, MARE and R values of the optimal ANFIS models are given in Table 2. As found for the ANN method, the ANFIS4 model comprising four has the lowest RMSE (1.187 mg/l), MAE (0.755 mg/l) and MARE (9.10%) and highest R (0.774) in testing phase. Comparison of optimal ANN and ANFIS models' validation results indicates that the ANN4 model performs better than the ANFIS4 model with respect to RMSE, MAE and MARE criteria. For both methods, fourth input combination showed the best accuracy.

Table 4. Preliminary Selection of Basic Functions for the Parse Tree

	Definition	SI
F1	$\{+,-,\times,\div\}$	0.15
F2	$\{+,-,\times,\div,\ln,e^x\}$	0.14
F3	$\{+,-,\times,\div,\sqrt[3]{},\sqrt{},x^3,x^2\}$	0.15
F4	$\{+,-,\times,\div,\sqrt[3]{},\sqrt{},\ln,e^x,x^2,x^3\}$	0.09
F5	$\{+,-,\times,\div,\sqrt[3]{},\sqrt{},\ln,e^x,x^2,x^3,\sin x,\cos x,Arctgx\}$	0.13

Table 5. Investigation on Various GEP Linking Functions by Using SI

Linking functions	SI
Addition	0.09
Multiplication	0.15
Subtraction	0.14
Division	0.18

Table 6. Train, Test and Validation Results of the GEP Model for the Optimal Input Combination

	R	RMSE (mg/l)	MAE (mg/l)	MARE (%)
Training	0.850	1.160	0.930	8.02
Testing	0.889	0.843	0.635	7.10
Validation	0.930	0.660	0.487	5.60

The optimal GEP model was obtained for the fourth input combination. The first step with GEP modeling is to select the appropriate fitness function. The best input combination (input combinations (iv)) was used with default function set of GeneXpro (i.e. +, -, ×, ÷, $\sqrt[3]{}$, $\sqrt{}$, ln, e^x, x^2, x^3, sin*x*, cos*x*, arctg*x*) for the selection of one of the fitness functions (Table 3). It is clear from the SI values given in Table 3 that the r-MAE fitness function gives the most accurate results among others. Therefore, it was decided to apply the r-MAE fitness function (based on relative error) in foregoing process. The next step is selecting the terminal set and function sets. For selecting the basic operators for building the parse tree a range of basic fun-

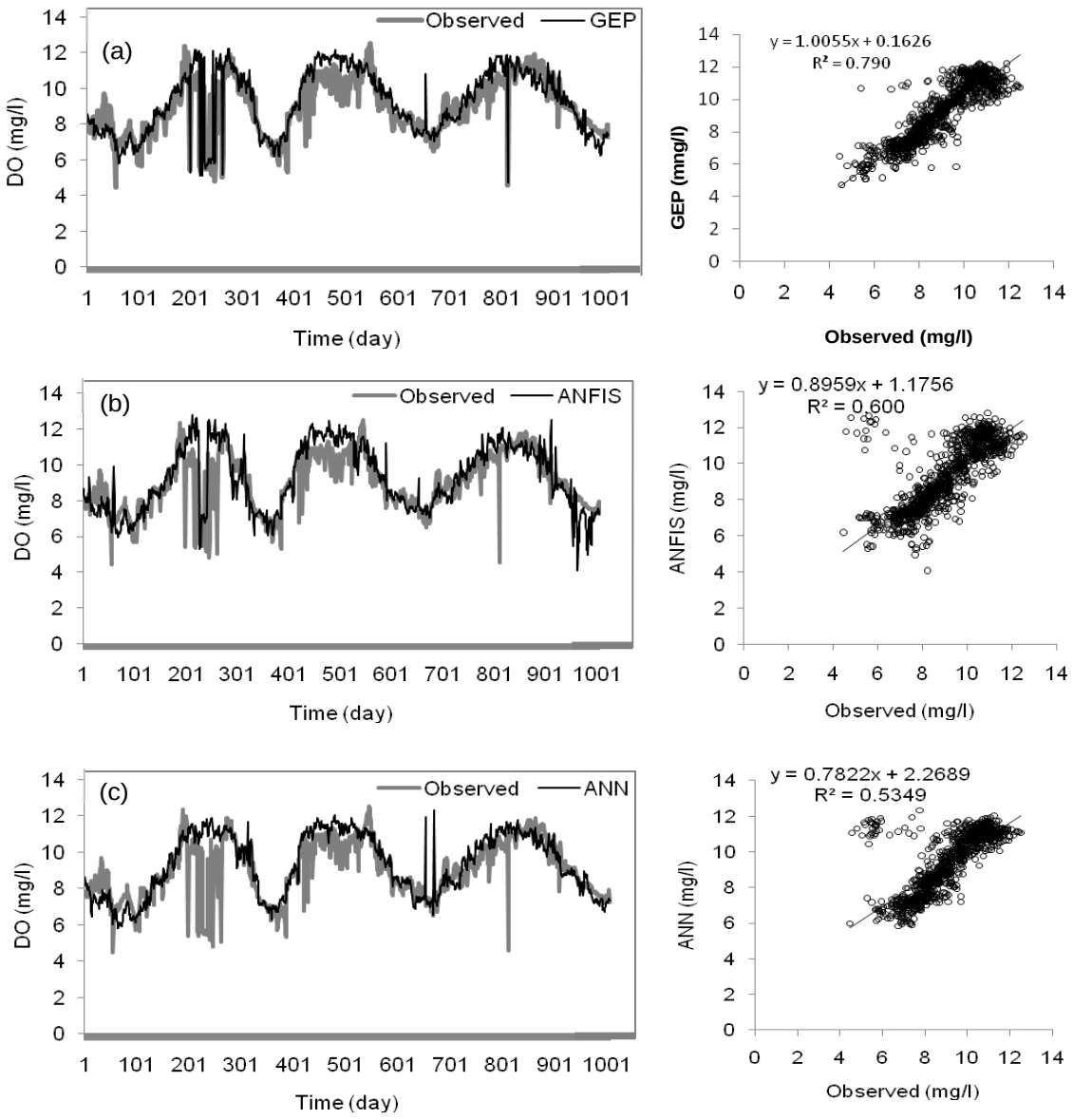

Figure 2. Observed and estimated DO concentrations by the GEP, ANN and ANFIS models in the test stage.

ctions were investigated as shown in Table 4. A set of preliminary model runs was carried out to test the performances of models with these function sets and select one in the next stage of the study. All of these procedures were performed for GEP model comprising fourth input combination by using the r-MAE fitness function. The results of the function sets are presented in Table 4 in terms of SI criterion. From the comparison of various GEP operators listed in this table, it can be said that the GeneXpro F4 function set surpasses all of the other four structures. Table 5 indicates the sensitivity analysis of various GEP setting options. It is clear from the table that the Addition linking function performs better than the others. Therefore, the addition is selected to link the sub trees. The test and validation results of the optimal GEP model for the fourth input combination are given in Table 6. Comparison of GEP and ANFIS4 model reveals that the GEP model performs better than the ANFIS4 model in both test and validation phases. From Table 2 and 6, it is clear that the GEP model give better accuracy than the ANN4 model from the RMSE and R viewpoint in validation stage. However ANN4 model shows slightly better accuracy than the GEP model with respect to MAE and MARE statistics. One of the key advantages of the GP (i.e. GEP) model over the other AI models is in giving explicit mathematical expression of the governing rule between the input-output variables. The optimal GEP equation is given:

$$DO = PH + arctg\left[9.28\left(\sqrt{Q}-\sqrt{SC}\right)^2\right] + \cos\left[\frac{T_{mean}}{0.38PH - 9.86}\right] \quad (23)$$

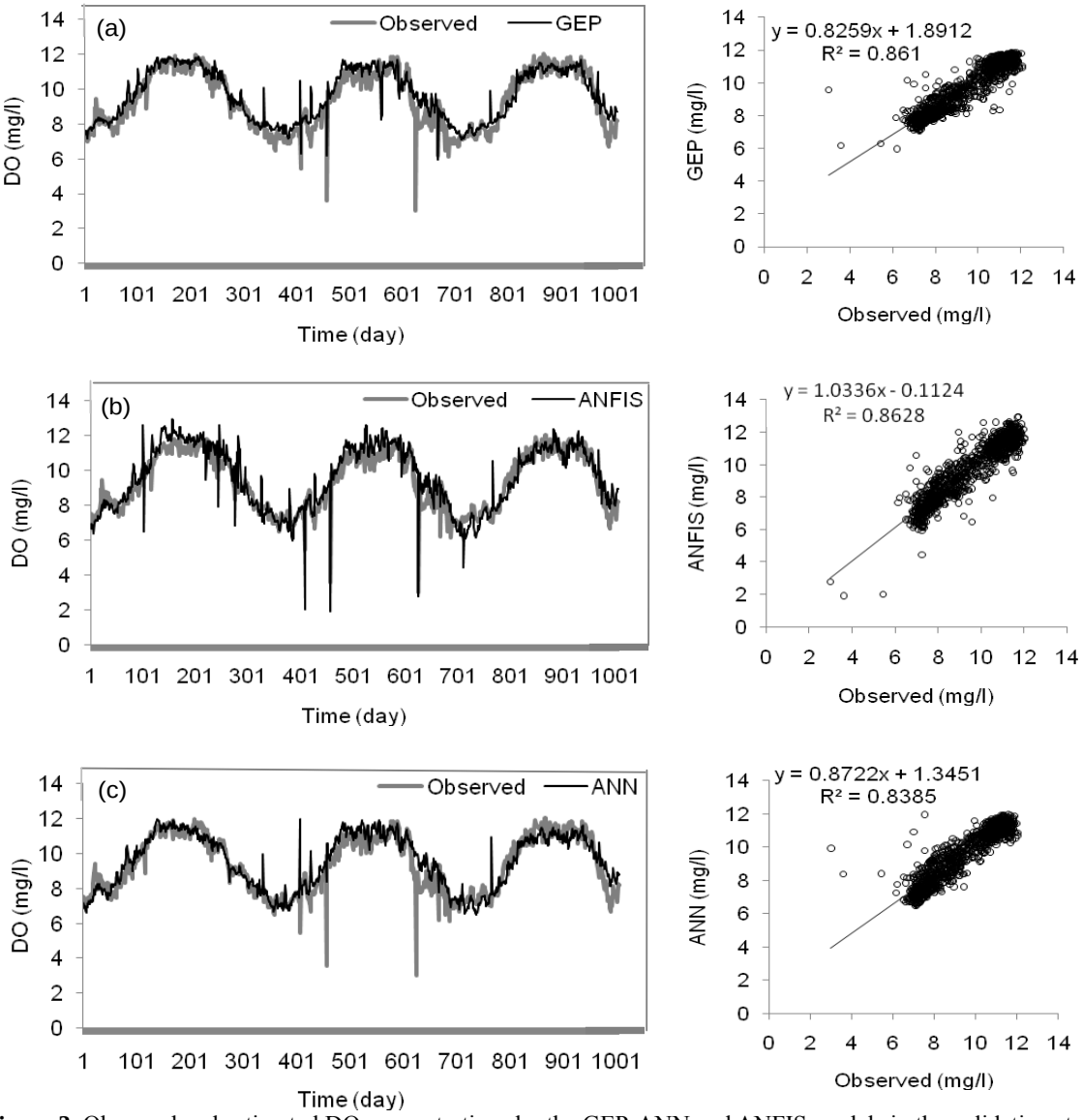

Figure 3. Observed and estimated DO concentrations by the GEP, ANN and ANFIS models in the validation stage.

Time variation graphs and scatterplots of the observed and estimated DO concentrations by the GEP, ANN and ANFIS models in the test stage are illustrated in Figure 2. It is clear from the time variation graphs that the GEP estimates closely follows the corresponding observed DO values. It can be seen from the fit line equations (assume that the equation is $y = a_0x + a_1$) given in scatterplots that a_0 and a_1 coefficients are closer to the 1 and 0 than those of the ANN and ANFIS models with a higher R^2 value, respectively. The validation results of each model are shown in Figure 3. From scatterplots, it is clear that GEP and ANFIS models have higher R^2 values than the ANN model and the ANFIS model has more scattered estimates than the GEP model. From time variation graphs, it is difficult to compare models with each model. As an example, the observed and estimated DO concentrations by the GEP, ANN and ANFIS models for the period of 700 ~ 800 days are shown in Figure 4. Figure 4 indicates that the GEP model's estimates are closer to the corresponding DO concentration values than those of the ANFIS and ANN models.

4. Conclusions

The ability of GEP, ANN and ANFIS models in modeling DO concentration was investigated in this study. Mean water temperature, specific conductance, pH, discharge and DO concentration data from South Platte River at Englewood, Colorado were used as a case study. First, various input combinations of Tmean, SC, pH and Q parameters were tried as inputs to the ANN and ANFIS methods to evaluate the degree of effect of each of variables on DO concentration. Out of four ANN and ANFIS models comprising different input combinations, the models with the Tmean, SC, pH and Q parameters were found

to be the best. Then, the optimal GEP model was obtained for the fourth input combination. The optimal GEP, ANN and ANFIS models were compared with each other with respect to correlation coefficient, root mean square error, mean absolute error and mean absolute relative error criteria. Comparison results indicated that the GEP model performed better than the other models in modeling DO concentration. The main advantage of GEP over the ANN and ANFIS techniques is that it has an explicit formulation and simple. It can be used by anyone not necessarily being familiar with GEP. The study only used data from one area and further studies using more data from various areas may be required to strengthen these conclusions.

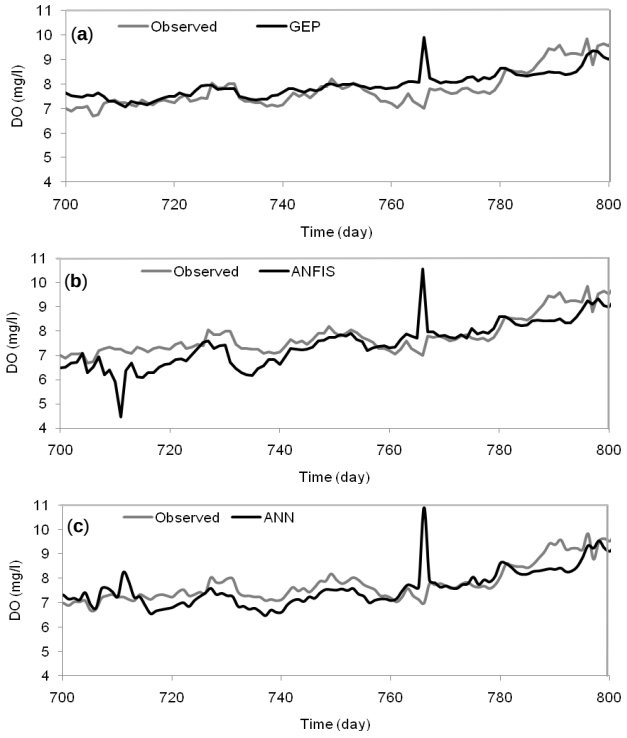

Figure 4. Observed and estimated DO concentrations by the GEP, ANN and ANFIS models for the period of 700-800 days in the validation stage.

References

Amercan Society of Civil Engineers (ASCE) Task Committee on Application of Artificial Neural Networks in Hydrology. (2000). Artificial neural networks in hydrology, I: Hydrologic applications. *J. Hydrol. Eng.,* 5(2), 124-137. http://dx.doi.org/10.1061/(ASCE)1084-0699(2000)5:2(124)

Ay, M., and Kisi, O. (2012). Modeling of dissolved oxygen concentration using different neural network techniques in Foundation Creek, El Paso County, Colorado, USA. *J. Environ. Eng.,* 138(6), 654-662. http://dx.doi.org/10.1061/(ASCE)EE.1943-7870.0000511

Azamathulla, H.M., and Ghani, A.A. (2011). ANFIS-Based Approach for Predicting the Scour Depth at Culvert Outlets. *J. Pipeline Syst. Eng. Pract.,* 2(1), 35-40. http://dx.doi.org/10.1061/(ASCE)PS.1949-1204.0000066

Bae, D.H., Jeong, D.M., and Kim, G. (2007). Monthly dam inflow forecasts using weather forecast information and neuro-fuzzy technique. *Hydrol. Sci. J.,* 52(1), 99-113. http://dx.doi.org/10.1623/hysj.52.1.99

Chang, F.J., and Chen, Y.C. (2001). Counter propagation fuzzy-neural network modeling approach to real time streamflow prediction. *J. Hydrol.,* 245, 153-164. http://dx.doi.org/10.1016/S0022-1694(01)00350-X

Cigizoglu, H.K. (2003). Estimation, forecasting and extrapolation of flow data by artificial neural networks. *Hydrol. Sci. J.,* 48(3), 349-361. http://dx.doi.org/10.1623/hysj.48.3.349.45288

Dawson, W.C., and Wilby, R. (1998). An artificial neural network approach to rainfall-runoff modeling. *Hydrol. Sci. J.,* 43(1), 47-66. http://dx.doi.org/10.1080/02626669809492102

Faruk, D.O. (2010). A hybrid neural netork and ARIMA model for water quality time series prediction. *Eng. Appl. Artificial Intell.,* 23, 586-594. http://dx.doi.org/10.1016/j.engappai.2009.09.015

Ferreira, C. (2001a). Gene expression programming in problem solving. In: 6th Online World Conference on Soft computing in Industrial Applications (invited tutorial).

Ferreira, C. (2001b). Gene expression programming: a new adaptive algorithm for solving problems. *Complex Syst.,* 13(2), 87-129.

Ferreira, C. (2006). Gene expression programming: Mathematical Modeling by an artificial intelligence. Springer, Berling, Heidelberg New York, NY, USA, 478 pp.

Giustolisi, O. (2004). Using GP to determine Chezzy resistance coefficient in corrugated channels. *J. Hydroinf.,* 157-173.

Goldberg, D.E. (1989). Genetic algorithms in search, optimization, and machine learning. Addison-Wesley, Reading MA 432 pp.

Harris, E.L, Babovic, V., and Falconer, R.A. (2003). Velocity predictions in compound channels with vegetated flood plains using genetic programming. *Int. J. River Basin Manage.,* 1(2), 117-123. http://dx.doi.org/10.1080/15715124.2003.9635198

Haykin, S. (1999). Neural Networks-A Comprehensive Foundation. (2nd. ed.). Prentice-Hall, Upper Saddle River, NJ, USA, 26-32.

Jang, J.S.R. (1993). ANFIS: adaptive-network-based fuzzy inference system. *IEEE Trans. Syst. Manag. Cyber.,* 23(3), 665-685. http://dx.doi.org/10.1109/21.256541

Jang, J.S.R., Sun, C.T., and Mizutani, E. (1997). Neurofuzzy and soft computing: a computational approach to learning and machine intelligence. Prentice-Hall, NJ, USA.

Karimi, S., Shiri, J., Kisi, O., and Makarynskyy, O. (2012). Forecasting water level fluctuations of Urmieh lake using gene expression programming and adaptive neuro-fuzzy inference system. *Int. J. Ocean Clim. Syst.,* 3(2), 109-125. http://dx.doi.org/10.1260/1759-3131.3.2.109

Kisi, O. (2006). Daily pan evaporation modeling using a neuro-fuzzy computing technique. *J. Hydrol.,* 329, 636-646. http://dx.doi.org10.1016/j.jhydrol.2006.03.015

Kisi, O, Yuksel, I., and Dogan, E. (2008). Modelling daily suspended sediment of rivers in Turkey using several data driven techniques. *Hydrol. Sci. J.,* 53(6), 1270-1285. http://dx.doi.org/10.1623/hysj.53.6.1270

Kisi, O., and Shiri, J. (2011). Precipitation forecasting using wavelet-genetic programming and wavelet-neuro-fuzzy conjunction models. *Water Resour. Manage.,* 25(13), 3135-3152. http://dx.doi.org/10.1007/s11269-011-9849-3

Kisi, O., and Shiri, J. (2012a). Wavelet and neuro-fuzzy conjunction model for predicting water table depth fluctuations. *Hydrol. Res.,* 43(3), 286-300.

Kisi, O., and Shiri, J. (2012b). River suspended sediment estimation by climatic variables implication: comparative study among soft computing techniques. *Comput. Geosci.,* 43, 73-82. http://dx.doi.org/10.1016/j.cageo.2012.02.007

Kisi, O., Shiri, J., and Tombul, M. (2013). Modeling rainfall-runoff process using soft computing techniques. *Comput. Geosci.,* 51,108-

117. http://dx.doi.org/10.1016/j.cageo.2012.07.001

Koza, J.R. (1992). Genetic Programming: On the programming of computers by means of Natural Selection. The MIT Press, Cambridge, MA 840 pp.

Legates, D.R., and McCabe, G.J. (1999). Evaluating the use of goodness-of-fit measures in hydrologic and hydroclimatic model validation. *Water Resour. Res.,* 35(1), 233-241. http://dx.doi.org/10.1029/1998WR900018

Maier, H.R., and Dany, G.C. (1996). The use of artificial neural networks for the preidiction of water quality parameters. *Water Resour. Res.,* 32(4), 1013-1022. http://dx.doi.org/10.1029/96WR03529

Mamdani, E.H., and Assilian, S. (1975). An experiment in linguistic synthesis with a fuzzy logic controller. *Int. J. Man Mach. Stud.,* 7(1), 1-13. http://dx.doi.org/10.1016/S0020-7373(75)80002-2

Ozger, M., and Yildirim, G. (2009). Determining turbulent flow friction coefficient using adaptive neuro-fuzzy computing technique. *Adv. Eng. Software,* 40, 281-287. http://dx.doi.org/10.1016/j.advengsoft.2008.04.006

Palani, S., Liong, S.Y., and Tkalich, P. (2008). An ANN application for water quality forecasting. *Mar. Pollut. Bull.,* 56, 1586-1597. Http://dx.doi.org/10.1016/j.marpolbul.2008.05.021

Ploi, R., and McPhee, N.F. (2008). Covariant parsimony pressure for genetic programming. Technical report CES-480, ISSN: 1744-8050.

Rankovic, V., Radulovic, J., Radojevic, I., Ostojic, A., and Comic, A. (2010). Neural network modeling of dissolved oxygen in the Gruza reservoir, Serbia. *Ecol. Model.,* 221, 1239-1244.

Schmid, B.H., and Koskiaho, J. (2006). Artificial neural network modeling of dissolved oxygen in a wetland pond: the case study Hovi, Finland.

Shiri, J., and Kisi, O. (2010). Short-term and long-term streamflow forecasting using a wavelet and neuro-fuzzy conjunction model. *J. Hydrol.,* 394(3-4), 486-493. http://dx.doi.org/10.1016/j.jhydrol.2010. 10.008

Shiri, J., and Kisi, O. (2011a). Comparison of genetic programming with neuro-fuzzy systems for predicting short-term water table depth fluctuations. *Comput. Geosci.,* 37(10), 1692-1701. http://dx.doi.org/10.1016/j.cageo.2010.11.010

Shiri, J., and Kisi, O. (2011b). Application of artificial intelligence to estimate daily pan evaporation using available and estimated climatic data in the Khozestan Province (South Western Iran). *J. Irrig. Drain. Eng.,* 137(7), 412-425. http://dx.doi.org/10.1061/(ASCE)IR.1943-4774.0000315

Shiri, J., Makarynskyy, O., Kisi, O., Dierickx, W., and Fakheri Fard, A. (2011a). Prediction of short term operational water levels using an adaptive neuro-fuzzy inference system. *J. Waterway Port Coast. Ocean Eng.,* 137(6), 344-354. http://dx.doi.org/10.1061/(ASCE)WW.1943-5460.0000097

Shiri, J., Dierickx, W., Pour-Ali Baba, A., Neamati, S., and Ghorbani, M.A. (2011b). Estimating daily pan evaporation from climatic data of the State of Illinois, USA using adaptive neuro-fuzzy inference system (ANFIS) and artificial neural networks (ANN). *Hydrol. Res.,* 42(6), 491-502. http://dx.doi.org/10.2166/nh.2011.020

Shiri, J., Kisi, O., Landeras, G., Lopez, J.J., Nazemi, A.H., and Stuyt, L.C.P.M. (2012). Daily reference evapotranspiration modeling by using genetic programming approach in the Basque Country (Northern Spain). *J. Hydrol.,* 414-415, 302-316. http://dx.doi.org/10.1016/j.jhydrol.2011.11.004

Singh, K.P., Basant, A., Malik, A., and Jain, G. (2009). Artificial neural network modeling of the river water quality-A case study. *Ecol. Model.,* 220, 888-895. http://dx.doi.org/10.1016/j.ecolmodel.2009.01.004

Takagi, T., and Sugeno, M. (1985). Fuzzy identification of systems and its application to modeling and control. *Trans. Syst. Man Cybernetics,* 15(1), 116-132. http://dx.doi.org/10.1109/TSMC. 1985.6313