EARTH SCIENCES
RESEARCH JOURNAL

Earth Sci. Res. J. Vol. 20, No. 2 (June, 2016) : D1-D11.

UNIVERSIDAD NACIONAL DE COLOMBIA
SEDE BOGOTÁ
FACULTAD DE CIENCIAS
DEPARTAMENTO DE GEOCIENCIAS
RESEARCH GROUP IN GEOPHYSICS

HYDROLOGY

Water temperature prediction in a subtropical subalpine lake using soft computing techniques

Saeed Samadianfard[1,*], Honeyeh Kazemi[1], Ozgur Kisi[2], Wen-Cheng Liu[3]

[1]Department of Water Engineering, University of Tabriz, Tabriz, Iran (*corresponding author, e-mail: s.samadian@tabrizu.ac.ir, Tel: +989141101845, Fax: +984113356007)

[2]Department of Civil Engineering, Canik Basari University, Samsun, Turkey

[3]Department of Civil and Disaster Prevention Engineering, National United University, Miao-Li, Taiwan

ABSTRACT

Lake water temperature is one of the key parameters in determining the ecological conditions within a lake, as it influences both chemical and biological processes. Therefore, accurate prediction of water temperature is crucially important for lake management. In this paper, the performance of soft computing techniques including gene expression programming (GEP), which is a variant of genetic programming (GP), adaptive neuro fuzzy inference system (ANFIS) and artificial neural networks (ANNs) to predict hourly water temperature at a buoy station in the Yuan-Yang Lake (YYL) in north-central Taiwan at various measured depths was evaluated. To evaluate the performance of the soft computing techniques, three different statistical indicators were used, including the root mean squared error (RMSE), the mean absolute error (MAE), and the coefficient of correlation (R). Results showed that the GEP had the best performances among other studied methods in the prediction of hourly water temperature at 0, 2 and 3 meter depths below water surface, but there was a different trend in the 1 meter depth below water surface. In this depth, the ANN had better accuracy than the GEP and ANFIS. Despite the error (RMSE value) is smaller in ANN than GEP, there is an upper bound in scatter plot of ANN that imposes a constant value, which is not suitable for predictive purposes. As a conclusion, results from the current study demonstrated that GEP provided moderately reasonable trends for the prediction of hourly water temperature in different depths.

Keywords: Soft computing techniques, statistical indicators, subalpine lake, water temperature.

RESUMEN

La temperatura del agua es uno de los parámetros básicos para determinar las condiciones ecológicas de un lago, ya que está influenciada por procesos químicos y biológicos. Además, la exactitud en la predicción de la temperatura del agua es esencial para el manejo del lago. En este artículo se evalúa el desempeño de técnicas de soft computing como la Programación de Expresiones de Genes (PEG), que es una variante de la Programación Genética (PG), el Sistema Neuro-fuzzy de Inferencia Adaptativa (Anfis, en inglés) y las Redes Neuronales Artificiales (RNA) para predecir la temperatura del agua en diferentes niveles de una estación flotante del lago Yuan-Yang (YYL), en el centro-norte de Taiwán. Se utilizaron tres indicadores estadísticos, el Error Cuadrático Medio (ECM), el Error Absoluto Medio (MAE, en inglés) y el Coeficiente de Correlación (R) para evaluar el desempeño de las técnicas de computación. Los resultados muestran que la PEG es más exacta en la predicción de la temperatura del agua entre 0,2 y 3 metros de profundidad. Sin embargo, se evidencia una tendencia diferente a partir del metro de profundidad. A esta distancia de la superficie, las RNA son más exactas que la PEG y el Anfis. Los resultados de este estudio probaron claramente la usabilidad del PEG y las RNA en la predicción de la temperatura del agua a diferentes profundidades.

Palabras clave: Técnicas soft computing, indicadores estadísticos, lago subalpino, temperatura del agua.

Record

Manuscript received: 29/04/2014
Accepted for publication: 26/02/2016

How to cite item

Earth Sciences Research Journal, 20(2), B-B.
doi:http://dx.doi.org/10.15446/esrj.v20n2.43199

1. Introduction

Water temperature is a fundamental physical property with a direct impact on all organisms inhabiting the aquatic environment (Webb et al., 2008). Because predicting the water temperature is important for maintaining water quality and for ecosystem management, several authors have investigated methods for simulating water temperature of lakes (Lawrence et al., 2002; Lee et al., 2009; Schwab et al., 2009). Hondzo and Stefan (1996) simulated daily water temperature and dissolved oxygen profiles in Minnesota lakes by deterministic process-based water quality models with daily meteorological conditions as input. Derived empirical formulas for lake water quality and stratification indicators from the simulation results gave good predictions of temperature and dissolved oxygen characteristics estimated from measurements in seven Minnesota lakes. Fang and Stefan (1996) substantially modified a water temperature and ice cover model for freshwater lakes and combined it with a summer model to simulate water temperature structures and ice thickness in two small lakes in the north central US. The best value of volume averaged water temperature for the two studied small lakes was higher than was previously found for a larger lake. Benyahya et al. (2007) provided an overview of the existing statistical water temperature models. They categorized them in two major groups: deterministic and statistical/stochastic models. They stated that the deterministic models require numerous input data and they are appropriate for analyzing different impact scenarios, but the main advantage of the statistical models is their relative simplicity and relative minimal data requirement. Sharma et al. (2008) developed models to predict annual maximum near-surface lake water temperatures for lakes across Canada using four statistical approaches: multiple regression, regression tree, artificial neural networks and Bayesian multiple regression. Although artificial neural networks were marginally better for three of the four data sets, multiple regression was considered to provide the best solution based on the combination of model performance and computational complexity. Trumpickas et al. (2009) tried to construct empirical relationships between surface water temperatures and local air temperatures that could be used to estimate future water temperatures using future air temperatures generated by global climate models. Zhao et al. (2011) calculated the surface water temperature, deep water temperature and mean annual epilimnetic temperature and compared the obtained values with empirical data in Lake Taihu by using the dynamic water temperature model. The simulated values were consistent with empirical data. Results showed that this model could be used for the temperature simulation in the studied lake. Thiery at al. (2014) evaluated a set of one dimensional lake models for Lake Kivu, East Africa. In this study, meteorological observations from two automatic weather stations were used to drive the models, whereas a unique dataset, containing over 150 temperature profiles recorded since 2002, was used to assess the model's performance. Simulations were performed over the freshwater layer only (60 m) and over the average lake depth (240 m). The good agreement between the deep simulations and the observed meromictic stratification also showed that a subset of models was able to account for the salinity- and geothermal induced effects upon deep-water stratification.

Also, some attempts have been made to relate lake water temperature to meteorological parameters such as air temperature. In this context, Piccolroaz et al. (2013) developed Air2Water, a simple physically based model to relate the temperature of the Lake Superior (USA–Canada), considering a 27 years record of measurements, to air temperature only. The results proved that their model was suitable to be applied over long timescales (from monthly to interannual), and could be easily used to predict the response of a lake to climate change, since projected air temperatures were usually available by large-scale global circulation models. Toffolon et al. (2014) reconstructed temperature of the surface layer of temperate lakes by means of a simplified model on the basis of air temperature alone. The comparison between calculated and observed data showed a remarkable agreement for all 14 lakes investigated (Mara, Sparkling, Superior, Michigan, Huron, Erie, Ontario, Biel, Zurich, Constance, Garda, Neusiedl, Balaton, and Baikal, in west-to-east order), which presented a wide range of morphological and hydrological characteristics.

Given the significant impact of lakes on surface atmosphere interactions, the need for an accurate simulations of lake temperatures at different depths arises. For this purpose, soft computing techniques such as gene expression programming (GEP), adaptive neuro fuzzy inference system (ANFIS) and artificial neural networks (ANN) can play undeniable roles in mentioned simulations.

Gene expression programming (GEP) has been applied to a wide range of problems in artificial intelligence, artificial life, engineering and science. GEP can be successively applied to areas where (i) the interrelationships among the relevant variables are poorly understood (or where it is suspected that the current understanding may well be wrong), (ii) finding the size and shape of the ultimate solution is difficult, (iii) conventional mathematical analysis does not, or cannot, provide analytical solutions, (iv) an approximate solution is acceptable (or is the only result that is ever likely to be obtained), (v) small improvements in performance are routinely measured (or easily measurable) and highly prized, (vi) there is a large amount of data in computer readable form, that requires examination, classification, and integration, e.g., molecular biology for protein and DNA sequences, astronomical data, satellite observation data, financial data, marketing transaction data, or data on the World Wide Web (Banzhaf et al. 1998). During the last decade, genetic programming has been used as a viable alternative approach to physical models. Aytek and Kisi (2008) applied GP to suspended sediment transport, and found it to perform better than conventional rating curve and multi-linear regression techniques. Shiri and Kisi (2011) compared GEP and ANFIS methods for predicting groundwater table depth fluctuations and found GEP to be better than ANFIS in this regard. Samadianfard (2012) examined the potential of the GEP technique in estimating flow friction factor in comparison with the most currently available explicit alternatives to the Colebrook–White equation. Results revealed that by using GEP, the friction factor could be identified precisely. Samadianfard et al. (2012) studied the capabilities of the GP in simulating the wetting patterns of drip irrigation. Results showed that the GP method had good agreement with the results of HYDRUS 2D software considering the full set of operators in estimation of radius and depth of wetting patterns. Also results obtained from field experimental in a sandy loam soil showed reasonable agreement with the GP results. Results of the study demonstrated the usefulness of the GP method for estimating wetting patterns of drip irrigation.

ANFIS is a neuro-fuzzy system, which uses a feed-forward network to search for fuzzy decision rules that perform well on a given task. Using a given input/output data set, ANFIS creates a fuzzy inference system whose membership function parameters are adjusted using a back-propagation algorithm alone or a combination of a back-propagation algorithm with a least mean square method. This allows the fuzzy systems to learn from the data being modeled. Kisi (2006) investigated the ability of ANFIS technique to improve the accuracy of daily evaporation estimation. Based on his results, the ANFIS computing technique could be used successfully in modeling evaporation process from the available climatic data. Shiri et al. (2011) compared ANFIS to ANN to estimate daily pan evaporation values from climatic data and found ANFIS to be better than ANN.

The artificial neural network (ANN) approach provides a viable solution to the environmental problems because it is based on training not on analytical models or statistical assumptions. ANN models can be trained to predict results from examples and once trained; they can perform predictions at very high speed (Mellit et al., 2006). ANN is an intelligent data-driven modeling tool that is able to capture and represent complex and non-linear input/output relationships. ANNs are massively parallel, distributed processing systems that can continuously improve their performance via dynamic learning. Moghaddamnia et al. (2009) explored evaporation estimation methods based on ANN and ANFIS techniques. They found that ANN and ANFIS techniques had much better performances than the empirical formulas. Zaier et al. (2010) developed ANN ensemble models to improve the results of single artificial neural network (single ANN) for the estimation of the ice thickness in a number of selected Canadian lakes during the early winter ice growth period. ANN ensemble models for the estimation of ice thickness proved to be more accurate than single ANN models. Kisi et al. (2012) applied three artificial intelligence approaches,

namely ANNs, ANFIS and GEP to forecast daily lake-level variations of Lake Iznik in Turkey. The results obtained by the GEP approach indicated that it performed better than ANFIS and ANNs in predicting lake-level variations. Liu and Chen (2012) compared the performance of the ANN technique with a physically based three-dimensional circulation model for prediction of water temperature at a buoy station in the Yuan-Yang Lake in north-central Taiwan at various measured depths. The simulated results revealed that the accuracy of the three-dimensional circulation model was better than the ANN model. Fallah-Mehdipour et al. (2013) investigated the capability of ANFIS and GP as two artificial intelligence tools to predict and simulate groundwater levels in three observation wells in the Karaj plain, Iran. Results indicated that GP yielded more appropriate results than ANFIS when different combinations of input data were employed in both prediction and simulation processes.

The main objective of this paper is to investigate the accuracy of soft computing techniques such as gene expression programming, adaptive neuro fuzzy inference system and artificial neural network methods for the prediction of hourly water temperature in a lake at different measured depths. Some statistical parameters for error estimation are used herein as comparing criteria for the evaluation of the performance of the studied models.

2. Material and Methods

2.1. Study site and data collection

Yuan-Yang Lake (YYL, 24°34'60.00'N, 121°24'0.00'E, area= 3.7×10^4 m^2; maximum depth= 4.5 m) in north-central Taiwan is a subtropical, humid lake in the Clilan National Forest Preserve (Figure 1). Figure 2 shows the bathymetry of the YYL (contours in m). Also, YYL is a subalpine lake situated in a natural reserve area that has been virtually undisturbed by human activity for a long time (Chen and Wu, 1999). It is surrounded by a cloud belt forest of Taiwan yellow cypress, an important timber tree and a relic species. The lake is located 1670 m above the mean sea level which is at the subalpine region. The geography of the drainage basin allows large quantities of terrestrial runoff from the surrounding mountains to enter the YYL (Wu et al., 2001). The mean annual air temperature is approximately 13°C (monthly averages range from -5 to 15°C), and annual precipitation can exceed 4000 mm. Wind speed over the lake, which was measured 1 m above the lake by an anemometer, (between 0 to 4.220 $m.s^{-1}$) is relatively weak. The dominant wind directions are from the east and the south west because of the V-shaped valley facing east to west. Concerns about the water quality in YYL, have been rapidly increasing recently due to the natural and anthropogenic pollution. In order to understand the underlying physical and chemical processes as well as their associated spatial distribution in YYL, Liu et al. (2011) analyzed fourteen physico-chemical water quality parameters recorded at the eight sampling stations by using multivariate statistical techniques and a geostatistical method. Their results showed that four principal components i.e., nitrogen nutrients, meteorological factor, turbidity and nitrate factors, account for 65.52% of the total variance among the water quality parameters. The spatial distribution of principal components further confirmed that nitrogen sources constitute an important pollutant contribution in the YYL. Since April 2004, the deepest point of YYL (approximately 4.5 m) was instrumented with a buoy that measures environmental parameters every 10 min (see Figure 2) ; these data are accessible at the Global Ecological Lakes Observatory Network (GLEON) website. All data from the instrumented buoy and associated meteorological variables were downloaded from the GLEON publicly accessible database. The YYL buoy measured surface dissolved oxygen, wind speed, wind direction, water temperature profiles, and air temperature. A weather station approximately 1 km from the lake also measured rainfall, humidity, and soil temperature. High-resolution water temperature profiles collected from the buoy were used for the training and validation of the studied models. Also meteorological variables given in Table 1 were used to develop the soft computing techniques.

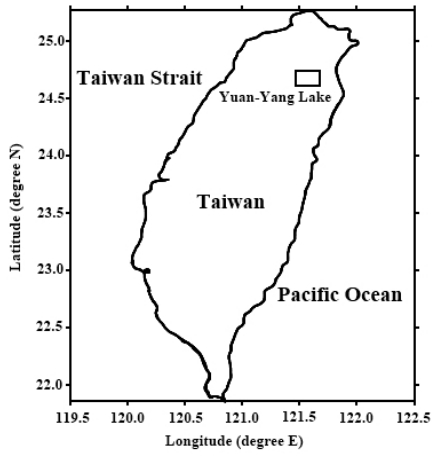

Figure 1. Location of Yuan-Yang Lake in Taiwan

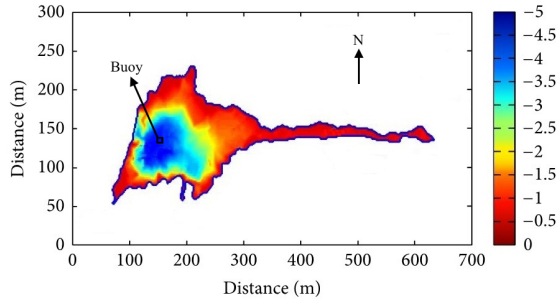

Figure 2. The bathymetry of the YYL (contours in m)

Table 1. Meteorological variables used to develop the models

Variables	Notation
Solar radiation ($W.m^{-2}$)	SR
Air pressure (h.pa)	AP
Relative humidity (%)	RH
Rainfall (mm)	RA
Wind speed ($m.s^{-1}$)	WS
Soil temperature (°C)	ST
Air temperature (°C)	AT
Water temperature at surface (°C)	WT0
Water temperature at 1 m below surface (°C)	WT1
Water temperature at 2 m below surface (°C)	WT2
Water temperature at 3 m below surface (°C)	WT3

Table 2 represents the hourly statistical parameters of the applied variables. In this table, the terms Xmean, X_{min}, X_{max}, S_x, C_v and C_{sx} denote the mean, minimum, maximum, standard deviation, coefficient of variation and skewness coefficient, respectively. As can be seen clearly, rainfall has the maximum skewness. Solar radiation and relative humidity also indicate a skewed distribution. Air pressure, soil temperature and air temperature show normal distributions because they have significantly low skewness values. Table 3 shows the correlations between the meteorological and hourly water temperature parameters. As can be seen from the table, the soil temperature has the highest correlations with hourly water temperatures in all depths. Air temperature also has higher correlations than the other variables. Time variation graphs of the meteorological variables used as inputs to the GEP, ANFIS and ANN models are illustrated in Figure 3.

Table 2. Hourly statistical parameters of the observed data.

Variables	Notation
Solar radiation (W.m^{-2})	SR
Air pressure (h.pa)	AP
Relative humidity (%)	RH
Rainfall (mm)	RA
Wind speed (m.s^{-1})	WS
Soil temperature (°C)	ST
Air temperature (°C)	AT
Water temperature at surface (°C)	WT0
Water temperature at 1 m below surface (°C)	WT1
Water temperature at 2 m below surface (°C)	WT2
Water temperature at 3 m below surface (°C)	WT3

Note: the terms X_{mean}, X_{min}, X_{max}, S_x, C_v *and* C_{sx} denote the mean, minimum, maximum, standard deviation, coefficient of variation and skewness, respectively.

Table 3. Correlations between meteorological and hourly water.temperature parameters.

	SR	AP	RH	RA	WS	ST	AT	WT0	WT1	WT2	WT
SR	1.00										
AP	0.43	1.00									
RH	-0.53	-0.11	1.00								
RA	-0.11	-0.14	0.07	1.00							
WS	0.61	0.14	-0.33	0.06	1.00						
ST	0.09	0.18	0.03	0.01	0.01	1.00					
AT	0.66	0.42	-0.43	-0.04	0.51	0.35	1.00				
WT0	0.11	0.45	-0.05	-0.03	0.13	0.59	0.66	1.00			
WT1	0.05	0.35	0.03	0.04	0.08	0.77	0.35	0.72	1.00		
WT2	0.06	0.14	0.06	0.04	0.03	0.92	0.26	0.44	0.70	1.00	
WT3	0.06	-0.11	0.05	0.05	0.03	0.83	0.22	0.25	0.44	0.90	1.00

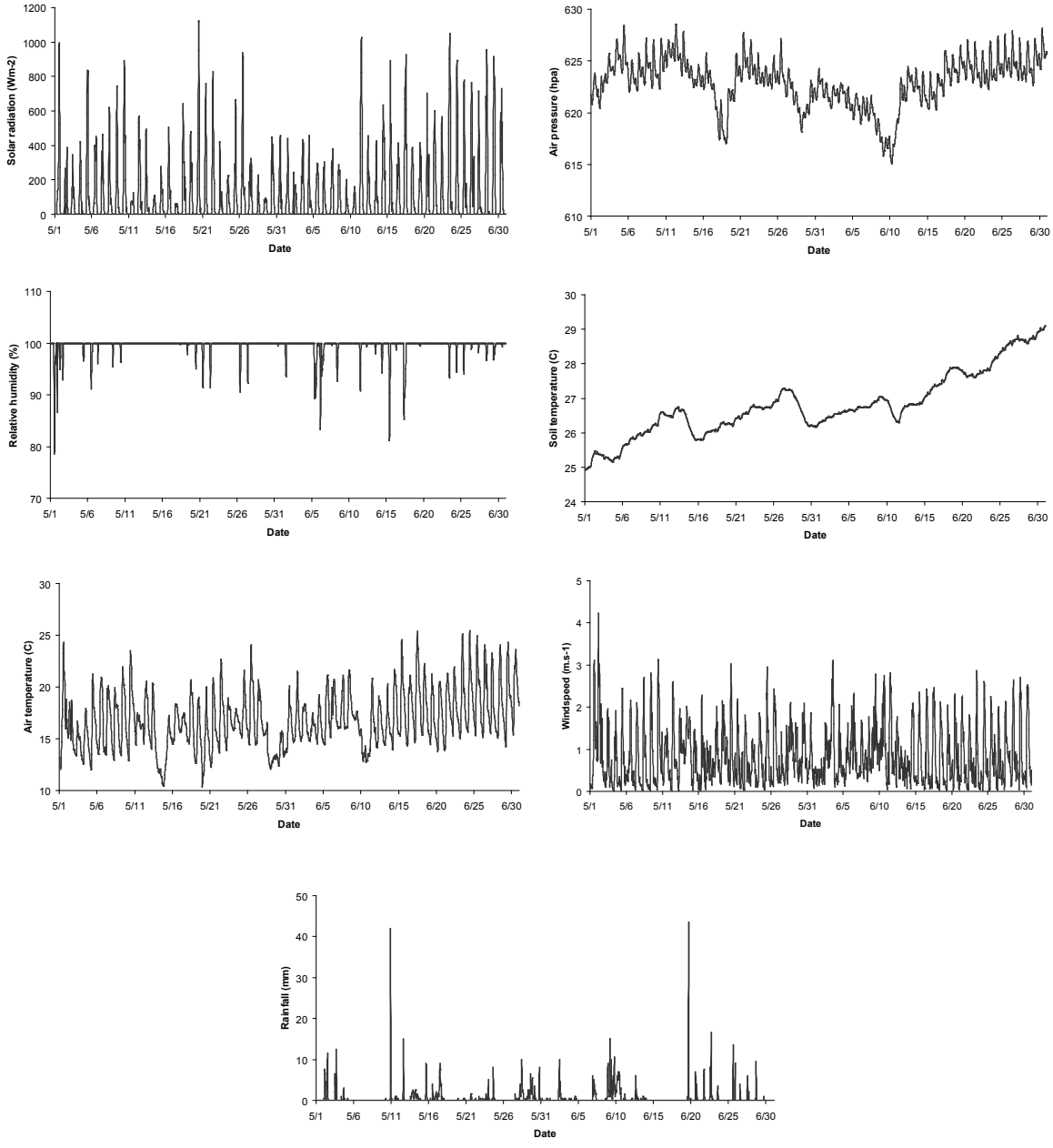

Figure 3. Time variation graphs of the meteorological conditions.

2.2 General overview of genetic programming

In this section, a brief overview of the GP and GEP is given. Detailed explanations of GP and GEP are provided by Koza (1992) and Ferreira (2006), respectively. GP was first proposed by Koza (1992). It is a generalization of genetic algorithms (GAs) (Goldberg, 1989). The fundamental difference between GA, GP, and GEP is due to the nature of the individuals. In the GA, the individuals are linear strings of fixed length (chromosomes). In the GP, the individuals are nonlinear entities of different sizes and shapes (parse trees), and in GEP the individuals are encoded as linear strings of fixed length (the genome or chromosomes), which are afterwards expressed as nonlinear entities of different sizes and shapes (Ferreira, 2001 a,b). GP is a search technique that allows the solution of problems by automatically generating algorithms and expressions. These expressions are coded or represented as a tree structure with its terminals (leaves) and nodes (functions). GP applies GAs to a "population" of programs, typically encoded as tree-structures. Trial programs are evaluated against a "fitness function" then the best solutions are selected for modification and re-evaluation. This modification-evaluation cycle is repeated until a "correct" program is produced.

There are five major preliminary steps for solving a problem by using GEP. These are the determination of (i) the set of terminals, (ii) the set of functions, (iii) the fitness measure, (iv) the values of the numerical parameters and qualitative variables for controlling the run, and (v) the criterion for designating a result and terminating a run (Koza, 1992). A GEP flowchart is presented in Figure 4.

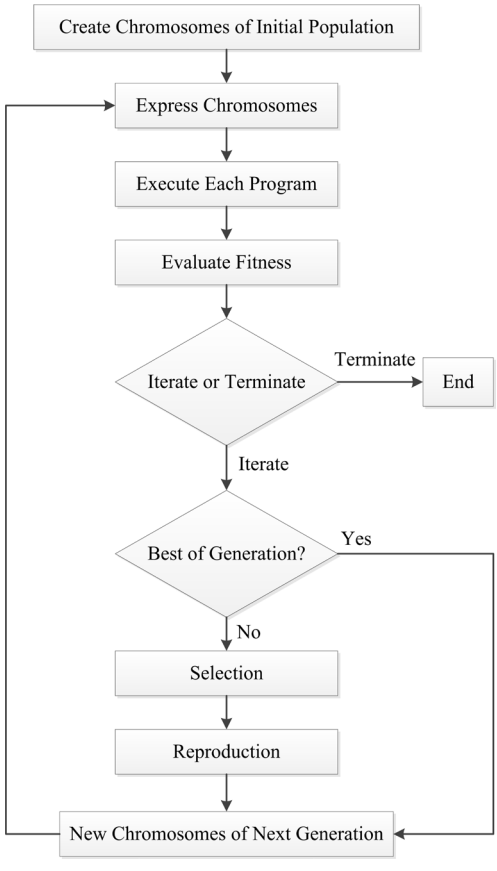

Figure 4. GEP flowchart

There are five major steps in preparing to use GEP of which the first is to choose the fitness function. The fitness of an individual program i for fitness case j is evaluated by Ferreira (2006) using:

$$If \quad E(i,j) \le p, \quad then \quad f_{(ij)} = 1; \quad else \quad f_{(ij)} = 0 \qquad (1)$$

Where p is the precision and E(i,j) is the error of an individual program i for fitness case j. For the absolute error, this is expressed by:

$$E(i,j) = \left| P_{(i,j)} - T_j \right| \qquad (2)$$

Where P(i,j) is the value predicted by the individual program i for fitness case j (out of n fitness cases) and T_j is the target value for fitness case j. Again for the absolute error, the fitness fi of an individual program i is expressed by:

$$f_i = \sum_{j=1}^{n} \left(R - \left| P_{(i,j)} - T_j \right| \right) \qquad (3)$$

Where R is the selection range. The second major step consists of choosing the set of terminals T and the set of functions F to create the chromosomes. For this study, the function set consists of 7 functions including four basic arithmetic operators, i.e., (+, -, ×, /) and some basic mathematical functions, i.e., (√, *Ln(x), exp*) selected among all the available functions in GEP. The function selection was based on simplicity and its relevance to the nature of the problem thus ensuring a simple and efficient final GEP model. The third major step is to choose the chromosomal architecture, i.e., the length of the head and the number of genes. Values of the length of the head, $h = 10$, and six genes per chromosome were employed based on the discussion in Ferreira (2001b). The fourth major step is to choose the linking function. In this study, the sub- programs were linked by addition on the basis of recommendations made by Ferreira (2001a) and findings of other studies (e.g. Guven and Aytek, 2009). Finally, the fifth major step is to choose the set of genetic operators that cause variation along with their rates. A combination of all genetic operators, i.e., mutation, transposition and recombination, was used for this purpose.

The parameters of the training of the GEP are given in Table 4.

Table 4. Parameters of the GEP model

Parameter	Value
Function set	+, -, ×, /, √, *Ln(x), exp*
Chromosomes	30
Head size	10
Number of Genes	6
Linking Function	Addition (+)
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

2.3. General overview of adaptive neuro-fuzzy inference system (ANFIS)

ANFIS (Jang, 1993), using a given input/output data set, constructs a fuzzy inference system (FIS) whose membership function parameters are tuned (adjusted) using either a back propagation algorithm alone or in combination with a least squares type of method. This adjustment allows fuzzy systems to learn from the data set. Figure 5 shows the ANFIS structure for a fuzzy inference system with two inputs.

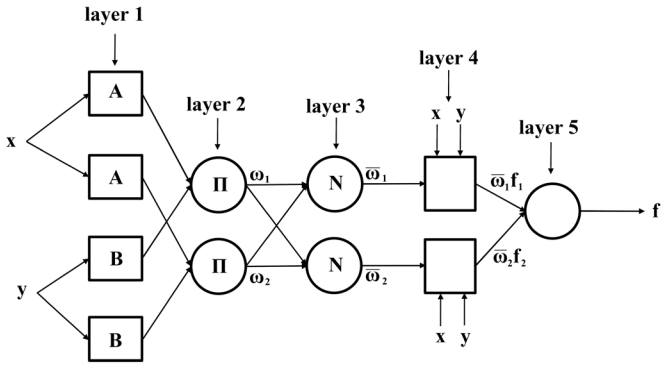

Figure 5. ANFIS structure

Firstly, the FIS type should be selected. In this regard, the Sugeno method was selected for the present study (Ozger, 2009; Ullah and Choudhury, 2013). To train a FIS, a training data set that contains the desired input/output data of the system to be modeled must be loaded. Before training, an initial FIS model structure must be specified:

(1) Hybrid algorithm was selected as the optimization method. The optimization methods train the membership function parameters to emulate the training data.

(2) The number of training Epochs and the training Error Tolerance were selected. These values are entered to set the stopping criteria for training. The training process stops whenever the maximum epoch number is reached or the training error goal is achieved. To validate the trained FIS, after loading the test data, the number of membership functions (MFs) and the type of input and output membership functions were selected. There are only two choices for the output membership function: constant and linear. This limitation of output membership function choices is because ANFIS only operates on Sugeno-type systems (Jang, 1993). The Built-in membership function composed of difference between two sigmoidal membership functions (dsigmf) and the constant membership function were selected as the input and output membership functions, respectively. The number of MFs assigned to each input was three. It should be noted that there is no basic rule for selecting the optimization method and determining the optimal number and type of MFs and they are usually considered by trial and error. Nevertheless, it should be taken into consideration that large numbers of MFs will increase the calculation time and efforts (Keskin et al., 2004). The values assigned to each parameter for the ANFIS model are given in Table 5.

Table 5. Parameters for the ANFIS model

Parameter	Value
FIS type	Sugeno
Optimization method	Hybrid algorithm
Number of MFs	3
Input MF type	Dsigmf
Output MF type	Constant

2.4. General overview of artificial neural networks (ANNs)

The Qnet neural network development system which is a complete solution for back propagation neural network modeling (Qnet, 2000) was used for the present study. Back propagation type neural networks process information in interconnecting processing elements termed nodes. These nodes are organized into groups termed layers. There are three distinct types of layers in a back propagation neural network: the input layer, the hidden layer(s) and the output layer. A network consists of one input layer, one or more hidden layers and one output layer. Connections exist between the nodes of adjacent layers to relay the output signals from one layer to the next. Information enters a network through the nodes of the input layer. The input layer nodes are unique in that their sole purpose is to distribute the input information to the next processing layer (i.e., the first hidden layer). Figure 6 shows the structure of the ANNs with one hidden layer including three nodes.

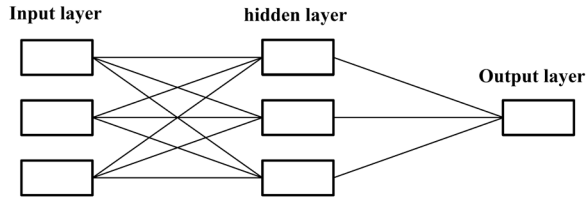

Figure 6. The structure of the artificial neural networks

To design the neural network, the modeler must specify the layer and node quantities, transfer functions and network connections. Including the input layer, one hidden layer and the output layer, the specified number of the layers for the network was three. The number of nodes in the input layer will be equal to the number of input data values in the model. According to table 1, It was specified seven for the prediction of hourly water temperature for the present study. The number of output nodes for the network must correspond to the number of outputs in the network. It was one because we only had one output. Choosing the number of hidden layers and the number of hidden nodes in each layer is not so trivial. The construction of the hidden processing structure of the network is arbitrary. Generally, it is best to start with simple network designs that use relatively few hidden layers and processing nodes. However, in practice, it is usually better to employ multiple hidden layers for solving complex problems. A single hidden layer including three single nodes was specified herein to avoid an unnecessary large and complex model. The sigmoid function is Qnet's default transfer function and it is the most widely used function for back propagation neural networks. Another network design consideration concerns how to control the network's connections. While the connection editor gives the modeler almost unlimited flexibility in designing a network, the fact is that the vast majority of designs work best fully connected. Qnet's connection editor is best suited for highly advanced models that require groups of input data to be processed through separate network pathways. The default fully connected configuration was used for the present study. The values assigned to each parameter for the ANN model are given in Table 6. As it is explained above, Qnet has a high and intelligent ability to simulate complex networks easily while preparing a code for these networks could be excruciating. For practical problems, using an easy method, which is usable for different cases, is more acceptable than sophisticated methods. In summary, Qnet is professional user friendly software and it has been used for simulating different complex problems (Kuo et al. 2004, Yang et al. 2009) and that is why it was used in this research.

Table 6. Parameters for the ANN model

Parameter	Value
Transfer function	Sigmoid
Number of layers	3
Number of hidden layer(s)	1
Number of hidden node(s)	3
Maximum iterations	10000

The performance of three soft computing techniques, namely GEP, ANFIS and ANN to predict the hourly water temperature at YYL at various measured depths was compared. Hourly water temperature data from May 1 to June 11, 2008 were taken as the training data set, while the measured

data from June 12 to 30, 2008 served as the validation data set. Furthermore, time-series meteorological conditions served as inputs for the studied models for data from May 1 to June 30, 2008 (Figure 3).

2.5. Evaluation parameters

Several parameters can be considered for the evaluation of the model-predicted values of hourly water temperature in the YYL Lake. In this study, root mean squared error (RMSE), mean absolute error (MAE) and correlation coefficient (R) were used as the evaluation criteria and they can be computed as follows (Liu and Chen, 2012):

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(T_p(i)-T_o(i)\right)^2} \quad (4)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|T_p(i)-T_o(i)\right| \quad (5)$$

$$R = \frac{(1/N)\sum_{i=1}^{n}\left(T_p(i)-\overline{T}_p\right)\left(T_o(i)-\overline{T}_o\right)}{\sqrt{(1/N)\sum_{i=1}^{n}\left(T_p(i)-\overline{T}_p\right)^2}\times\sqrt{(1/N)\sum_{i=1}^{n}\left(T_o(i)-\overline{T}_o\right)^2}} \quad (6)$$

where $T_P(i)$ and To(i) represent the model-predicted and the observed water temperature, respectively, and n is the number of observations.

3. Results and discussion

A comprehensive comparison was made to compare the accuracy of GEP, ANFIS and ANN methods for the prediction of hourly water temperature in different depths in the YYL. RMSE, MAE and R values of each model for the prediction of hourly water temperature at different depths are shown in Table 7. It is clear from the table that in the case of surface water temperature, the GEP model has the lowest RMSE (1.49 °C), MAE (1.21 °C) and the highest R (0.73) values. The ANN model seems to be the second best from the RMSE and MAE viewpoints. Also, in the case of 1 meter depth, the ANN model shows better accuracy than the GEP and ANFIS models. ANFIS model is ranked as the second best. Despite the error (RMSE value) is smaller in ANN than GEP, there is an upper bound in scatter plot of ANN (see figure 7.e) that imposes a constant value, which is not suitable for predictive purposes. Furthermore, at the 2 meter depth below water surface, the GEP model has the lowest RMSE (0.35 °C) and MAE (0.24 °C) values. Also, the ANN model has a better accuracy than the ANFIS. Finally, at the 3 meter depth below water surface, the superiority of the GEP model over the other models is clearly seen from Table 7. The ANN model seems to be the second best from the RMSE, MAE and R viewpoints.

Figure 7 shows the observed (x-axis) and predicted (y-axis) hourly water temperature values in different depths and in validation period in the form of scatter plots. GEP model seems to be better than the other soft computing techniques. In the case of surface water temperature, the estimates of GEP model seem to be closer to the exact line than those of the ANN and ANFIS models. The ANFIS model's estimates are also less scattered than the ANN. In the case of 1 meter depth, although the ANN model has the least scattered estimates, but there is an upper bound in scatter plot of ANN that imposes a constant value, which will make additional problems in future predictions. So, GEP model having less scattered estimations seems to be the best.

Table 7. Performance assessment of different models for predicting hourly water temperature at different depths

Depth	Model	Statistical parameters		
		RMSE (°C)	MAE (°C)	R
0	GEP	1.49	1.21	0.73
	ANN	1.64	1.32	0.64
	ANFIS	1.77	1.34	0.63
1	GEP	0.68	0.57	0.64
	ANN	0.44	0.35	0.80
	ANFIS	0.64	0.51	0.41
2	GEP	0.35	0.24	0.82
	ANN	0.40	0.31	0.84
	ANFIS	0.81	0.67	0.59
3	GEP	0.32	0.25	0.78
	ANN	0.55	0.44	0.75
	ANFIS	0.68	0.58	0.52

At the 2 meter depth below water surface, the estimates of the GEP model closer to the exact line than those of the other models. The ANN and ANFIS models underestimate all the high values (>13.5 °C). The ANN model has a better accuracy than the ANFIS because ANFIS model significantly underestimates high values (>13 °C). It is clear from the figure, at the 3 meter depth below water surface, the estimates of the GEP model are closer to the ideal line than those of the other models. The ANN model's estimates seem to be less scattered than the ANFIS model.

It should be noted that actual physical processes of the lakes such as vertical mixing, which controls the vertical distribution of temperature, have undeniable roles in changing water temperatures at different depths. Vertical mixing, in small, seasonally stratified lakes such as YYL, might result in partial or complete mixing of the water column, depending on the thermal conditions of the lake and the strength of meteorological forces driving the mixing process. Due to the fact that the purpose of the current research is statistical investigation of the effects of metrological parameters in predicting hourly water temperatures of YYL, the roles of physical phenomena including vertical mixing and phase lag between the sub-daily variations of the variables, affecting the heat flux, have been ignored. So, ignoring the physical factors might be one of the sources, which increased the error parameters of the predictions.

Figure 8 shows the observed and predicted values as time series plot in the validation period. From the figure, it can be said that the GEP model is generally more successful than the ANN and ANFIS model especially for the high water temperature values. One of the advantages of GEP in comparison to the other theories is its ability in producing analytical formula for determination of output parameters. Table 8 summarizes the GEP mathematical equations for the prediction of hourly water temperature in different depths. From the table, the *AP* seems to be not effective variable on the water temperature in the case of 2- and 3-meter depth below water surface. It can be said that the *AP* has a more effect on WT0 than the WT1. Decreasing *AP* effect by increasing depth is clearly seen from the table.

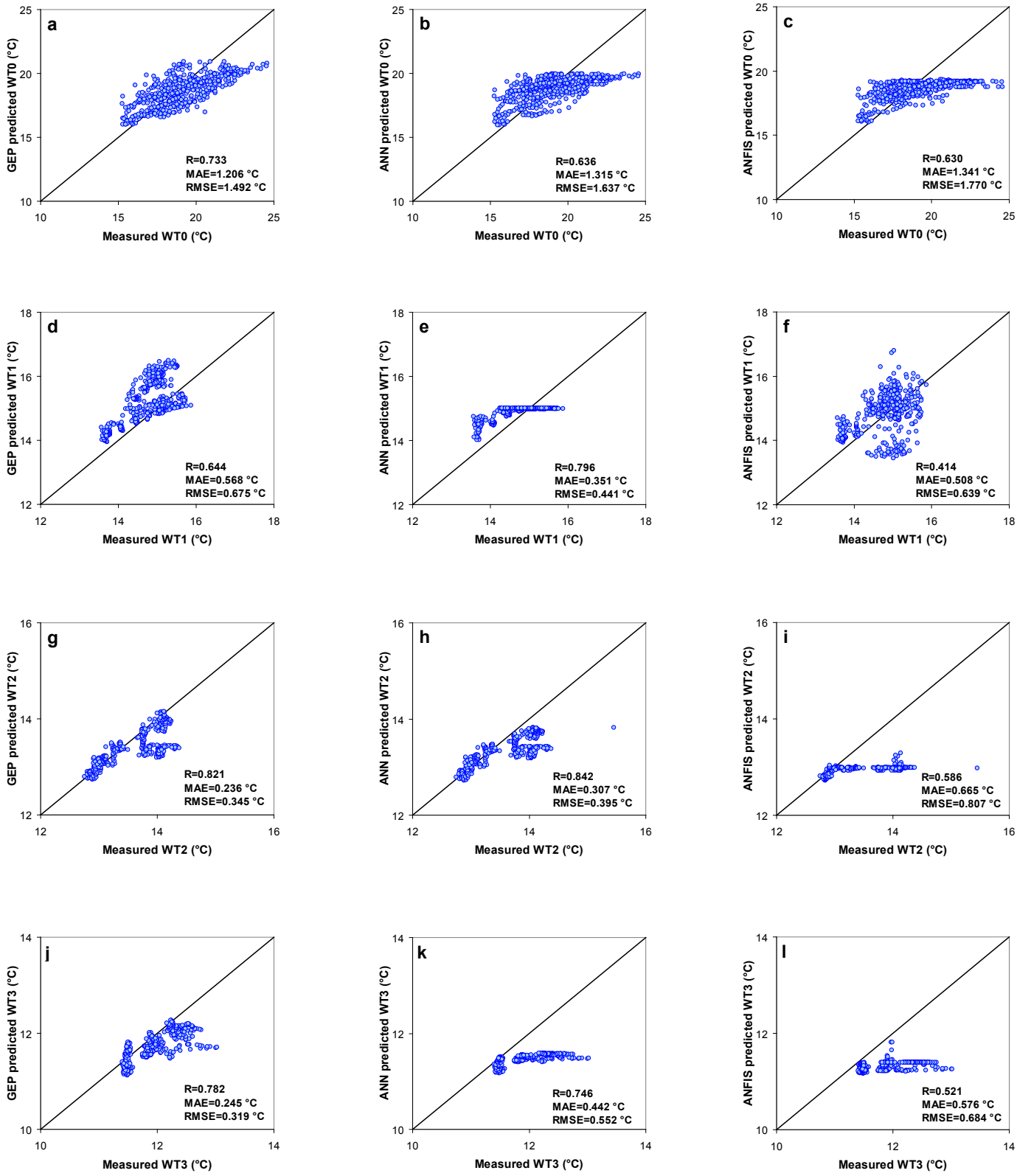

Figure 7. Scatter plots of observed (x-axis) and predicted values (y-axis) of hourly water temperature in different depths.

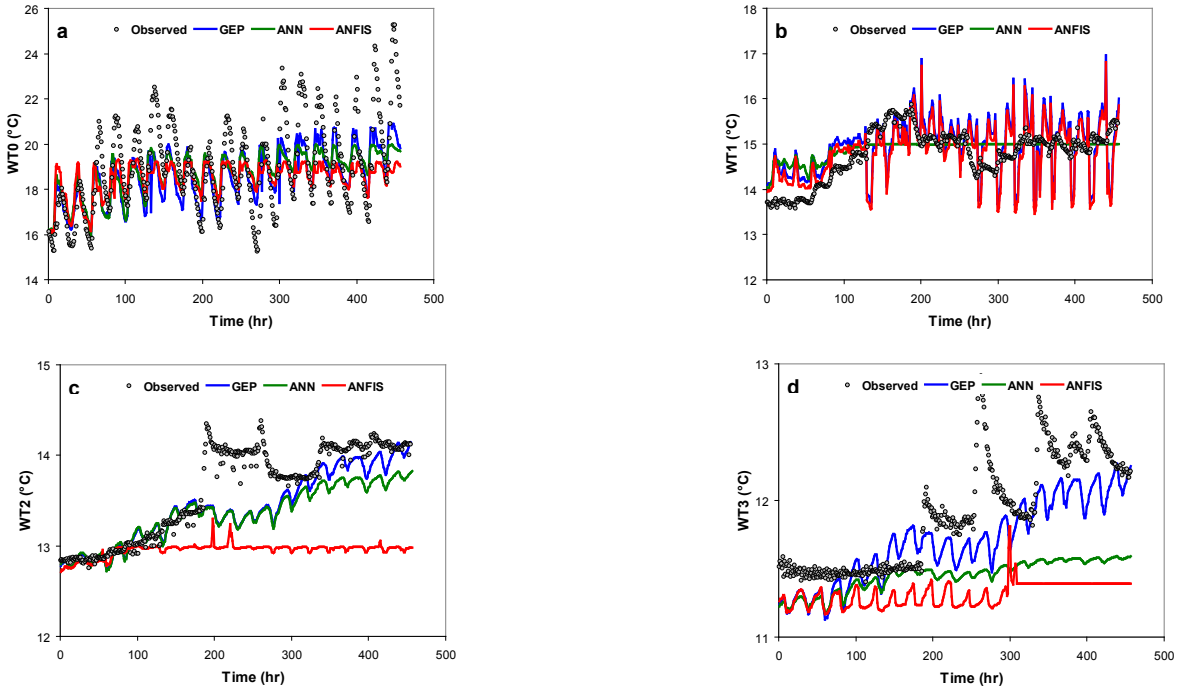

Figure 8. Time series plots of observed and predicted values of hourly water temperature in different depths.

Table 8. Mathematical expressions of GEP model.

Predicted	Mathematical Expression of the model
WT0	$-21.906-\frac{1.57009\times Exp[0.798065AT]}{AP^3}+3.30075\left(AT(3.30075+AT)-3\times ST\right)^{1/4}-\frac{6.77805(9.72513-AT)+ST}{2.27471+AT^2}+ST+Exp[-AT]\times(-AT\times ST-3.64459ST(ST-AP))+\frac{4.83991ST+ST(AT+ST)}{6.62537-AP+ST}+\frac{ST}{Ln[Exp[ST]-AP]-Ln[2AP]}$
WT1	$-15.6273+0.0253065AT+2.8006\sqrt{ST}-\frac{AT\times ST}{2AP}+\frac{AT\times ST}{2AT(AP+AT)-3AP}-\frac{4.42575ST(ST-8.32956)}{AT-8.32956ST}+Exp[-AT]\left(AT\times ST^2-9.59219\right)\times Ln[2.31732+AT]+0.00227366AP\times Ln[ST]$
WT2	$-3.42694-8.38251(1.03943-0.000228836(AT-ST))-\sqrt{ST}+0.97287ST+\frac{132.822-AT+4.88901\times Ln[AT]}{ST}+Exp[-0.286869ST]\left(9.95398AT+ST^2+\sqrt{AT\times ST}\right)+(AT-7.30499)\times Exp[3.63193-AT]\times(AT+ST)\times Ln[AT]+Ln\left[\frac{8.28603}{ST}\right]$
WT3	$-15.2691-Exp\left[-\frac{2.27249(-7.40158+AT)\times AT^2}{ST^2}\right]+\sqrt{\frac{5.5354+0.548691ST}{1+5.5354AT}}+\frac{6.48978(5.4389+AT+ST)}{ST^2}+\sqrt{\frac{AT}{(-1.49371-AT)\times(8.11706-0.669473AT)ST}}+ST+\sqrt{8.66666ST-0.807352AT}+\sqrt{5.13181+ST}$

As a conclusion, output results showed that GEP provided reasonable and moderately accurate trends for the prediction of hourly water temperature in different depths.

4. Conclusion

In the present study, the performance of some soft computing techniques, namely gene expression programming, adaptive neuro fuzzy inference system and artificial neural network to predict hourly water temperatures at different layers of the Yuan-Yang Lake in north-central Taiwan has been compared. A time-series set of data for the hourly water temperature at different measured depths from May 1 to June 11, 2008 was taken as training dataset, while the data measured from June 12 to June 30, 2008 were served as a validation dataset for the studied models. The relative performances of these models were comprehensively evaluated using various statistical indices including RMSE, MAE and R coefficients. Results showed that the GEP had the best performances in predicting hourly water temperatures at the surface and both 2 and 3 meter depths below water surface, whereas, a different trend was seen for the 1 meter depth below water surface. In this depth, in spite of smaller error in ANN than GEP, there is an upper bound in scatter plot of ANN that imposes a constant value, which is not appropriate for predictive purposes. Conclusively, results obtained from this study showed that GEP can provide reasonable trends for the prediction of hourly water temperature in different depths especially in shallow waters.

Acknowledgments

Sincere thanks are given to the two anonymous reviewers for their invaluable and constructive comments and suggestions for improving the paper quality.

References

Aytac, A. and Kisi, O., (2008). A genetic programming approach to suspended sediment modelling. Journal of Hydrology, 351, 288-298.

Banzhaf, W., Nordin, P., Keller, R.E. and Francone, F.D., (1998). Genetic programming. Morgan Kaufmann, San Francisco, CA.

Benyahya, L., Caissie, D., St-Hilaire, A., Ouarda, T.B.M.J. and Bobee, B., (2007). A review of statistical water temperature models. Canadian Water Resource Journal, 32(3), 179-192.

Chen, S.H. and Wu, J.T., (1999). Paleolimnological environment indicated by the diatom and pollen assemblages in an alpine lake of Taiwan. Journal of Paleolimnology, 22, 149-158.

Fallah-Mehdipour, E, Borzorg Haddad, O. and Marino, M.A., (2013). Prediction and simulation of monthly groundwater levels by genetic programming. Journal of Hydro-environment Research, 7(4), 253-260.

Fang, X. and Stefan, H.G., (1996). Long-term lake water temperature and ice cover simulations/measurements. Cold Regions Science and Technology, 24, 289-304.

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

Ferreira, C., (2001b). Gene expression programming, A new adaptive algorithm for solving problems. Complex Systems, 13 (2), 87.

Ferreira, C., (2006). Gene expression programming: mathematical modeling by an artificial intelligence. Springer, Berlin 2006, p. 478.

GLEON, Global Lake Ecological Observatory Network. http://gleon.org/.

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

Guven, A. and Aytek, A., (2009). New approach for stageedischarge relationship: geneexpression programming. Journal of Hydrologic Engineering, 14, 812-820.

Hondzo, M. and Stefan, H.G., (1996). Long-term lake water quality predictors. Water research, 30(12), 2835-2852.

Jang, R., (1993). ANFIS: adaptive network-based fuzzy inference system. IEEE Transactions on Systems, Man, and Cybernetics, 23, 665-685.

Keskin, M.E., Terzi, O. and Taylan, D., (2004). Fuzzy logic model approaches to daily pan evaporation estimation in Western Turkey. Hydrological Sciences Journal, 49 (6), 1001–1010.

Kisi, O., (2006). Daily pan evaporation modeling using a neuro-fuzzy computing technique. Journal of hydrology, 329, 636-646.

Kisi, O., Shiri, J. and Nikoofar, B., (2012). Forecasting daily lake levels using artificial intelligence approaches. Computers & Geosciences, 41, 169-180.

Koza, J.R., (1992). Genetic programming, on the programming of computers by means of natural selection. MIT Press, Cambridge, MA, ISBN 0-262-11170-5.

Kuo, Y.M., Liu, C.W. and 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 Research, 38 (1), 148-158.

Lawrence, S.P., Hogeboom, K. and Le Core, H.L., (2002). A three-dimensional general circulation model of the surface layer of Lake Baikal. Hydrobiologia, 487 (1), 95-110.

Lee, H.S., Yamashita, T. and Haggag, M., (2009). Modelling hydrodynamic in Yachiyo Lake using a non-hydrostatic general circulation model with spatially and temporally varying meteorological conditions. Hydrological Processes, 23 (14), 1973-1987.

Liu, W.C. and Chen, W.B., (2012). Prediction of water temperature in a subtropical subalpine lake using an artificial neural network and three-dimensional circulation models. Computers & Geosciences, 45, 13-25.

Liu, W.C., Yu, H.L. and Chung, C.E., (2011). Assessment of Water Quality in a Subtropical Alpine Lake Using Multivariate Statistical Techniques and Geostatistical Mapping: A Case Study. International Journal of Environmental Research and Public Health, 8, 1126-1140.

Mellit, A., Benghanen, M. and Kalogirou, S.A., (2006). An adaptive wavelet network model for forecasting daily total solar radiation. Applied Energy, 83, 705-722.

Moghaddamnia, A., Ghafari Gousheh, M., Piri, J., Amin, S. and Han, D., (2009). Evaporation estimation using artificial neural networks and adaptive neuro-fuzzy inference system techniques. Advances in Water Resources, 32, 88-97.

Ozger, M., (2009). Comparison of fuzzy inference systems for streamflow prediction. Hydrological Sciences Journal, 54(2), 261-273.

Piccolroaz, S., Toffolon, M. and Majone, B., (2013). A simple lumped model to convert air temperature into surface water temperature in lakes. Hydrology and Earth System Sciences, 17, 3323-3338.

Qnet, (2000). Qnet 2000 Neural Network Modelling for Windows 95/98/NT, QnetToll User's Guide and Datapro User's Guide. Vesta Services, Incorporated, USA.

Samadianfard, S., (2012). Gene expression programming analysis of implicit Colebrook-White equation in turbulent flow friction factor calculation. Journal of Petroleum Science and Engineering, 92-93, 48-55.

Samadianfard, S., Sadraddini, A.A., Nazemi, A.H., Provenzano, G. and Kisi, O., (2012). Estimating soil wetting patterns for drip irrigation using genetic programming. Spanish Journal of Agricultural Research, 10(4), 1155-1166.

Schwab, D.J., Beletsky, D., DePinto, J. and Dolan, D.M., (2009). A hydrodynamic approach to modeling phosphorus distribution in Lake Erie. Journal of Great Lakes Research, 35 (1), 50-60.

Sharma, S., Walker, S.C. and Jackson, D.A., (2008). Empirical modelling of lake water-temperature relationships: a comparison of approaches. Freshwater Biology, 53, 897-911.

Shiri, J. and Kisi, O., (2011). Comparison of genetic programming with neuro-fuzzy systems for predicting short-term water table depth fluctuations. Computers & Geosciences, 37 (10), 1692-1701.

Shiri, J., Dierickx, W., Pour-Ali Baba, A., Nemati, S. and Ghorbani, M.A., (2011). Estimating daily pan evaporation from climatic data of the state of Illinois, USA using adaptive neuro-fuzzy inference system and artificial neural network. Hydrological research, 42 (6), 491-502.

Thiery, W., Stepanenko, V.M., Fang, X., Johnk, K.D., Li, Z., Martynov, A., Perroud M., Subin, Z.M., Darchambeau, F., Mironov, D. And Lipzig, N.P.M.V, (2014). LakeMIP Kivu: evaluating the representation of a large, deep tropical lake by a set of one-dimensional lake models. Tellus A, 66, 21390, http://dx.doi.org/10.3402/tellusa.v66.21390

Toffolon, M., Piccolroaz, S., Majone, B., Soja, A.M., Peeters, F., Schmid, M. and Wuest, A., (2014). Prediction of surface temperature in lakes with different morphology using air temperature. Limnology and Oceanography, 59 (6), 2185-2202.

Trumpickas, J., Shuter, B.J. and Minns, C.K., (2009). Forecasting impacts of climate change on Great Lakes surface water temperatures. Journal of Great Lakes Research, 35, 454-463.

Ullah, N. And Choudhury, M., (2013). Flood Flow Modeling in a River System Using Adaptive Neuro-Fuzzy Inference System. Environmental Management and Sustainable Development, 2(2), 54-68.

Webb, B.W., Hannah, D.M., Moore, R.D., Brown, L.E. and Nobilis, F., (2008). Recent advances in stream and river temperature research. Hydrological Processes, 22, 902-918.

Wu, J.T., Chang, S.C., Wang, Y.S., Wang, Y.F. and Hsu, M.K., (2001). Characteristics of the acidic environment of the Yuan yang Lake (Taiwan). Botanical Bulletin of Academia Sinica, 42, 17-22.

Yang, C.T., Marsooli, R. and Aalami, M.T., (2009). Evaluation of Total Load Sediment Transport Formulas Using ANN. International Journal of Sediment Research, 24(3), 274-286.

Zaier, I., Shu, C., Ouarda, T.B.M.J, Seidou, O. and Chebana, F., (2010). Estimation of ice thickness on lakes using artificial neural network ensembles. Journal of Hydrology, 383, 330-340.

Zhao, L., Yu, Z. and Acharya, K., (2011). Modeling water temperature of Lake Taihu. American Geophysical Union, Fall meeting.