Environ Earth Sci (2016) 75:139
DOI 10.1007/s12665-015-4905-6

ORIGINAL ARTICLE

Prediction of water quality index in free surface constructed wetlands

Reza Mohammadpour[1,2] · **Syafiq Shaharuddin**[2] · **Nor Azazi Zakaria**[2] · **Aminuddin Ab. Ghani**[2] · **Mohammadtaghi Vakili**[3] · **Ngai Weng Chan**[4]

Received: 6 March 2015 / Accepted: 6 August 2015 / Published online: 6 January 2016

Abstract Water quality and its effects on human life have become one of the major concerns in aquatic ecosystems. The water quality index (WQI) is defined as a parameter to interpret water-monitoring data and clarify the quality of water. In this study, the gene expression programming (GEP) and artificial neural networks (ANNs) were employed to predict WQI in free surface constructed wetlands. Seventeen points of a selected wetland were monitored twice a month over a period of 14 months, and an extensive data set was collected for 11 water quality variables (WQVs). A principal factor analysis (PFA) indicated that WQI was greatly affected by pH and SS, while temperature no has significant effect on the WQI in tropical areas. A sensitivity analysis was carried out to reduce the number of 11 WQVs in prediction of the WQI. Subsequently, five significant parameters, pH, suspended solid (SS), ammoniacal nitrogen (AN), dissolved oxygen (DO) and chemical oxygen demand were selected to develop a GEP and ANNs. The GEP was able to successfully predict the WQI with high accuracy ($R^2 = 0.983$ and MAE $= 0.295$). The statistical parameters indicate that, although the ANNs with $R^2 = 0.988$ and MAE $=$ 0.013 produced better results compared with GEP, the GEP-based formula is more useful for practical purposes. The GEP and ANNs are recommended as rapid and powerful WQI evaluation techniques to reduce substantial effort and time by optimizing the calculations.

Keywords Constructed wetland · Gene expression programming · Water quality index · Surface water · Principal factor analysis · Artificial neural networks

Introduction

Poor quality of surface water is a serious problem in the world which threatens human health, ecosystems and plants/animals life. Water quality (WQ) is, therefore, a main concern in water resource, environmental systems and ecosystem. It is a terminology used to describe the chemical, physical, and biological characteristics of water in connection with a set of standards (Liou et al. 2004). WQ assessment can be used to evaluate water properties in reference to natural quality and human health effects (Fernández et al. 2004). It can be assessed by measuring a broad range of variables to represent the water pollution level. Hence, a robust mathematical technique is required to combine the physico-chemical characterization of water into a single variable which describes the water quality. In view of this, a water quality index (WQI) was developed as a single number which uses a set of physico-chemical water variables to explain the water quality at a certain place and time (Zandbergen and Hall 1998).

✉ Reza Mohammadpour
reza564@gmail.com

Mohammadtaghi Vakili
farshid3601@gmail.com

Ngai Weng Chan
nwchan1@gmail.com

[1] Department of Civil Engineering, Estahban Branch, Islamic Azad University, Estahban, Iran

[2] River Engineering and Urban Drainage Research Centre (REDAC), Universiti Sains Malaysia, Engineering Campus, Seri Ampangan, 14300 Nibong Tebal, Penang, Malaysia

[3] School of Industrial Technology, Universiti Sains Malaysia, 11800 Penang, Malaysia

[4] School of Humanities, Universiti Sains Malaysia, 11800 Penang, Malaysia

WQI is a unit-less number which reflects the status of water quality in wetlands, lakes, streams, rivers, and reservoirs. The concept of WQI is based on the comparison of the water quality parameter with respective regulatory standards (Khan et al. 2003). There several equations for WQI in different countries such as the US, Canada, and Malaysia which are developed based on the standards of the US National Sanitation Foundation (Said et al. 2004). In 1974, the Department of Environment (DoE) Malaysia recommended an index to assess the quality of surface waters in Malaysia. Totally, six parameters were chosen as main water quality variables (WQVs) to develop WQI for surface water such as dissolved oxygen (DO), biochemical oxygen demand (BOD), chemical oxygen demand (COD), ammoniacal nitrogen (AN), suspended solid (SS) and pH (DoE 2005; Khuan et al. 2002; Norhayati et al. 1997). These variables should be converted into non-dimensional parameters by sub-index functions. The conventional method recommended by DoE requires long-lasting transformations to calculate sub-indices. In addition, the sub-indices required the inclusion of different equations, which need lengthy effort and time to estimate the final WQI. Therefore, estimation of such a WQI is cumbersome and can lead to occasional mistakes (Gazzaz et al. 2012), and robust techniques can be employed to solve these problems (Mohammadpour et al. 2013a). The gene expression programming (GEP) and artificial neural networks (ANNs) can be suggested as alternative techniques for estimation of WQI, as both employ the raw data instead of sub-indices.

In the last decade, the GEP and genetic programming (GP) have been successfully used in water resources modelling issues (Azamathulla et al. 2010; Zakaria et al. 2010; Azamathulla and Ghani 2011). Furthermore, these methods were recommended as significant tools in environmental and river engineering problems (Chen et al. 2008; Aras et al. 2007; Mohammadpour et al. 2013b; Ghani and Azamathulla 2014; Mohammadpour et al. 2015b). Vink and Schot (2002) developed GP for optimization of drinking water. The performance of the GP was compared with analytic solution of a series of hypothetical case studies. Hashmi et al. (2011) developed GEP for downscaling of watershed precipitation in Canada. Azamathulla (2012) applied GEP for prediction of scour depth at downstream of sills. Ni et al. (2012) evaluated water storage in wetlands using the GP technique. The result indicated that the GP method can be used for estimation of water fluctuation in the wetlands. Azamathulla and Ahmad (2012) used GEP approach to predict the transverse mixing coefficient in open channel flows. Xu and Qin (2013) solved the problem related to agricultural water quality management by using a combination of GA and fuzzy simulation. Orouji et al. (2013) investigated the performance of GP and ANFIS-GP to estimate water quality parameters. Different combinations of data set were employed in their study, and the results showed that GP is superior to ANFIS for prediction of water quality parameters. Zaman Zad Ghavidel and Montaseri (2014) employed GEP and other artificial intelligence approaches to predict total dissolved solids in river basin. A comparison between all selected approaches emphasized the superiority of GEP over the other intelligent methods.

Recently, a lot of studies have been reported in literature regarding the application of ANNs in different fields such as water quality, wastewater treatment and other water resources problems (Singh et al. 2009; Civelekoglu et al. 2009; Verma and Singh 2013; Mohammadpour et al. 2013c, 2014b, 2015a). In the area of river management, ANNs was used to simplify and speed up the calculation of water quality index (Khuan et al. 2002; Juahir et al. 2004; Gazzaz et al. 2012). The ANNs was employed to determine water quality parameters and simulate wetlands processes (Wang et al. 2012; Kashefi Alasl et al. 2012; Li et al. 2013; Song et al. 2013). Schmid and Koskiaho (2006) developed ANNs to model concentrations of dissolved oxygen in free surface wetlands. They have also used ANNs to estimate the relative influence of flow rate and wind shear on near bottom oxygen saturation. The results indicated that ANNs was able to produce estimates of convective oxygen transport. Dadaser-Celik and Cengiz (2013) simulated the water level in wetlands using ANNs. It was found that the ANN method can successfully be employed to predict water levels in wetlands. Karthikeyan et al. (2013) developed ANNs to predict ground water levels in the upland of a tropical coastal wetland with fairly accurate results.

The main objective of this research is to reduce substantial time and effort for calculation of WQI in the free surface constructed wetlands. The GEP and ANNs were employed as the robust techniques to determine WQI. Seventeen points in a wetland were monitored twice a month over a period of 14 months and an extensive data set was collected for 11 water quality variables. A principal factor analysis (PFA) was used to determine and interpret the correlation between variables. To develop GEP and ANN, the significant variables were chosen using sensitivity analysis. Finally, accuracy of each method was evaluated using a comparison between the obtained results.

Materials and methods

Study area

In this research, the free surface constructed wetland (FSCW) in the Universiti Sains Malaysia (USM) in Penang (Malaysia) was chosen as a case study. The landscape area is about 320 hectares, and it is covered by oil palm plantation (Shaharuddin et al. 2013; Mohammadpour et al.

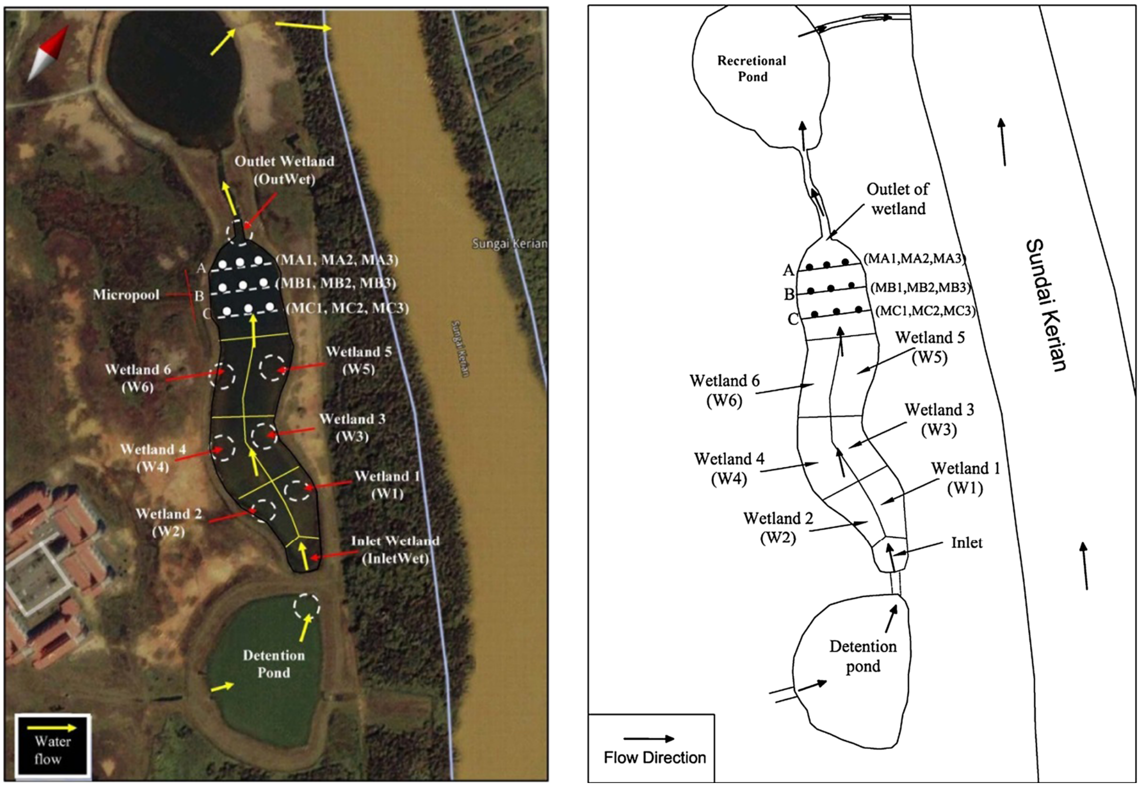

Fig. 1 Seventeen sample points in the constructed wetland of USM

2014a). The wetland is located at latitude 5° 9′ 7.8294″ North and longitude 100° 29′ 53.1672″ East. The FSCW was designed based on the Stormwater Management Manual for Malaysia (Zakaria et al. 2003). Seventeen sampling points with different plant species and water depths were chosen to monitor the water quality. These points include the inlet, six stations in the macrophyte area (W1–W6), nine points in micropool (MA1–MC3), and the outlet (Fig. 1). These points have been chosen in such a way that covers all range of plants and the water depths in the wetland (Table 1).

The data were collected twice a month over a period of 14 months (from Oct. 2010 to Dec. 2011). Totally, 11 water quality variables (WQVs) were collected in the wetland, including dissolved oxygen (DO), pH, temperature, conductivity, suspended solid (SS), nitrite, nitrate, ammoniacal nitrogen (AN), chemical oxygen demand (COD), biochemical oxygen demand (BOD), and phosphate. Table 2 indicates statistical parameters of the collected data.

The local water quality index

As mentioned earlier, to determine WQI of water surface, the DoE (2005) recommended six variables such as, DO, BOD, COD, AN, SS and pH. These variables should be converted into non-dimensional variables using sub-index functions (SI). Table 3 shows the required functions which can be used to estimate sub-indices. In this table, X is the concentration parameter in terms of mg/L, except for pH and DO. For DO, the X refers to percentage of saturation and for pH it refers to the pH value. Finally, the WQI can be calculated using the following equation (DoE 2005; Khuan et al. 2002):

$$\begin{aligned}\mathrm{WQI} = {} & 0.22\,\mathrm{SI}_{\mathrm{DO}} + 0.19\,\mathrm{SI}_{\mathrm{BOD}} + 0.16\,\mathrm{SI}_{\mathrm{COD}} + 0.15\,\mathrm{SI}_{\mathrm{AN}} \\ & + 0.16\,\mathrm{SI}_{\mathrm{SS}} + 0.12\,\mathrm{SI}_{\mathrm{pH}} \qquad (5)\end{aligned}$$

where SI stands for sub-index.

WQI is a unit-less number which varies between 0 and 100, where a high value of WQI represents high (good) water quality and a low value of WQI represents low (poor) water quality. Based on this index, the water quality can be classified into five classes. Table 4 shows the water quality classes suggested by the DoE.

Principal factor analysis

In this study, principal factor analysis (PFA) has been employed to determine the correlation between variables and WQI. Furthermore, insignificant variables can be clarified in this analysis. To avoid the effect of strong variables with high values on PFA, the z scale

Table 1 Plant species and the water depth in the USM wetland

Site	Wetland plant species	Water depth (m)
Wetland 1	Dominant: *Hanguana malayana, Lepironia articulata*	0.25–0.3
Wetland 2	Dominant: *Hanguana malayana, Typha angustifolia* Less dominant*: Scirpus grossus*	0.27–0.32
Wetland 3	Dominant: *Lepironia articulata, Eleocharis variegata* Less Dominant*: Eriocaulon longifolium*	0.51–0.62
Wetland 4	Dominant: *Hanguana malayana, Lepironia articulata, Eleocharis variegata*	0.47–0.54
Wetland 5	Dominant: *Lepironia articulata*	0.51–0.64
Wetland 6	Dominant: *Lepironia articulata* Less dominant: *Typha angustifolia*	0.31–0.54
Micropool (MA, MB and MC)	Without plant	2.48–2.54

Table 2 Descriptive statistics of wetland parameters

WQV	Min	Max	Mean	SD
Temperature (°C)	27.30	35.15	31.12	1.52
pH	6.11	9.19	7.73	0.69
DO (mg/l)	4.96	11.06	8.24	0.87
Conductivity (µs/cm)	94.00	206.00	136.59	25.49
Nitrite (mg/l)	0.00	0.06	0.02	0.01
Nitrate (mg/l)	0.20	4.90	2.15	0.91
Phosphate (mg/l)	0.11	0.58	0.25	0.10
AN (mg/l)	0.10	0.47	0.22	0.07
BOD (mg/l)	1.32	4.12	2.53	0.47
COD (mg/l)	9.00	44.00	21.96	5.60
SS (mg/l)	2.00	39.00	17.75	7.79

Number of data = 442

transformation was used to standardize the collected data set. The KMO (Kaiser–Meyer–Olkin) and Barlett's tests of sphericity were employed to evaluate sampling size adequacy and verification of PFA, respectively.

The PFA was applied to a matrix with the dimension of 442 objects and twelve variables (a WQI and 11 WQVs). The KMO test produces a value equal to 0.822 which indicates the number of collected data is adequate. In addition, the Bartlett's test of sphericity with approximate Chi Square of 3792.804 ($\rho = 0.000 < 0.05$ and $df = 66$) reveals that the principal factor analysis can be used to explain the WQVs.

As shown in Table 5, three factors were extracted by the PFA with eigenvalue bigger and equal to one. To estimate the effect of each variable in the PFA, the Varimax rotation was employed to determine values of rotated factor loadings. However, a factor loading less than 0.4 was recognized as a weak factor (Lambrakis et al. 2004; Gazzaz et al. 2012).The strong and moderate factors (bigger than 0.40) are shown in bold in Table 5.

Eight variables including the WQI are loaded on the first factor with a variation of 49 %. The WQVs and their factor loadings are SS (0.85), nitrate (0.84), phosphate (0.84), AN (0.81), nitrite (0.81), BOD (0.79), COD (0.77), and WQI (−0.62). A negative factor loading for WQI indicates that the WQI increases with decreasing values in the mentioned variables in the first factor. Among all variables, SS has higher correlation with WQI. Consequently, it is a significant parameter on WQI.

Suspended solids (SS) is solid materials, including organic and inorganic, that are suspended in the water. High concentrations of SS increases the amount of light which can be absorbed by the water. In this condition, the water becomes warmer and loses its ability to hold oxygen. Aquatic plants also receive less light and less oxygen that is produced by photosynthesis. The combination of less light, warmer water and less oxygen decreases the water quality.

The loaded variables on the second factor are pH (0.90), conductivity (0.82), DO (0.47) and WQI (−0.69). The high correlation between pH (0.90) and WQI (−0.69) illustrates that pH is another significant variable. In addition, negative coefficient indicates that WQI decreases with increasing pH in range of 6.11 and 9.19 (Table 2).

The third factor received the highest factor loading from DO (0.68) and temperature (0.82). The WQI is loaded on this factor with very low value (−0.09). In the second extracted factor, it was observed that DO has a correlation with WQI indicating that temperature alone has no effect on the WQI. It may be due to low variation of wetland temperature in tropical areas with minimum, maximum and average value of 27.3, 35.15 and 31.12, respectively, (Table 2). Consequently, temperature is an insignificant variable for wetlands which are located in tropical areas.

Table 3 The sub-index equation for WQI in Malaysia (DoE 2005)

WQVs	Value[a]	Sub-index
DO (%saturation)	$X \le 8$	$\mathrm{SI}_{\mathrm{DO}} = 0$
	$8 < X < 92$	$\mathrm{SI}_{\mathrm{DO}} = -0.395 + 0.03\,X^2 - 0.0002\,X^3$
	$X \ge 92$	$\mathrm{SI}_{\mathrm{DO}} = 100$
BOD	$X \le 5$	$\mathrm{SI}_{\mathrm{BOD}} = 100.4 - 4.23\,X$
	$X > 5$	$\mathrm{SI}_{\mathrm{BOD}} = (108\,e^{-0.055X}) - 0.1\,X$
COD	$X \le 20$	$\mathrm{SI}_{\mathrm{COD}} = 99.1 - 1.33\,X$
	$X > 20$	$\mathrm{SI}_{\mathrm{COD}} = (103\,e^{-0.0157X}) - 0.04X$
AN	$X \le 0.3$	$\mathrm{SI}_{\mathrm{AN}} = 100.5 - 105\,X$
	$0.3 < X < 4$	$\mathrm{SI}_{\mathrm{AN}} = (94\,e^{-0.573X}) - 5\lvert X - 2\rvert$
	$X \ge 4$	$\mathrm{SI}_{\mathrm{AN}} = 0$
SS	$X \le 100$	$\mathrm{SI}_{\mathrm{SS}} = (97.5\,e^{-0.00676X}) + 0.05X$
	$100 < X \le 1000$	$\mathrm{SI}_{\mathrm{SS}} = (71e^{-0.0016X}) + 0.015$
	$X \ge 1000$	$\mathrm{SI}_{\mathrm{SS}} = 0$
pH	$X < 5.5$	$\mathrm{SI}_{\mathrm{pH}} = 17.2 - 17.2\,X + 5.02\,X^2$
	$5.5 \le X < 7$	$\mathrm{SI}_{\mathrm{pH}} = -242 + 95.5\,X - 6.67\,X^2$
	$7 \le X < 8.75$	$\mathrm{SI}_{\mathrm{pH}} = -181 + 82.4\,X - 6.05\,X^2$
	$X \ge 8.75$	$\mathrm{SI}_{\mathrm{pH}} = 536 - 77\,X + 2.76\,X^2$

[a] X is the concentration parameter in terms of mg/L, except for pH and DO. For DO, X refers to DO percentage saturation and for pH it refers to the pH value

Table 4 Water quality classes, WQI and water status (DoE 2005)

Parameters	Unit	Classes				
		I	II	III	IV	V
AN	mg/l	<0.1	0.1–0.3	0.3–0.9	0.9–2.7	>2.7
BOD	mg/l	<1	1–3	3–6	6–12	>12
COD	mg/l	<10	10–25	25–50	50–100	>100
DO	mg/l	>7	5–7	3–5	1–3	<1
pH	–	>7	6–7	5–6	<5	>5
SS	mg/l	<25	25–50	50–150	150–300	>300
Water quality index	–	>92.7	76.5–92.7	51.9–76.5	31.0–51.9	<31.0
Water Status	–	Very good	Good	Average	Polluted	Very polluted

Artificial neural networks (ANNs) methods

Artificial neural networks (ANNs) are a computational process which attempts to represent and compute a mapping from multivariate data set as inputs to another as outputs. A neuron is the smallest part of the neural network, with artificial neurons arranged in the structure like a network. In this study, feed forward back propagation neural network (FFBP) was used to predict WQI in the wetland. The network consists of a set of neurons in three, inputs, hidden and output layers to approximate a multivariant function of $f(x)$. The number of neurons in hidden layers can be detected by trial and error. The learning procedure includes the best weight vector to achieve the best approximation of $f(x)$. Firstly, a set of input data (x_1, x_2,…x_R) is fed to the input layer, and the output of each neuron can be determined from the following equation:

$$n = \sum w_{ij} x_i + b_i \tag{2}$$

where n is the neuron output, w_{ij} is weight of the connection between the jth neuron in the present layer and ith neuron in the previous layer, x_i is neuron value in the previous layer and b_i is the bias. The sigmoid function can be used as a transfer function to generate the output of each neuron (Bateni et al. 2007) given by:

$$y_i = \frac{1}{1 + e^{-C_1\left(\sum w_{ij} x_i + b_i\right)}}, \quad C_1 > 0 \tag{3}$$

A comparison between the target value and obtained results was used to estimate network errors, while the back propagation algorithm corrects the weight between neurons. The back-propagation (BP) method is a descent algorithm, which tries to minimize the error at each iteration. The network weights are set by the algorithm such

Table 5 Matrix of the weights for the principal components

Item	Component (factor)		
	1	2	3
SS	**0.85**	0.18	0.19
Nitrate	**0.84**	0.11	0.13
Phosphate	**0.84**	0.13	0.16
AN	**0.81**	0.11	0.24
Nitrite	**0.81**	0.25	0.00
BOD	**0.79**	0.05	-0.06
COD	**0.77**	−0.11	0.19
pH	−0.03	**0.90**	0.20
Conductivity	0.16	**0.82**	−0.01
WQI	**−0.62**	**−0.69**	−0.09
Temperature	0.34	−0.03	**0.82**
DO	0.00	**0.47**	**0.68**
Eigenvalue	5.88	2.00	1.00
Initial variance (%)	49.00	16.72	8.32
Cumulative variance (%)	49.00	65.72	74.05
Total variance (%)			74.05

that the network error decreases along a descent direction (gradient descent). Generally two parameters, called momentum factor (MF) and learning rate (LR), are used to control the weight adjustment in the descent direction.

Sensitivity analysis using ANNs

In this study, the ANNs was employed to reduce the number of independent variables for prediction of the WQI. Range of data for sensitivity analysis is shown in Table 6.

A network with feed forward back propagation method (FFBP) was developed for sensitivity analysis. The number of neurons in the input layer was determined based on the number of input variables. Since the WQI was chosen as the network output, then the number of neurons in the output layer was selected equal to one. One layer was chosen in hidden layer and the optimum number of neurons in this layer was found equal to 5 using trial and error approach.

The leave-one-out method was used to assess the effect of each variable on the WQI. In this method, two indicators, the ratio of error and its rank, were estimated by removing each input variable at a time (Ha and Stenstrom 2003). The ratio of the error is obtained after elimination of individual variable to the error obtained using all variables. The high ratio illustrates the importance of individual variable and vice versa (Table 7).

Another attempt was conducted to determine the significance or influence of input variables on WQI. Table 8 compares the ANNs models with one of the independent variables removed in each case. As shown in this table, pH, COD, DO, AN and SS are significant variables with $R^2 = 0.9882$, RMSE $= 0.0179$ and MAE $= 0.0136$ and have a non-negligible influence on WQI. These parameters were chosen to developed GEP and ANNs in this study. Other parameters such as BOD, phosphate, nitrate, nitrite and conductivity do not have any significant effect on WQI and can be ignored.

In light of these findings, the pH with the highest rank can be considered as a main parameter for WQI in the wetlands (Table 7), although it is ranked only as the 6th variable in the conventional WQI equation. This equation (Eq. 1) is suggested for estimation of WQI in the rivers, and the difference between ranking of pH in Eq. (1) and the present study may be due to the discharge of the point source and non-point source pollution loads to rivers. However, the selected wetland is mainly polluted by discharge from non-point source pollution due to storm water.

Table 6 Range of data for training and testing

Parameters	Training		Testing	
	Min	Max	Min	Max
Temperature	27.30	35.15	28.38	34.76
Conductivity	94.00	206.00	96.00	189.00
Nitrite	0.00	0.06	0.00	0.05
Nitrate	0.20	4.50	0.20	4.90
Phosphate	0.11	0.58	0.12	0.48
pH	6.11	9.19	6.40	9.17
DO	4.96	11.06	5.13	10.81
AN	0.10	0.47	0.10	0.42
BOD	1.32	4.12	1.46	3.75
COD	9.00	44.00	11.00	44.00
SS	2.00	39.00	2.00	34.00
WQI	73.51	93.21	73.76	92.84

Table 7 Sensitivity analysis using ANNs

All variable without	Ratio	Rank
pH	1.226	1
COD	1.081	2
DO	1.048	3
AN	1.044	4
SS	1.020	5
BOD	1.019	6
Phosphate	1.000	7
Nitrate	1.000	8
Conductivity	0.999	9
Nitrite	0.998	10
Temperature	0.997	11

Table 8 Sensitivity analysis using different variables

Variables	Training			Testing		
	R^2	RMSE	MAE	R^2	RMSE	MAE
pH-COD	0.8235	0.0703	0.0530	0.7870	0.0758	0.0602
pH-COD-DO	0.8835	0.0571	0.0443	0.8130	0.0710	0.0555
pH-COD-DO-AN	0.9757	0.0261	0.0208	0.9636	0.0313	0.0255
pH-COD-DO-AN-SS	0.9926	0.0144	0.0112	0.9882	0.0179	0.0136
pH-COD-DO-AN-SS-BOD	0.9988	0.0058	0.0047	0.9911	0.0155	0.0073
pH-COD-DO-AN-SS-BOD-phosphate	0.9988	0.0059	0.0048	0.9970	0.0090	0.0068
pH-COD-DO-AN-SS-BOD-phosphate-nitrate	0.9989	0.0055	0.0044	0.9974	0.0084	0.0062
pH-COD-DO-AN-SS-BOD-phosphate-nitrate-conduc.	0.9991	0.0050	0.0039	0.9980	0.0073	0.0056
pH-COD-DO-AN-SS-BOD-phosphate-nitrate-conduc.-nitrite	0.9992	0.0048	0.0037	0.9981	0.0074	0.0055
All variables	0.9992	0.0047	0.0035	0.9981	0.0074	0.0055

This point can be considered for re-establishment of a new equation for WQI in the wetlands and other water resources with discharge from non-point pollution.

Development of GEP for water quality index

Gene expression programming (GEP) is a learning algorithm which was developed based on genetic programming (GP) and genetic algorithms (GA). In each individual population, the chromosomes are generated randomly and evaluated using a fitness function. Mutation is found as effective genetic operators to modify chromosomes. The following steps were used to develop GEP model.

In the first step, the size of the population was chosen equal to 30 as optimum size. Ferreira (2001) recommended a population size between 30 and 100 chromosomes as being able to provide an accurate result.

Secondly, the root relative squared error (RRSE) was chosen as fitness function in the GEP.

In the third step, a basic mathematical function (power), and four basic arithmetic operators (+, −, ×, /) were chosen to create chromosomes in each gene.

In next step, the chromosome architecture was chosen based on the length of the head, number of genes, and tail. The optimum result was determined for length head of seven and three genes per chromosome (Ferreira 2001, Mohammadpour et al. 2011, 2013b).

In the fourth step, both addition and multiplication operators were evaluated to find the best linking function, and the result showed that the addition function is more accurate. This function was employed to make a link between the sub-expression (chromosomes) in the GEP.

In the last step, the operators of GEP such as, mutation, transpositions, inversion, and cross-over, were employed to develop the GEP model.

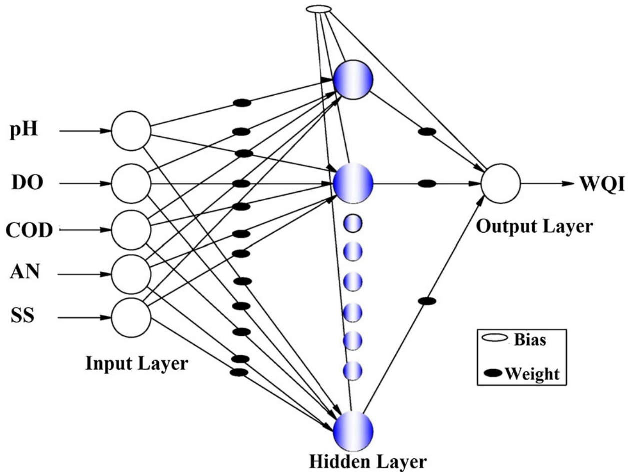

Fig. 2 Architecture of ANNs-FFBP for free constructed wetland

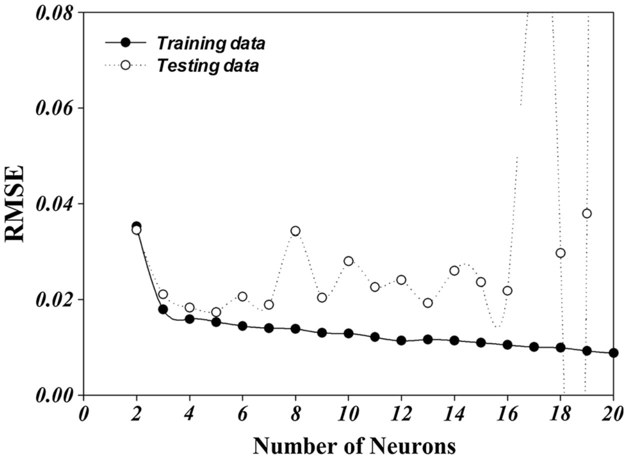

Fig. 3 Variation of RMSE for training and testing data in terms of number of neurons

Table 9 Performance of ANN with different neurons in hidden layer

No. of neurons	Training			Testing		
	R^2	RMSE	MAE	R^2	RMSE	MAE
2	0.9559	0.0352	0.0243	0.9552	0.0345	0.0247
3	0.9886	0.0179	0.0142	0.9832	0.0211	0.0156
4	0.9910	0.0159	0.0127	0.9873	0.0183	0.0139
5	0.9917	0.0153	0.0121	0.9887	0.0173	0.0130
6	0.9926	0.0145	0.0113	0.9840	0.0206	0.0156
7	0.9930	0.0140	0.0109	0.9866	0.0189	0.0147
8	0.9932	0.0139	0.0107	0.9559	0.0342	0.0181
9	0.9940	0.0130	0.0098	0.9843	0.0204	0.0148
10	0.9941	0.0129	0.0098	0.9704	0.0280	0.0177
11	0.9948	0.0121	0.0095	0.9808	0.0226	0.0161
12	0.9954	0.0114	0.0087	0.9782	0.0241	0.0174
13	0.9952	0.0116	0.0087	0.9860	0.0193	0.0146
14	0.9954	0.0114	0.0087	0.9745	0.0260	0.0183
15	0.9957	0.0110	0.0082	0.9790	0.0236	0.0182
16	0.9961	0.0105	0.0081	0.9820	0.0218	0.0165
17	0.9964	0.0101	0.0075	0.7040	0.0886	0.0278
18	0.9965	0.0099	0.0073	0.9666	0.0298	0.0213
19	0.9970	0.0093	0.0072	0.9458	0.0379	0.0204

Performance of the GEP and ANNs was assessed through the statistical parameters such as, coefficient of determination (R^2), mean absolute error (MAE) and root mean square error (RMSE). Expressions for these measures are given as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{p} (O_i - P_i)^2}{\sum_{i=1}^{p} (O_i - \overline{O}_i)^2} \tag{4}$$

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{p} |O_i - P_i| \tag{5}$$

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{p} (O_i - P_i)^2}{n}} \tag{6}$$

where O_i is observed values, P_i is predicted value, $\overline{O}_i$ is average of observed value and n is the number of samples.

Results and discussion

The total 442 datasets were divided randomly into training and testing subsets, 80 % (353 data set) for training and 20 % (89 data set) for testing (Table 6). Regarding the sensitivity analysis, five main variables of pH, COD, DO, AN and SS were employed to develop ANNs and GEP.

Figure 2 indicates an architecture of FFBP with five neurons as input and one neuron at output layer. Based on trial and error, the ANNs-FFBP network with 2000 epochs provided better results in comparison with the other networks.

The ANNs was developed with a different number of neurons in the hidden layer to find ANNs with the best performance. To assess over-fitting of network (low training error but high test error), the root mean square error (RMSE) was employed as a criterion. As shown in Fig. 3, the RMSE decreases dramatically with increasing number of neurons in the hidden layer. Table 9 indicates the performance of ANN-FFBP with different neurons in the hidden layer. The testing data was assessed to find the optimum number of neurons in hidden layer.

The best performance was provided for networks with five neurons in hidden layer. In this network, ANNs-FFBP predicts WQI with high accuracy in the wetland ($R^2 = 0.9887$, RMSE $= 0.0173$ and MAE $= 0.0130$). An over-fitting was observed in testing data for a number of neurons bigger than 5.

To evaluate the WQI, the GEP model has been developed using the same data set employed for the ANNs. The GEP expression tree is shown in Fig. 4. The simplified analytic form of the GEP model can be expressed as:

$$\begin{aligned}\mathrm{WQI} = {} & \left[\left(\frac{8.5 + 0.85SS}{\mathrm{DO}}\right)(\mathrm{AN} - 0.81)(\mathrm{pH}) + \mathrm{AN} - 7.68\right] \\ & \times (\mathrm{AN}) - \frac{\mathrm{DO}^2 - 7.63\ \mathrm{DO}}{\mathrm{COD}} - 0.19\mathrm{COD} \\ & - (\mathrm{pH} - 7.31)^2 + 96.63\end{aligned} \tag{7}$$

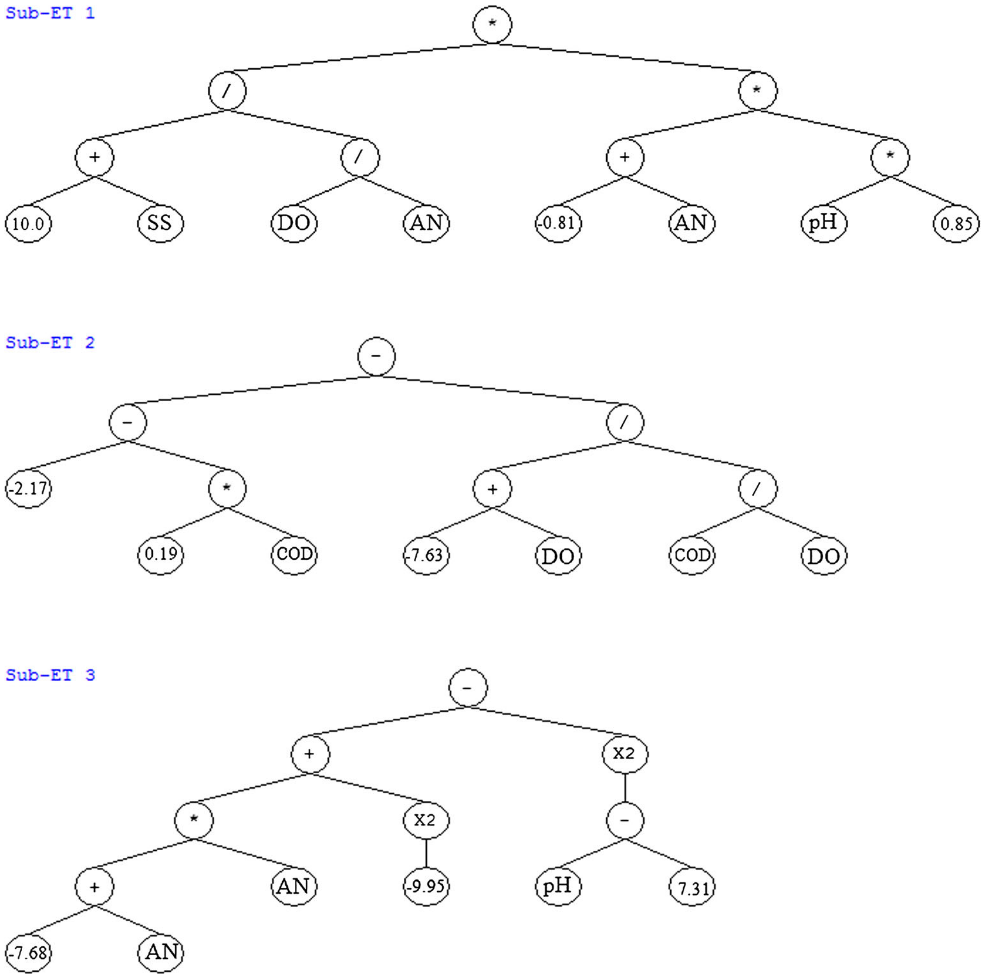

Fig. 4 Expression trees for the GEP equation

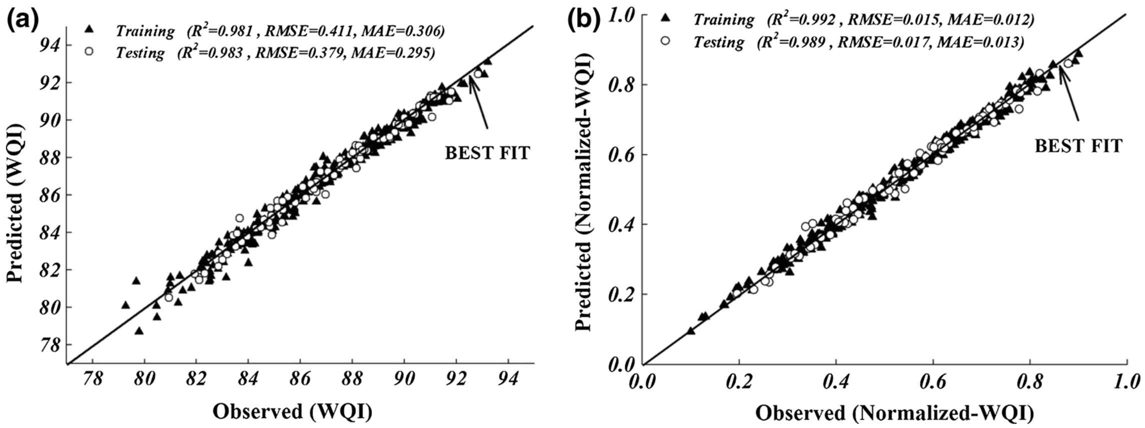

Fig. 5 Comparison between predicted and observed WQI using **a** GEP; **b** ANN-FFBP

This equation predicts WQI in constructed wetlands with only five direct variables instead of sub-index variables. Therefore, this equation is more useful and rapid in comparison with Eq. (1).

A comparison between predicted and observed WQI for both GEP and ANN-FFBP is shown in Fig. 5 and Table 10. It should be noted that the raw dataset was used to develop GEP (Fig. 5a) while the normalized dataset was employed

Table 10 Statistical parameters to predicted WQI using the GEP and ANNs

Method	Dataset	R^2	RMSE	MAE
GEP	Training	0.981	0.411	0.306
	Testing	0.983	0.379	0.295
ANNs	Training	0.9917	0.0153	0.0121
	Testing	0.9887	0.0173	0.0130

for prediction of WQI in ANN-FFBP (Fig. 5b). Prediction of proposed GEP with $R^2 = 0.983$, RMSE = 0.379 and MAE = 0.295 is comparable with ANNs ($R^2 = 0.988$, RMSE = 0.017 and MAE = 0.013). The results indicate that the both GEP (Eq. 7) and ANN-FFBP can be used as a reliable and precise method in the range of the collected data (Table 2). Furthermore, these methods propose some advantages in comparison with the traditional method.

Firstly, the BOD is excluded in both GEP and ANN, and these methods have been developed using five variables. Therefore, the number of variables required is less than those in traditional methods which required six sub-indices. Furthermore, measurement of BOD requires significant time, cost and commitment. The BOD test is run in the dark at 20 °C for 5 days. The temperature is specified because the rate of oxygen consumption is temperature dependent, and with no light source to eliminate the possibility of photosynthesis. However, determination of BOD is a very time-consuming process in comparison with other variables. Therefore, the recommended methods are more rapid and cost effective.

Secondly, the conventional equation recommended by DoE (2005) employs six sub-indices parameters, which requires a more cumbersome attempt and longer time to convert the six raw data into its sub-indices (Table 3). In addition, instead of using the original parameters, all parameters are based on the sub-indices (Eq. 1) which should be obtained from rating curves. In contrast, both the GEP and ANN approaches use the raw variables rather than the sub-indices which lead to a direct prediction of the WQI. Most importantly, the GEP and ANN techniques are more direct, rapid, and convenient compared to the conventional method.

Thirdly, the proposed GEP (Eq. 7) is more practical in comparison to ANNs, and raw data without normalization can be used in this equation. In comparison with conventional equation, GEP is more direct, convenient, and rapid. An example is mentioned in the Appendix to compare calculation of WQI based on proposed and traditional methods. The WQI obtained by GEP with a value of 84.15 is close to the value obtained by the traditional equation (84.34). The results show that GEP is accurate, simple and quick to calculate WQI. In this sample, the water was classified as group-II with a range of WQI between 76.5 and 92.7 (Table 4).

Accordingly, this research highlights that the GEP and ANN-FFBP can be employed as valuable techniques for estimation of water quality in the FSCW. These methods simplify the calculation of the WQI and reduce substantial time and effort by optimizing the computations. These approaches are highly recommended to be used for water quality assessment of any aquatic system in the world. This research should encourage the researchers and managers to apply the GEP and ANN-FFBP methods as more direct and reliable alternatives to estimate water quality in wetlands and other water bodies.

Conclusions

In this study, GEP and ANNs techniques were employed to develop the WQI in the free surface constructed wetlands. Seventeen points of the wetland were monitored twice a month over a period of 14 months, and an extensive data set was collected for 11 water quality variables. The PFA was employed to interpret correlation between WQI and other variables. This analysis indicated that WQI was greatly affected by pH and SS, while temperature had no significant effect on the WQI in tropical areas. A sensitivity analysis was carried out using ANNs to reduce the number of variables. Subsequently, five significant parameters including pH, COD, DO, AN and SS were chosen to develop GEP and ANN methods. A high value of the coefficient of correlation ($R^2 = 0.983$) and low error (MAE = 0.295) indicated that the GEP method was able to successfully predict the WQI with high accuracy. The statistical parameters indicate that, although the ANN-FFBP with $R^2 = 0.988$ and MAE = 0.013 produced better results compared with GEP, the GEP-based formula is more useful for practical purposes. This research highlights that the GEP and ANN-FFBP can be employed as powerful and highly reliable methods to estimate water quality in wetlands and other water bodies. These two techniques are highly recommended to be used for accurate, quick and cost effective water quality assessments for any aquatic system in the world.

Compliance with ethical standards

Conflict of interest The authors would like to acknowledge the financial assistance from the Ministry of Education under the Long Term Research Grant (LRGS) No. 203/PKT/672004 entitled “Urban Water Cycle Processes, Management and Societal Interactions: Crossing from Crisis to Sustainability”. This study is funded under a subproject entitled “Sustainable Wetland Design Protocol for Water Quality Improvement” (Grant number: 203/PKT/6724002).

Appendix

A data sampling was collected in the free surface constructed wetland with Temp. = 33.12 °C, DO = 9.05 mg/l, BOD = 2.54 mg/l, COD = 34 mg/l, AN = 0.21 mg/l, SS = 26 mg/l and pH = 8.10. Determine the WQI in the wetland using the conversional equation (DoE 2005) and GEP equation?

1. Determine WQI based on DoE (2005):
   DO = 9.05 mg/l and Temperature = 33.12 °C then DO (%saturation) = 126.34 > 92 $\mathrm{SI}_{\mathrm{DO}} = 100$.
   BOD = 2.54 ≤ 5, then $\mathrm{SI}_{\mathrm{BOD}} = 89.66$.
   COD = 34 > 20, then $\mathrm{SI}_{\mathrm{COD}} = 59.04$.
   AN = 0.21 ≤ 0.3, then $\mathrm{SI}_{\mathrm{AN}} = 78.45$
   SS = 26 ≤ 100, then $\mathrm{SI}_{\mathrm{SS}} = 83.09$.
   pH = 8.10, 7 ≤ pH 100 < 8.75, then $\mathrm{SI}_{\mathrm{pH}} = 89.49$.

$$\begin{aligned} \mathrm{WQI} = {} & 0.22\,(100) + 0.19\,(89.99) + 0.16\,(59.04) \\ & + 0.15\,(78.45) + 0.16\,(83.09) \\ & + \; 0.12\,(89.49) \\ = {} & 84.34 \end{aligned}$$

2. Determine WQI based on GEP equation:

$$\begin{aligned} \mathrm{WQI} = {} & \left[\left(\frac{8.5 + 0.85(26)}{9.05}\right)(0.21 - 0.81)(8.1) \right. \\ & \left. + 0.21 - 7.68\right](0.21) - \frac{9.05^2 - 7.63(9.05)}{34} - 0.19(34) \\ & - (8.1 - 7.31)^2 + 96.63 = 84.15 \end{aligned}$$

References

Aras E, Toğan V, Berkun M (2007) River water quality management model using genetic algorithm. Environ Fluid Mech 7:439–450

Azamathulla HM (2012) Gene expression programming for prediction of scour depth downstream of sills. J Hydrol 460–461:156–159

Azamathulla HM, Ahmad Z (2012) Gene-expression programming for transverse mixing coefficient. J Hydrol 434–435:142–148

Azamathulla HM, Ghani AA (2011) Genetic programming for predicting longitudinal dispersion coefficients in streams. Water Resour Manag 25:1537–1544

Azamathulla HM, Ghani AA, Zakaria NA (2010) Prediction of Scour below Flip Bucket using Soft Computing Techniques. Iscm Ii and Epmesc Xii, Pts 1 and 2, 1233, 1588–1593

Chen L, Tan CH, Kao SJ, Wang TS (2008) Improvement of remote monitoring on water quality in a subtropical reservoir by incorporating grammatical evolution with parallel genetic algorithms into satellite imagery. Water Res 42:296–306

Civelekoglu G, Yigit NO, Diamadopoulos E, Kitis M (2009) Modelling of COD removal in a biological wastewater treatment plant using adaptive neuro-fuzzy inference system and artificial neural network. Water Sci Technol 60:1475–1487

Bateni SM, Borghei SM, Jeng DS (2007) Neural network and neuro-fuzzy assessments for scour depth around bridge piers. Eng Appl Artif Intell 20(3):401–414

Dadaser-Celik F, Cengiz E (2013) A neural network model for simulation of water levels at the Sultan Marshes wetland in Turkey. Wetlands Ecol Manag 21:297–306

Department of Environment (2005) Malaysia Environmental Quality Report. Ministry of Natural Resources and Environment, Petaling Jaya

Fernández N, Ramírez A, Solano F (2004) Physico-chemical water quality indices—a comparative review *Bistua: Revista de la Facultad de Ciencias Básicas,* num. pp 19–30

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

Gazzaz NM, Yusoff MK, Aris AZ, Juahir H, Ramli MF (2012) Artificial neural network modeling of the water quality index for Kinta River (Malaysia) using water quality variables as predictors. Mar Pollut Bull 64:2409–2420

Ghani AA, Azamathulla HM (2014) Development of GEP-based functional relationship for sediment transport in tropical rivers. Neural Comput Appl 24:271–276

Ha H, Stenstrom MK (2003) Identification of land use with water quality data in stormwater using a neural network. Water Res 37:4222–4230

Hashmi MZ, Shamseldin AY, Melville BW (2011) Statistical downscaling of watershed precipitation using gene expression programming (GEP). Environ Model Softw 26:1639–1646

Juahir H, Zain SM, Toriman ME, Mokhtar M, Man HC (2004) Application of artificial neural network models for predicting water quality index. J Kejuruteraan Awam 16:42–55

Karthikeyan L, Kumar DN, Graillot D, Gaur S (2013) Prediction of ground water levels in the uplands of a tropical coastal riparian wetland using artificial neural networks. Water Resour Manag 27:871–883

Kashefi Alasl M, Khosravi M, Hosseini M, Pazuki GR, Nezakati Esmail Zadeh R (2012) Measurement and mathematical modelling of nutrient level and water quality parameters. Water Sci Technol 66:1962–1967

Khan F, Husain T, Lumb A (2003) Water quality evaluation and trend analysis in selected watersheds of the Atlantic region of Canada. Environ Monit Assess 88(1–3):221–248

Khuan LY, Hamzah N, Jailani R (2002) Prediction of water quality index (WQI) based on artificial neural network (ANN). In: Proceedings of the student conference on research and development, Shah Alam, Malaysia

Lambrakis N, Antonakos A, Panagopoulos G (2004) The use of multicomponent statistical analysis in hydrogeological environmental research. Water Res 38:1862–1872

Li W, Cui L, Zhang Y, Zhang M, Zhao X, Wang Y (2013) Statistical modeling of phosphorus removal in horizontal subsurface constructed wetland. Wetlands:1–11

Liou SM, Lo SL, Wang SH (2004) A generalized water quality index for Taiwan. Environ Monit Assess 96:35–52

Mohammadpour R, Ghani AA, Azamathullah HM (2011) Estimating time to equilibrium scour at long abutment by using genetic programming. 3rd international conference on managing rivers in the 21st century, Rivers 2011. Penang, Malaysia

Mohammadpour R, Ghani AA, Azamathullah HM (2013a) Numerical modeling of 3-D flow on porous broad crested weirs. Appl Math Model 37:9324–9337

Mohammadpour R, Ghani AA, Azamathullah HM (2013b) Prediction of equilibrium scour time around long abutments. Proc Inst Civil Eng Water Manag 166:394–401

Mohammadpour R, Ghani AA, Azamathullah HM (2013c) Estimation of dimension and time variation of local scour at short abutment. Int J River Basin Manag 11:121–135

Mohammadpour R, Ghani AA, Shaharuddin S, Kiat C, Chang NZ (2014a) Nitrogen removal assessment by multivariable statistical technique in free surface wetland. 13th international conference on urban drainage. Sarawak, Malaysia

Mohammadpour R, Shaharuddin S, Chang CK, Zakaria NA, Ghani AA (2014b) Spatial pattern analysis for water quality in free surface constructed wetland. Water Sci Technol 70:1161–1167

Mohammadpour R, Shaharuddin S, Chang C, Zakaria N, Ghani AA, Chan N (2015a) Prediction of water quality index in constructed wetlands using support vector machine. Environ Sci Pollut Res 22:6208–6219

Mohammadpour R, Ghani A, Vakili M, Sabzevari T (2015b) Prediction of temporal scour hazard at bridge abutment. Nat Hazard 1–21. doi:10.1007/s11069-015-2044-8

Ni Q, Wang L, Zheng B, Sivakumar M (2012) Evolutionary algorithm for water storage forecasting response to climate change with small data sets: the Wolonghu Wetland, China. Environ Eng Sci 29:814–820

Norhayati MT, Goh SH, Tong SL, Wang CW, Abdul Halim S (1997) Water quality studies for the classification of Sungai Bernam and Sungai Selangor. J Ensearch 10:27–36

Orouji H, Bozorg Haddad O, Fallah-Mehdipour E, Mariño MA (2013) Modeling of water quality parameters using data-driven models. J Environ Eng (United States) 139:947–957

Said A, Stevens DK, Sehlke G (2004) An innovative index for evaluating water quality in streams. Environ Manag 34(3):406–414

Schmid BH, Koskiaho J (2006) Artificial neural network modeling of dissolved oxygen in a Wetland Pond: the case of Hovi, Finland. J Hydrol Eng 11:188–192

Shaharuddin S, Zakaria NA, Ghani AA,Chang CK(2013) Performance evaluation of constructed Wetland in Malaysia for water security enhancement. In: Proceedings of 2013 IAHR world congress, China

Singh KP, Basant A, Malik A, Jain G (2009) Artificial neural network modeling of the river water quality—a case study. Ecol Model 220:888–895

Song K, Park YS, Zheng F, Kang H (2013) The application of Artificial Neural Network (ANN) model to the simulation of denitrification rates in mesocosm-scale wetlands. Ecol Inform 16:10–16

Verma AK, Singh TN (2013) Prediction of water quality from simple field parameters. Environ Earth Sci 69:821–829

Vink K, Schot P (2002) Multiple-objective optimization of drinking water production strategies using a genetic algorithm. Water Resour Res 38:201–2015

Wang L, Li X, Cui W (2012) Fuzzy neural networks enhanced evaluation of wetland surface water quality. Int J Comput Appl Technol 44:235–240

Xu TY, Qin XS (2013) Solving water quality management problem through combined genetic algorithm and fuzzy simulation. J Environ Inform 22:39–48

Zakaria NA, Ghani AA, Abdullah R, Mohd Sidek L, Ainan A (2003) Bio-ecological drainage system (BIOECODS) for water quantity and quality control. Int J River Basin Manag 1:237–251

Zakaria NA, Azamathulla HM, Chang CK, Ghani AA (2010) Gene expression programming for total bed material load estimation—a case study. Sci Total Environ 408:5078–5085

Zaman Zad Ghavidel S, Montaseri M (2014) Application of different data-driven methods for the prediction of total dissolved solids in the Zarinehroud basin. Stoch Environ Res Risk Assess 28(8):2101–2118. doi:10.1007/s00477-014-0899-y

Zandbergen PA, Hall KJ (1998) Analysis of the British Columbia Water Quality Index for watershed managers: a case study of two small watersheds. Water Qual Res J Can 33:519–549