

# Prediction of water quality index in constructed wetlands using support vector machine

Reza Mohammadpour · Syafiq Shaharuddin · Chun Kiat Chang ·  
Nor Azazi Zakaria · Aminuddin Ab Ghani · Ngai Weng Chan

Received: 8 July 2014 / Accepted: 2 November 2014 / Published online: 19 November 2014  
© Springer-Verlag Berlin Heidelberg 2014

**Abstract** Poor water quality is a serious problem in the world which threatens human health, ecosystems, and plant/animal life. Prediction of surface water quality is a main concern in water resource and environmental systems. In this research, the support vector machine and two methods of artificial neural networks (ANNs), namely feed forward back propagation (FFBP) and radial basis function (RBF), were used to predict the water quality index (WQI) in a free constructed wetland. Seventeen points of the wetland were monitored twice a month over a period of 14 months, and an extensive dataset was collected for 11 water quality variables. A detailed comparison of the overall performance showed that prediction of the support vector machine (SVM) model with coefficient of correlation ( $R^2$ )=0.9984 and mean absolute error (MAE)=0.0052 was either better or comparable with neural networks. This research highlights that the SVM and FFBP can be successfully employed for the prediction of water quality in a free surface constructed wetland environment. These methods simplify the calculation of the WQI and reduce substantial efforts and time by optimizing the computations.

**Keywords** Support vector machine · Constructed wetland · Water quality index · Neural networks · Surface water

## Introduction

Municipal and industrial wastewaters from human activities are the major factors that contribute to the deterioration of water quality in urban areas. Water quality (WQ) can be used to assess the water properties in reference to human health and natural quality effects. The poor quality of surface water is a serious problem in the world which threatens human health, ecosystems, and plant/animal life. Consequently, water quality analysis has become a main concern in water resource and environmental systems (Espejo et al. 2012; Vanlandeghem et al. 2012; Zhang et al. 2013a, b; Wang et al. 2012). In terms of environmental and ecological problems, the number of water quality parameters is quite extensive. Hence, a robust mathematical technique is required to combine the physicochemical characterization of water into a single variable which describes the water quality. A water quality index (WQI) is a single number which uses a set of physicochemical water parameters to express the quality of water at a certain place and time.

In 1974, the Department of Environment of Malaysia (DoE) recommended the WQI parameter for categorizing and estimating water quality. Based on this parameter, the water quality was classified into five different classes according to the water's suitability for various uses such as water supplies, irrigation, and fish culture. The conventional method suggested by DoE requires lengthy transformations to estimate subindices. In addition, the subindices required the inclusion of different equations, which need lengthy effort and time to calculate the final WQI. Therefore, estimation of such a WQI is cumbersome and can lead to occasional mistakes. However, the support vector machine (SVM) and artificial neural networks (ANNs) can be suggested as alternatives for estimation of WQI, as both employ the raw data instead of subindices.

Responsible editor: Michael Matthies

R. Mohammadpour (✉) · S. Shaharuddin · C. K. Chang ·  
N. A. Zakaria · A. A. Ghani · N. W. Chan  
River Engineering and Urban Drainage Research Centre (REDAC),  
Universiti Sains Malaysia, Engineering Campus, Seri Ampangan,  
14300 Nibong Tebal, Penang, Malaysia  
e-mail: reza564@gmail.com

The performance of free surface wetlands to enhance water quality and reduce a wide range of wastewater was reported in several studies (Zedler and Kercher 2005; Vymazal 2011; Zhang et al. 2012; Wang et al. 2012; Shih et al. 2013; Mohammadpour et al. 2014). Wetlands have high ability to absorb and reduce agriculture and municipal wastewater and can also highly decrease various nutrients (Mitsch and Gosselink 2007; Kadlec and Wallace 2008). To remove pollutants and nutrients associated with fine particulates, several processes occur in the constructed wetlands such as settling, filtration, absorption, and biological uptake (Guardo 1999). Generally, three parts can be recognized in the free surface wetlands such as inlet, macrophyte, and open water area (Zakaria et al. 2003). The macrophyte area (wetland plants) has a high effect on the wetland ecosystem and water quality (Brix 1997; Kadlec and Wallace 2008). Since the number of variables which affect water quality is too high, the SVM can be proposed as a robust technique for prediction of water quality in a free constructed wetland environment.

Recently, soft computing techniques such as SVM, genetic programming (GP), and ANNs have been successfully employed to solve the problems related to engineering (Mohammadpour et al. 2011; Tabari et al. 2012; Kakaei Lafdani et al. 2013; Mohammadpour et al. 2013b; Ghani and Azamathulla 2014; He et al. 2014). The SVM is proposed as a leading technique which can be used for regression and classification purposes (Noori et al. 2011; Singh et al. 2011a, b). The SVM has high ability for generalization and is less prone to overfitting. Furthermore, it simultaneously minimizes the estimation of error and model dimensions (Singh et al. 2011a, b; Li et al. 2013). Sivapragasam and Muttill (2005) applied the SVM for prediction of water level in rivers by extension of rating curves. Khan and Coulibaly (2006) recommended the SVM as the appropriate tool to forecast lake water levels and obtained quite acceptable results. Singh et al. (2011a, b) employed the SVM for classification and regression of water quality. The SVM was applied as a classification tool for some studies related to wetlands (Dronova et al. 2012; Betbeder et al. 2013; Zhang and Xie 2013). Sadeghi et al. (2012) predicted the distribution pattern of *Azolla filiculoides* (Lam.) in wetlands using the SVM.

The ANNs have been recommended as an effective tool for the prediction of water pollution and water quality in the wetlands (Schmid and Koskiahoo 2006; Wang et al. 2012; Dong et al. 2012; Dadaser-Celik and Cengiz 2013; Li et al. 2013; Song et al. 2013). The ANNs were employed to model the constructed wetlands in different fields (Tomenko et al. 2007; Singh et al. 2011a, b; Zhang et al. 2013a, b). The ANNs are a useful technique that was used to speed up the calculation of water quality index in rivers (Khuan et al. 2002; Juahir et al. 2004; Gazzaz et al. 2012). Mohammadpour et al. (2013a) examined two kinds of ANNs, feed forward back

propagation (FFBP) and radial basis function (RBF), for the prediction of time variation of local scour in rivers. Nourani et al. (2013) applied the FFBP network to determine the quality of treated water. Diamantopoulou et al. (2005) used the neural networks to predict water quality in a river in Greece. Khalil et al. (2011) estimated water quality characteristics at ungauged sites using ANN and ensemble ANN (EANN). The results showed that the EANN provides better prediction than the ANN.

In this research, both SVM and ANNs were used as robust techniques for rapid and direct prediction of the WQI in the constructed wetlands which can be used as another alternative for some long-lasting conventional methods. Seventeen points in the wetland were monitored twice a month over a period of 14 months and an extensive dataset was collected for 11 water quality variables. A sensitivity analysis was conducted to find more significant variables on WQI. Finally, the SVM result was compared with two models of neural networks, namely the FFBP and RBF.

## Materials and methods

### Study area

The free surface constructed wetland in the Universiti Sains Malaysia (USM) was chosen as a case study in this research. The wetland is covered by different kinds of plant species where the *Hanguana malayana* was the dominant species. The volume of wetland is 16,312 m<sup>3</sup> with a water depth between 0.25 and 2.54 m from inlet to outlet of the wetland (Table 1). The wetland is located at latitude 5° 9' 7.8294" North and longitude 100° 29' 53.1672" East. It was designed based on the Stormwater Management Manual for Malaysia to improve water quality and provide better wildlife habitat (Zakaria et al. 2003; Shaharuddin et al. 2013). Seventeen sampling points with different plant species and water depths were chosen to monitor the water quality. As shown in Fig. 1, these points included the inlet, six stations in macrophyte area (W1 to W6), nine points in micropool (MA1 to MC3), and the outlet.

The data collection was carried out twice a month over a period of 14 months (from October 2010 to December 2011). Totally, 11 water quality variables (WQVs) were collected in the wetland, including temperature, pH, dissolved oxygen (DO), conductivity, suspended solid (SS), nitrite, nitrate, ammoniacal nitrogen (AN), biochemical oxygen demand (BOD), chemical oxygen demand (COD), and phosphate. The final dataset consisted of 442 samples and 11 WQVs. Table 2 indicates the statistical parameters of WQVs in the wetland.

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

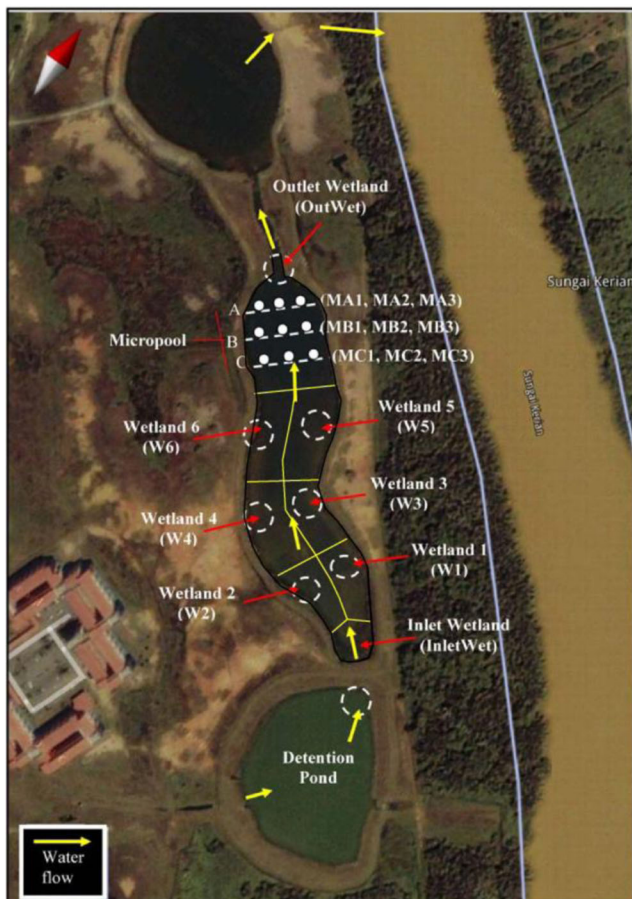
Site	Wetland plant species	Water depth (m)
Wetland 1	Dominant: <i>Hanguana malayana</i> , <i>Lepironia articulata</i>	0.25–0.3
Wetland 2	Dominant: <i>Hanguana malayana</i> , <i>Typha angustifolia</i> Less dominant: <i>Scirpus grossus</i>	0.27–0.32
Wetland 3	Dominant: <i>Lepironia articulata</i> , <i>Eleocharis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

The local water quality index

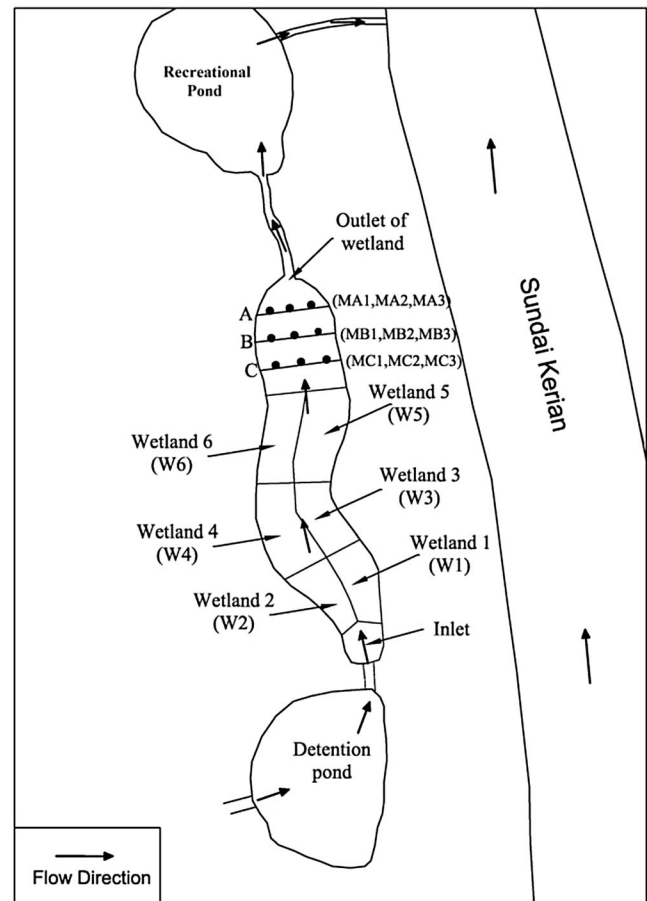
To determine the WQI, six physicochemical parameters were proposed by Department of Environment (DoE 2005), namely DO, BOD, COD, AN, SS, and pH. The WQI was recognized as a unitless variable with a value from 0 to 100, where a high value of WQI represents high water quality. As shown in Table 2, the mentioned

WQVs should be converted into nondimensional variables using subindex functions (SI). Finally, the WQI can be estimated using the following equation (DoE 2005; Khuan et al. 2002):

$$\text{WQI} = 0.22 \text{SI}_{\text{DO}} + 0.19 \text{SI}_{\text{BOD}} + 0.16 \text{SI}_{\text{COD}} + 0.15 \text{SI}_{\text{AN}} + 0.16 \text{SI}_{\text{SS}} + 0.12 \text{SI}_{\text{pH}} \quad (1)$$



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



**Table 2** Descriptive statistics of wetland parameters (number of data=442)

WQV	Min	Max	Mean	Mean standard error	Std. deviation
Temperature (°C)	27.30	35.15	31.12	0.07	1.52
pH	6.11	9.19	7.73	0.03	0.69
DO (mg/l)	4.96	11.06	8.24	0.04	0.87
Conductivity (µs/cm)	94.00	206.00	136.59	1.21	25.49
Nitrite (mg/l)	0.00	0.06	0.02	0.00	0.01
Nitrate (mg/l)	0.20	4.90	2.15	0.04	0.91
Phosphate (mg/l)	0.11	0.58	0.25	0.00	0.10
AN (mg/l)	0.10	0.47	0.22	0.00	0.07
BOD (mg/l)	1.32	4.12	2.53	0.02	0.47
COD (mg/l)	9.00	44.00	21.96	0.27	5.60
SS (mg/l)	2.00	39.00	17.75	0.37	7.79

The subindices equations were provided in Table 3, where  $X$  is the concentration parameter in terms of mg/l, except for pH and DO. For DO, the  $X$  refers to the percentage of saturation, and for pH, it refers to the pH value. According to the calculated WQI, the water can be classified into one of five classes. Table 4 shows the water quality classes suggested by the DoE.

Support vector machine technique

The SVM proposed by Vapnik (1995, 1998) was developed based on the statistical learning theory. The SVM is a novel classification technique which uses the principle of structural

risk minimization (SRM) and transforms it into quadratic programming.

The SVM uses suitable kernel function to map the original data into a high-dimensional feature space where a maximal separating plane (SP) is constructed. To separate the data, two parallel hyperplanes can be developed on each side of the SP. The SVM simultaneously maximizes the geometric margin and minimizes the empirical classification error (Singh et al. 2011a, b). The SVM was extended to solve the regression problems with the introduction of  $\epsilon$ -insensitive loss function (Pan et al. 2008). In this case, the SVM attempts to determine the optimal hyperplane which minimizes the distance of all data points (Qu and Zuo 2010; Lin et al. 2008).

A detailed expression of the SVM may be found elsewhere (Vapnik 1995; Smola and Scholkopf 2004). However, a brief discussion of this technique is mentioned here. For a set of training data  $\{x_i, y_i\}$ , the objective of SVM is to determine a function  $f(x)$  with high deviation ( $\epsilon$ ) from targets ( $y_i$ ) and it should be flat as possible at the same time. If  $f(x)$  is introduced as a linear discriminant function, then the SVM can be presented as (Smola and Scholkopf 2004):

$$f(x) = (w, x) + b \tag{2}$$

where  $w$  is the weight vector ( $w \in R^n$ ), and  $b$  is the bias. The  $f(x)$  function is flatted by minimizing the values of  $w$ . Using a convex optimization problem, minimization can be expressed as (Singh et al. 2011a, b):

$$\begin{cases} \text{Minimize } \frac{1}{2} \|w\|^2 \\ \text{subject to } \begin{cases} y_i^-(w, x_i) - b \leq \epsilon \\ (w, x_i) + b - y_i \leq \epsilon \end{cases} \end{cases} \tag{3}$$

The last equation in some cases with more errors can be introduced with slack variables  $\xi_i$ , then minimization formula changes as follows: (Vapnik 1995):

**Table 3** The subindex equation for WQI in Malaysia (DoE 2005)

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

**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

$$\left\{ \begin{array}{l} \text{Minimize } \frac{1}{2} \|w\|^2 + C \sum_{i=1}^m (\xi_i + \xi_i^*) \\ \text{subject to } \begin{cases} y_i - (w, x_i) - b \leq \varepsilon + \xi_i \\ (w, x_i) + b - y_i \leq \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \geq 0 \end{cases} \end{array} \right. \quad (4)$$

The  $C$  is a penalty parameter, and it should be defined by the user. This parameter determines the trade-off between the tolerable amount larger than  $\varepsilon$  and flatness of  $f(x)$ . The Lagrangian form of minimization formula can be express as:

$$L = \frac{1}{2} \|w\|^2 + C \sum_{i=1}^m (\xi_i + \xi_i^*) - \sum_{i=1}^m (\eta_i \xi_i + \eta_i^* \xi_i^*) - \sum_{i=1}^m \alpha_i (\varepsilon_i + \xi_i - y_i + (w, x_i) + b) - \sum_{i=1}^m \alpha_i^* (\varepsilon_i + \xi_i - y_i + (w, x_i) + b); \begin{cases} \alpha_i, \eta_i \geq 0 \\ \alpha_i^*, \eta_i^* \geq 0 \end{cases} \quad (5)$$

where are  $\alpha_i$ ,  $\eta_i$ ,  $\alpha_i^*$ , and  $\eta_i^*$  are Lagrangian parameters. The saddle points of Eq. (5) can be estimated as:

$$\frac{\partial L}{\partial b} = 0 \quad \text{then} \quad \sum_{i=1}^m (\alpha_i^* - \alpha_i) = 0 \quad (6)$$

$$\frac{\partial L}{\partial \xi_i^*} = 0 \quad \text{then} \quad C - \alpha_i^* - \eta_i^* = 0 \quad (8)$$

$$\frac{\partial L}{\partial w} = 0 \quad \text{then} \quad w - \sum_{i=1}^m (\alpha_i^* - \alpha_i) x_i = 0 \quad (7)$$

Dual maximization problem is determined by substituting Eqs. (6) to (8) into Eq. (5) as (Smola and Scholkopf 2004):

$$\begin{aligned} & \text{Maximize} \left\{ -\frac{1}{2}, \sum_{i=1}^m (\alpha_i^* - \alpha_i) (\alpha_j - \alpha_j^*) (x_i, x_j), -\varepsilon, \sum_{i=1}^m (\alpha_j + \alpha_j^*) + \sum_{i=1}^m y_i (\alpha_j - \alpha_j^*) \right\} \\ & \text{subject to } \sum_{i=1}^m (\alpha_j - \alpha_j^*) = 0 \quad \text{and} \quad \alpha_i, \alpha_i^* \in [0, C] \end{aligned} \quad (9)$$

Finally the SVM function can be expressed as:

$$f(x) = \sum_{i=1}^m (\alpha_i - \alpha_i^*) (x_i, x_j) + b \quad (10)$$

The kernel function was used to solve the nonlinear problem in the support vector regression. This function maps the data into higher dimension feature space (Vapnik 1995). The support vector regression problem in the feature space was

expressed using  $K(x_i, x_j)$  instead of  $(x_i, x_j)$ , then the SVM can be written as:

$$f(x) = \sum_{i=1}^m (\alpha_i - \alpha_i^*) K(x_i, x_j) + b \quad (11)$$

Four possible choices for the kernel function are suggested, namely linear, polynomial, sigmoid, and radial Gaussian. The



most common kernel functions are polynomial and Gaussian or RBF which were expressed as:

$$K(x_i, x_j) = (1 + x_i \cdot x_j)^p \text{ Polynomial kernel function} \quad (12)$$

$$K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2) \text{ Gaussian kernel function} \quad (13)$$

where  $p$  and  $\gamma$  are adjustable kernel parameters. The performance of SVM depends on the combination of several parameters, such as the type of kernel function and its adjustable parameters, penalty parameter,  $C$ , and  $\epsilon$ -insensitive loss function. The selection of kernel function depends on the distribution of the data and generally can be selected through the trial and error approach (Widodo and Yang 2007). Since the RBF is employed in most of the applications (Xie et al. 2008), then in this study, the RBF (Gaussian kernel function) was chosen as kernel function for the prediction of WQI.

The  $\gamma$  is the most important parameter for the RBF kernel function. The amplitude of the kernel can be controlled by this parameter, and it can lead to overfitting and underfitting in prediction. The best value of  $\gamma$  can be found by trial and error (Noori et al. 2011). The  $C$  is a regularization parameter that controls the trade-off between maximizing the margin and minimizing the training error. For a low value of  $C$ , insufficient fitting will be placed on the training data, while the algorithm will overfit for too large of  $C$  (Wang et al. 2007). A well-performing and robust regression model is dependent on a proper choice of  $C$  in combination with  $\epsilon$  (Ustun et al. 2005). A range of 0.001–20,000 was chosen for the  $C$  parameter to investigate the optimal value (Fig. 3). Furthermore, since the exact contribution of the noise in the training set is usually unknown, the  $\epsilon$  was optimized in the range of 0.00001 and 0.1 (Ustun et al. 2005). To achieve a good combination of the two variables ( $C$  and  $\epsilon$ ), an internal cross-validation was performed during the construction of the SVM model. In this study, the SVM-base classification was performed using

Library for Support Vector Machines (LIBSVM) in MATLAB (Chang and Lin 2011).

Artificial neural network methods

ANNs are a computational process which attempts to represent and compute a mapping from multivariate dataset as inputs to another as outputs. A neuron is the smallest part of the neural network; these artificial neurons are arranged in the structure like a network. In this study, two models of neural networks, FFBP and RBF, were presented and a brief description of these methods are given here.

Feed forward back propagation neural network

The network consists of a set of neurons in three, inputs, hidden, and output, layers to approximate a multivariate function  $f(x)$ . 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  $(x_1, x_2, \dots, x_R)$  is fed to the input layer, and the output of each neuron can be determined from the following relation:

$$n = \sum w_{ij}x_i + b_i \quad (14)$$

where  $n$  is the neuron output,  $w_{ij}$  is the weight of the connection between the  $j$ th neuron in the present layer and  $i$ th neuron in the previous layer,  $x_i$  is the 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:

$$y_i = \frac{1}{1 + e^{-C_1 \left( \sum w_{ij}x_i + b_i \right)}}, C_1 > 0 \quad (15)$$

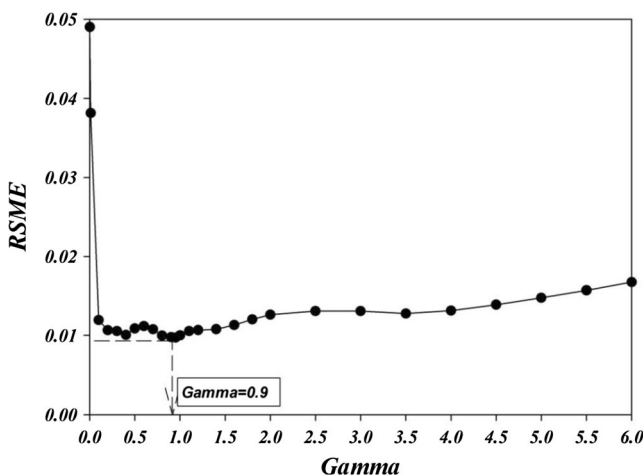


Fig. 2 Variation of gamma values in terms of RMSE in the SVM model

Table 5 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

The network error calculation uses a comparison between the target value and the obtained results, 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 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.

### Radial basis function neural network

The RBF network is a general regression tool for approximate function that uses a radial basis function as activation function. In this study, the Softmax transfer function is used to estimate the  $\varnothing$  value at each node.

$$\phi_j(x) = \frac{\exp\left(-\frac{\|x-\mu_j\|^2}{2\sigma_j^2}\right)}{\sum \exp\left(-\frac{\|x-\mu_j\|^2}{2\sigma_j^2}\right)} \quad (16)$$

where  $x$  is the input dataset,  $\mu_j$  is the center of the radial basis function for the  $j$ th hidden node,  $\sigma_j$  is a radius of the radial basis function for the  $j$ th hidden node, and  $\|x-\mu_j\|^2$  is the Euclidean norm. The linear interconnectedness between network outputs and hidden nodes can be explained by the following equation:

$$y_k = \sum w_{kj} \phi_j(x) \quad (17)$$

**Table 7** Range of data used for training and testing

Parameters	Training		Testing	
	Min	Max	Min	Max
pH	6.11	9.19	6.40	9.17
COD	9.00	44.00	11.00	44.00
DO	4.96	11.06	5.13	10.81
AN	0.10	0.47	0.10	0.42
SS	2.00	39.00	2.00	34.00
BOD	1.32	4.12	1.46	3.75
Phosphate	0.11	0.58	0.12	0.48
WQI	73.51	93.21	73.76	92.84

where  $y_k$  is the  $k$ th component of the output layer, and  $w_{kj}$  is the weight between the  $j$ th hidden node and  $k$ th node of the output layer.

In this study, the collected dataset was normalized within the range of 0.1–0.9. Three common statistical measures, namely the coefficient of correlation ( $R^2$ ), root mean square error (RMSE), and mean absolute error (MAE), were used to validate the provided results.

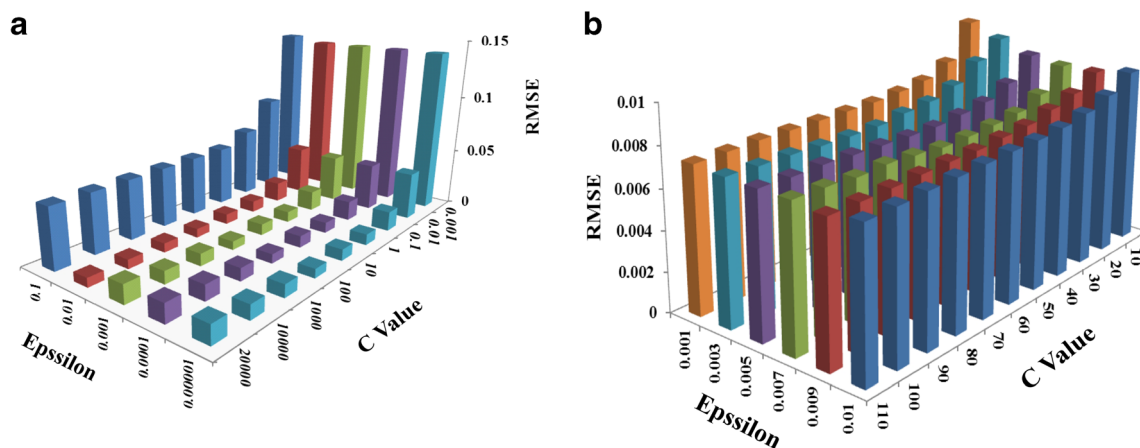
## Results and discussion

### Determining the main water quality variables

To determine the main WQVs in the prediction of the WQI, a sensitivity analysis was carried out using ANN. The sensitivity analysis was conducted using the FFBP network with one hidden layer. The number of neurons in the input layer was

**Table 6** Sensitivity analysis using different variables

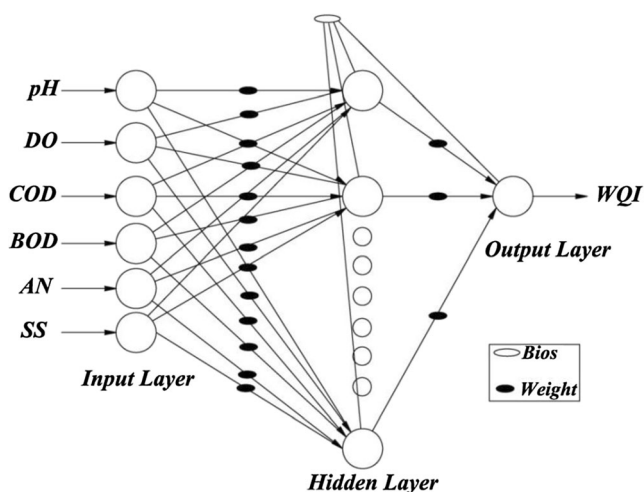
Variables	Training			Testing		
	$R^2$	RMSE	MAE	$R^2$	RMSE	MAE
All 11 variables	0.9992	0.0047	0.0035	0.9981	0.0074	0.0055
pH-COD-DO-AN-SS-BOD-phosphate-nitrate-conductivity-nitrite	0.9992	0.0048	0.0037	0.9981	0.0074	0.0055
pH-COD-DO-AN-SS-BOD-phosphate-nitrate-conductivity	0.9991	0.0050	0.0039	0.9980	0.0073	0.0056
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	0.9988	0.0059	0.0048	0.9970	0.0090	0.0068
pH-COD-DO-AN-SS-BOD	0.9988	0.0058	0.0047	0.9911	0.0155	0.0073
pH-COD-DO-AN-SS	0.9926	0.0144	0.0112	0.9882	0.0179	0.0136
pH-COD-DO-AN	0.9757	0.0261	0.0208	0.9636	0.0313	0.0255
pH-COD-DO	0.8835	0.0571	0.0443	0.8130	0.0710	0.0555
pH-COD	0.8235	0.0703	0.0530	0.7870	0.0758	0.0602



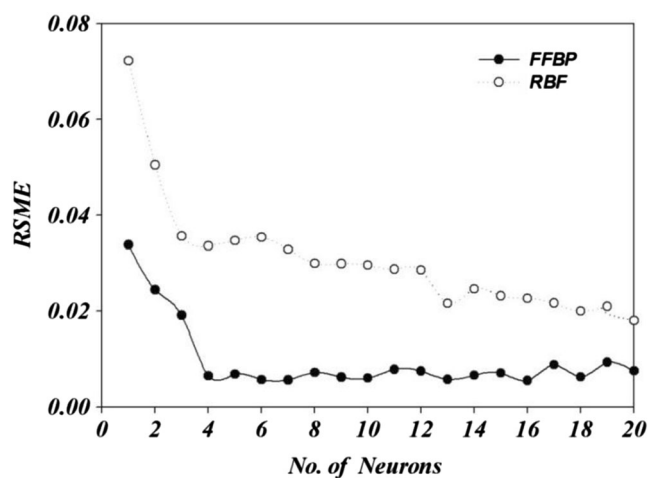
**Fig. 3** The accuracy of SVM model in terms of  $C$  and  $\epsilon$ : **a** coarse grid, **b** fine grid

determined based on the number of WQVs as input to ANN. In the output, a layer with one neuron was chosen for WQI. The leave-one-out technique was employed to assess the effect of each variable on the WQI. By removing one variable in the input at each time, two indicators were determined, namely the ratio of ANN error and its rank (Ha and Stenstrom 2003). The ratio of ANN error was found by removing an individual variable to the error obtained using all variables. The high value for this ratio can be interpreted as high significance of individual variable and vice versa. The result of this analysis is shown in Table 5. Generally, in the environmental works, the simple equations with a few variables are more practical and useful in comparison with complex equations. Therefore, another sensitivity analysis was conducted to simplify the classification and reduce the number of variables on WQI. Table 6 compares the ANN 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 a high performance of ANNs ( $R^2=0.9882$ ,  $RMSE=0.0179$ , and  $MAE=0.0136$ ). The performance of ANNs decreases with removal of one of these variables in the next rows. Therefore, these variables were chosen to develop the SVM, FFBP, and ANNs-RBF models. It was observed that removing more variables such as BOD, phosphate, nitrate, nitrite, and conductivity has no considerably effect on the performance of ANNs. In light of these findings, the pH with the highest rank in the sensitive analysis can be considered as a main parameter (Table 5), while it is ranked as the sixth variable in the conventional WQI equation (Eq. 1). This equation is recommended for estimation of WQI in the rivers, and the difference between ranking of pH in Eq. 1 and the present study can be due to the discharge at the point source and nonpoint pollution to rivers. However, the selected wetland is discharged just by nonpoint source pollution due to storm water. Same ranking results were observed by Gazzaz et al. (2012). This point may be considered for re-



**Fig. 4** Architecture of neural network for free constructed wetland



**Fig. 5** Variation of RMSE for testing data in terms of the number of neurons



**Table 8** List of statistical data for training and testing

Network	No. of neurons	Training			Testing		
		$R^2$	RMSE	MAE	$R^2$	RMSE	MAE
FFBP	1	0.9494	0.0377	0.0253	0.9573	0.0338	0.0265
	2	0.9836	0.0214	0.0146	0.9778	0.0244	0.0162
	3	0.9882	0.0182	0.0126	0.9864	0.0191	0.0141
	4	0.9983	0.0069	0.0055	0.9985	0.0064	0.0052
	5	0.9985	0.0065	0.0052	0.9983	0.0068	0.0055
	6	0.9989	0.0057	0.0043	0.9988	0.0056	0.0044
	7	0.9987	0.0061	0.0048	0.9988	0.0056	0.0044
	8	0.9989	0.0054	0.0041	0.9981	0.0071	0.0052
	9	0.9992	0.0048	0.0037	0.9986	0.0061	0.0047
RBF	1	0.8752	0.0666	0.0517	0.8460	0.0722	0.0553
	2	0.9141	0.0552	0.0388	0.9247	0.0505	0.0395
	3	0.9515	0.0415	0.0290	0.9626	0.0356	0.0293
	4	0.9571	0.0390	0.0267	0.9668	0.0336	0.0247
	5	0.9583	0.0385	0.0271	0.9645	0.0347	0.0253
	10	0.9737	0.0306	0.0215	0.9742	0.0296	0.0221
	20	0.9902	0.0166	0.0122	0.9879	0.0180	0.0129
	30	0.9948	0.0121	0.0091	0.9932	0.0134	0.0104
	40	0.9965	0.0099	0.0077	0.9960	0.0103	0.0080
50	0.9965	0.0099	0.0077	0.9960	0.0103	0.0080	

establishment of a new equation for WQI in the wetlands and other water resource with nonpoint pollution discharge.

#### The SVM result

Based on the provided result by sensitivity analysis, all 11 water quality variables can be reduced to just six significant variables including pH, COD, DO, AN, SS, and BOD. In the present study, the selected dataset (442 samples $\times$ 6 variables) was randomly divided into training dataset (354 samples $\times$ 7 variables) and testing subsets dataset (88 samples $\times$ 7 variables). Thus, the training and validation (testing) dataset were comprised of 80 and 20 % of samples, respectively. Table 7 summarizes the range of training and testing dataset.

The SVM models with different values of  $\gamma$  were developed to determine the best value for this parameter. The RMSE was chosen to assess the accuracy of SVM models. Figure 2 indicates that the minimum error was obtained for  $\gamma=0.9$ , and this value was chosen to determine  $C$  and  $\varepsilon$  in the next steps. Furthermore, a tenfold cross-validation and grid search algorithm were employed to find the optimal values of  $C$  and  $\varepsilon$  (Hsu et al. 2003).

In tenfold cross-validation, the collected data were randomly divided into ten equal groups, where eight groups were used for training, and the rest of the groups were employed for validation. The grid search algorithm takes different samples

from the space of the independent variables. In each step, the prediction of the model was compared with the best value provided from the previous iterations. If the newly found values were better than the previous one, the new values were used. The grid search algorithm is an unguided technique, and it was developed based on trial and error (Hsu et al. 2003; Noori et al. 2011). A two-step grid search with cross-validation was employed to solve this problem and determine the tune values of  $C$  and  $\varepsilon$  (Chen and Yu 2007). In the first step, a coarse grid search was applied to determine the best region of demanded parameters. In the next step, a finer grid search was employed to recognize the optimal combination of parameters. In Fig. 3, the accuracy of SVM was assessed regarding  $C$  and  $\varepsilon$  values. Through a coarse grid search (Fig. 3a), the optimum values of  $C$  and  $\varepsilon$  were determined

**Table 9** Comparison of three techniques to predict WQI in constructed wetland

Method	Training			Testing		
	$R^2$	RMSE	MAE	$R^2$	RMSE	MAE
FFBP-ANNs	0.9987	0.0061	0.0048	0.9988	0.0056	0.0044
SVM	0.9987	0.0061	0.0053	0.9984	0.0066	0.0052
RBF-ANNs	0.9965	0.0099	0.0077	0.9960	0.0103	0.0080

over a space of 10–100 and 0.0001–0.01, respectively. Finally, the best value of  $C$  and  $\epsilon$  was found equal to 57 and 0.007 through the fine grid search (Fig. 3b).

The neural network results

The ANN model was developed using the same dataset employed for the SVM. Figure 4 indicates the architecture of FFBP with six neurons in the input layer for WQVs and one neuron in the output layer for WQI. Two types of ANNs, FFBP and RBF, were developed to investigate the best network for the reliable prediction of WQI. Based on trial and error, the FFBP network with 2000 epochs and the RBF network with the spread constant of 1.1 provided better results in comparison to the other networks.

Since the ANNs are sensitive to the number of neurons in the hidden layer and these neurons were unknown in the first step, then the ANNs were developed with different numbers of neurons in the hidden layer. The RMSE was employed to assess overfitting of network (low training error but high test error). As shown in Fig. 5, the RMSE decreased dramatically with increasing number of neurons in the hidden layer, especially in the FFBP method. An overfitting was observed in the FFBP networks when the number of neurons was more than 16. The performance of both FFBP and RBF with different neurons in the hidden layer is indicated in Table 8.

The testing data was assessed to find the optimum number of neurons. The best performance for FFBP and RBF was provided for networks with six and 40 neurons in the hidden layer. For testing data, the FFBP with  $R^2=0.9988$ ,  $RMSE=0.0056$ , and  $MAE=0.0044$  predicts the WQI with high accuracy. As shown in this table, the accuracy of RBF ( $R^2=0.9960$ ,  $RMSE=0.0103$ , and  $MAE=0.0080$ ) is a little lower than the FFBP method.

Comparison between SVM and ANNs

In Table 9, the performance of the SVM was compared with both the FFBP and RBF models. The SVM with  $R^2=0.9984$ ,  $RMSE=0.0066$ , and  $MAE=0.0052$  forecasts the WQI in the wetland better than the RBF ( $R^2=0.9960$ ,  $RMSE=0.0103$ , and  $MAE=0.0080$ ). Furthermore, statistical parameters and scatter plot (Fig. 6) show that the prediction of SVM is comparable with FFBP with  $R^2=0.9988$ ,  $RMSE=0.0056$ , and  $MAE=0.0044$ .

In light of this research, it can be concluded that to predict the WQI in free surface constructed wetland environment, both the SVM and FFBP propose some advantages over the conventional method. The method recommended by DoE (2005) employs six subindices parameters, which need more effort and a long time to convert the six raw data (DO, BOD, COD, AN, SS, and pH) into its subindices (Table 3). Furthermore, instead of using the original parameters, all calculations are based on the subindices (Eq. 1) which are obtained from rating curves. In contrast, both the SVM and FFBP approaches use the raw WQVs for training and testing rather than the subindices which led to a direct prediction of the WQI. Therefore, the SVM and FFBP methods are more direct, rapid, and convenient techniques than the conventional method.

Accordingly, this research highlights that the SVM and FFBP can be used as valuable methods for the prediction of water quality in the constructed wetland as they simplify the calculation of the WQI and reduce substantial efforts and time by optimizing the computations. These approaches can be commonly used for any aquatic system in the world. This research should encourage the managers and authorities to use the SVM and FFBP methods as more direct and highly reliable alternatives to predict water quality in wetlands and other water bodies.

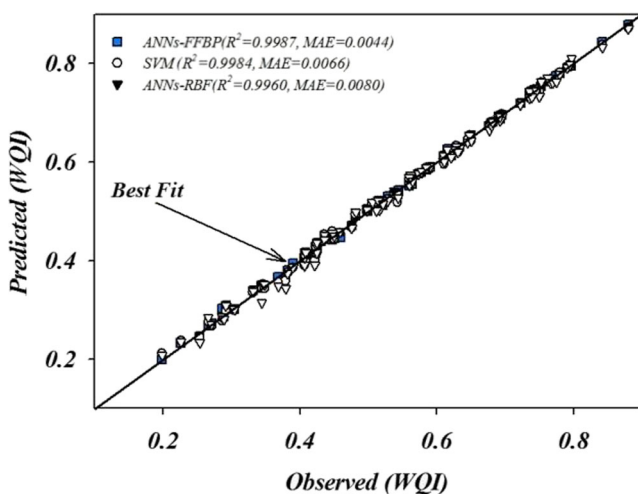


Fig. 6 Comparison between predicted and observed WQI

Conclusions

In this study, the SVM and two methods of ANNs, namely FFBP and RBF, were employed to investigate the WQI in the free surface constructed wetland. Seventeen points of the wetland were monitored twice a month over a period of 14 months, and an extensive dataset was collected for 11 water quality variables. A sensitivity analysis was carried out using ANN and six significant variables that included pH, COD, DO, AN, SS, and BOD to develop the SVM, FFBP, and ANNs-RBF models. The results illustrate that the SVM technique was able to successfully predict the WQI with high accuracy. The high value of the coefficient of correlation ( $R^2=0.9984$ ) and low error ( $MAE=0.0052$ ) indicated that the SVM model provides better prediction compared to the

RBF network with  $R^2=0.9960$  and  $MAE=0.0080$ . Furthermore, the result provided by SVM was comparable with that of the FFBP network ( $R^2=0.9988$  and  $MAE=0.0044$ ). This research highlights that the SVM and FFBP can be successfully used as valuable methods for the prediction of water quality in the wetlands. These methods simplify the calculation of the WQI and decrease the substantial efforts and time by optimizing the computations. The mentioned approaches can be commonly used as more direct and highly reliable techniques to predict water quality at any aquatic system worldwide.

**Acknowledgments** 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).

## References

- Batani SM, Borghei SM, Jeng DS (2007) Neural network and neuro-fuzzy assessments for scour depth around bridge piers. *Eng Appl Artif Intell* 20:401–414
- Betbeder J, Rapinel S, Corpetti T, Pottier E, Corgne S & Hubert-Moy L (2013) Multi-temporal classification of TerraSAR-X data for wetland vegetation mapping. *Proc SPIE Int Soc Opt Eng*
- Brix H (1997) Do macrophytes play a role in constructed treatment wetlands? *Water Sci Technol* 35:11–17
- Chang C-C, Lin C-J (2011) LIBSVM: a library for support vector machines. *ACM Trans Intell Syst Technol* 2:27
- Chen S-T, Yu P-S (2007) Real-time probabilistic forecasting of flood stages. *J Hydrol* 340:63–77
- Dadaser-Celik F, Cengiz E (2013) A neural network model for simulation of water levels at the Sultan Marshes wetland in Turkey. *Wetl Ecol Manag* 21:297–306
- Department of Environment. Malaysia Environmental Quality Report (2005) Department of Environment, Ministry of Natural Resources and Environment, Petaling Jaya, Malaysia
- Diamantopoulou MJ, Antonopoulos VZ, Papamichail DM (2005) The use of a neural network technique for the prediction of water quality parameters of Axios River in Northern Greece. *J Oper Res, Springer-Verlag*, 115–125
- Dong Y, Scholz M, Harrington R (2012) Statistical modeling of contaminants removal in mature integrated constructed wetland sediments. *J Environ Eng (United States)* 138:1009–1017
- Dronova I, Gong P, Clinton NE, Wang L, Fu W, Qi S, Liu Y (2012) Landscape analysis of wetland plant functional types: the effects of image segmentation scale, vegetation classes and classification methods. *Remote Sens Environ* 127:357–369
- Espejo L, Kretschmer N, Oyarzún J, Meza F, Núñez J, Maturana H, Soto G, Oyarzo P, Garrido M, Suckel F, Amezaga J, Oyarzún R (2012) Application of water quality indices and analysis of the surface water quality monitoring network in semiarid North-Central Chile. *Environ Monit Assess* 184:5571–5588
- 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 & Applic* 24:271–276
- Guardo M (1999) Hydrologic balance for a subtropical treatment wetland constructed for nutrient removal. *Ecol Eng* 12:315–337
- Ha H, Stenstrom MK (2003) Identification of land use with water quality data in stormwater using a neural network. *Water Res* 37:4222–4230
- He Z, Wen X, Liu H, Du J (2014) A comparative study of artificial neural network, adaptive neuro fuzzy inference system and support vector machine for forecasting river flow in the semiarid mountain region. *J Hydrol* 509:379–386
- Hsu CW, Chang CC, Lin CJ (2003) A practical guide to support vector classification. <<http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf>>
- Juahir H, Zain SM, Toriman ME, Mokhtar M, Man HC (2004) Application of artificial neural network models for predicting water quality index. *J KejuruteraanAwam* 16:42–55
- Kadlec RH, Wallace SD (2008) *Treatment wetlands*, 2nd edn. CRC, Boca Raton
- Kakaei Lafdani E, Moghaddam Nia A, Ahmadi A (2013) Daily suspended sediment load prediction using artificial neural networks and support vector machines. *J Hydrol* 478:50–62
- Khalil B, Ouarda TBMJ, St-Hilaire A (2011) Estimation of water quality characteristics at ungauged sites using artificial neural networks and canonical correlation analysis. *J Hydrol* 405:277–287
- Khan MS, Coulibaly P (2006) Application of support vector machine in lake water level prediction. *J Hydrol Eng* 11:199–205
- 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
- LiW, Cui L, Zhang Y, Zhang M, Zhao X & Wang Y (2013) Statistical modeling of phosphorus removal in horizontal subsurface constructed wetland. *Wetlands*, 1–11
- Lin S-W, Lee Z-J, Chen S-C, Tseng T-Y (2008) Parameter determination of support vector machine and feature selection using simulated annealing approach. *Appl Soft Comput* 8:1505–1512
- Mitsch WJ, Gosselink JG (2007) *Wetlands*, 4th edn. Wiley, New York
- Mohammadpour R, Ghani AA, Azamathulla HM (2013a) Prediction of equilibrium scours time around long abutments. *Proc Inst Civ Eng Water Manag* 166(7):394–401
- Mohammadpour R, Ab. Ghani A & Azamathulla 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, 6th–9th December, Penang, Malaysia
- Mohammadpour R, Ghani AA, Azamathulla HM (2013b) Estimation of dimension and time variation of local scour at short abutment. *Int J Rivers Basin Manag* 11:121–135
- Mohammadpour R, Shaharuddin S, Chang CK, Zakaria NA, Ghani AA (2014) Spatial pattern analysis for water quality in free surface constructed wetland. *Water Sci and Technol* 70:1161–1167
- Noori R, Karbassi AR, Moghaddamia A, Han D, Zokaei-Ashtiani MH, Farokhnia A, Gousheh MG (2011) Assessment of input variables determination on the SVM model performance using PCA, gamma test, and forward selection techniques for monthly stream flow prediction. *J Hydrol* 401:177–189
- Nourani V, Rezapour Khanghah T & Sayyadi M (2013) Application of the artificial neural network to monitor the quality of treated water. *Int J Manag Inf Technol*, 3
- Pan Y, Jiang J, Wang R, Cao H (2008) Advantages of support vector machine in QSPR studies for predicting auto-ignition temperatures of organic compounds. *Chemom Intell Lab Syst* 92:169–178
- Qu J, Zuo MJ (2010) Support vector machine based data processing algorithm for wear degree classification of slurry pump systems. *Measurement* 43:781–791

- Sadeghi R, Zarkami R, Sabetraftar K, Van Damme P (2012) Use of support vector machines (SVMs) to predict distribution of an invasive water fern *Azolla filiculoides* (Lam.) in Anzali wetland, southern Caspian Sea, Iran. *Ecol Model* 244:117–126
- 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. Proceedings of 2013 IAHR World Congress, China
- Shih SS, Kuo PH, Fang WT, Lepage BA (2013) A correction coefficient for pollutant removal in free water surface wetlands using first-order modeling. *Ecol Eng* 61:200–206
- Singh KP, Basant N, Gupta S (2011a) Support vector machines in water quality management. *Anal Chim Acta* 703:152–162
- Singh G, Kandasamy J, Shon HK, Chob J (2011b) Measuring treatment effectiveness of urban wetland using hybrid water quality—artificial neural network (ANN) model. *Desalin Water Treat* 32:284–290
- Sivapragasam C, Muttill N (2005) Discharge rating curve extension—a new approach. *Water Resour Manag* 19:505–520
- Smola A, Scholkopf B (2004) A tutorial on support vector regression. *Stat Comput* 14:199–222
- 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
- Tabari H, Kisi O, Ezani A, Hosseinzadeh Talaee P (2012) SVM, ANFIS, regression and climate based models for reference evapotranspiration modeling using limited climatic data in a semi-arid highland environment. *J Hydrol* 444–445:78–89
- Tomenko V, Ahmed S, Popov V (2007) Modelling constructed wetland treatment system performance. *Ecol Model* 205:355–364
- Ustun B, Melssen WJ, Oudenhuijzen M, Buydens LMC (2005) Determination of optimal support vector regression parameters by genetic algorithms and simplex optimization. *Anal Chim Acta* 544:292–305
- Vanlandeghem MM, Meyer MD, Cox SB, Sharma B, Patiño R (2012) Spatial and temporal patterns of surface water quality and ichthyotoxicity in urban and rural river basins in Texas. *Water Res* 46:6638–6651
- Vapnik VN (1995) The nature of statistical learning theory. Springer, New York
- Vapnik VN (1998) Statistical learning theory. Wiley, New York
- Vymazal J (2011) Enhancing ecosystem services on the landscape with created, constructed and restored wetlands. *Ecol Eng* 37:1–5
- Wang J, Du H, Liu H, Yao X, Hu Z, Fan B (2007) Prediction of surface tension for common compounds based on novel methods using heuristic method and support vector machine. *Talanta* 73:147–156
- 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
- Widodo A, Yang BS (2007) Application of nonlinear feature extraction and support vector machines for fault diagnosis of induction motors. *Expert Syst Appl* 33:241–250
- Xie X, Liu WT, Tang B (2008) Spacebased estimation of moisture transport in marine atmosphere using support vector regression. *Remote Sens Environ* 112:1846–1855
- 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 Rivers Basin Manag* 1:237–251
- Zedler JB, Kercher S (2005) Wetland resources: status, trends, ecosystem services, and restorability. *Annu Rev Environ Resour* 30:39–74
- Zhang C & Xie Z (2013) Object-based vegetation mapping in the Kissimmee River watershed using HyMap data and machine learning techniques. *Wetlands*, 1–12
- Zhang T, Xu D, He F, Zhang Y, Wu Z (2012) Application of constructed wetland for water pollution control in China during 1990–2010. *Ecol Eng* 47:189–197
- Zhang H, Sun L, Sun T, Li H, Luo Q (2013a) Spatial distribution and seasonal variation of polycyclic aromatic hydrocarbons (PAHs) contaminations in surface water from the Hun River, Northeast China. *Environ Monit Assess* 185:1451–1462
- Zhang Y, Cui L, Li W, Zhang M, Zhao X, Wang Y (2013b) Modeling phosphorus removal in horizontal subsurface constructed wetland based on principal component analysis. *Nongye Gongcheng Xuebao Trans Chin Soc Agric Eng* 29:200–207