Neural Comput & Applic
DOI 10.1007/s00521-016-2404-7

ORIGINAL ARTICLE

Application of artificial intelligence (AI) techniques in water quality index prediction: a case study in tropical region, Malaysia

Mohammed Hameed[1] · Saadi Shartooh Sharqi[2] · Zaher Mundher Yaseen[1] · Haitham Abdulmohsin Afan[1] · Aini Hussain[3] · Ahmed Elshafie[4]

Received: 26 February 2016 / Accepted: 30 May 2016

Abstract The management of river water quality is one the most significant environmental challenges. Water quality index (WQI) describes several water quality variables at a certain aquatic environment and time. Classically, WQI is commonly computed using the traditional methods which involved lengthy computation, consume timing and occasionally associated with accidental errors during subindex calculation. Thus, providing an accurate prediction model for WQI is highly required. Recently, the artificial neural networks (ANNs) have been examined for similar prediction applications and exhibited a remarkable ability to capture the nonlinearity pattern between predictors and predictand. In the current research, two different ANN algorithms, namely radial basis function neural network (RBFNN) and back propagation neural networks models, have been applied to examine and mimic the relationship of WQI with the water quality variables in a tropical environment (Malaysia). The input variables categorized into two different architectures and have been inspected. In addition, comprehensive analysis for the performance evaluation and the sensitivity analysis of the variables have been conducted. The results achieved are positively promising with high performance accuracy belonging to RBFNN model for both scenarios. Furthermore, the proposed approach offers an effective alternative to compute and predict WQI, to the fact that WQI manual calculation methods involved lengthy computations, transformations, use of various subindex formulae for each value of the constituent water quality variables, and consuming time.

Keywords Artificial neural networks · Water quality index · Tropical environment · RBFNN · BPNN · Water quality variables

1 Introduction

Last two decades, water river pollution considered as one of the global concern to the communities which is required full attention from the environmental researchers [1]. However, water river quality is one of the main characteristics that needs full attention from environmental scholars. The quality of the water becomes a growing concern throughout the developing world. The process of abstraction water for domestic use, agricultural production, mining industrial production, power generation, and forestry practices can lead to deterioration in water quality and quantity that impact not only the aquatic ecosystem (i.e., the assemblage of organisms living and interacting together within an aquatic environment), but also the availability of safe water for human consumption. Thus, assessment of the quality of surface waters is important in hydro-environmental management and it is very significant in monitoring the concentration of pollutants in rivers. The

✉ Zaher Mundher Yaseen
zahermundher@gmail.com

[1] Civil and Structural Engineering Department, Faculty of Engineering and Built Environment, Universiti Kebangsaan Malaysia (UKM), 43600 Bangi, Selangor Darul Ehsan, Malaysia

[2] Department of Civil Engineering, Al Anbar University, Al Anbar, Iraq

[3] Department of Electrical, Electronics and Systems Engineering, Faculty of Engineering, Universiti Kebangsaan Malaysia (UKM), 43600 Bangi, Selangor Darul Ehsan, Malaysia

[4] Civil Engineering Department, Faculty of Engineering, University of Malaya, 50603 Kuala Lumpur, Malaysia

Published online: 08 June 2016

evaluation of the river water quality is a phrase to describe the chemical, biological, and physical characteristic of water.

Over the past few decades, increasing in number of industrial areas blooms throughout the country resulting to enormous amount of anthropogenic activities being introduce to the environment which eventually effect the water bodies. Nowadays, with the advanced science and technology, the increase in human population, industries, agriculture activities, and urban development have causes the riverbanks to widen. Nonetheless, unstoppable human activities such as agricultural using pesticides, domestic sewage, factories effluents, and even soil erosion due to improper development have led to pollution in the water bodies [2]. However, water pollution mainly occurs due to the overloading of waste in the water system. The contamination of streams, lakes, underground water, bays, or oceans by substances can be harmful to living things, not only human, but the wildlife and plants. Pollutants that have been discharges into water body via point source and nonpoint source are difficult to identify due to indefinite origin of the pollutants. However, factors such as effluent from industrial area, sewage disposals, and land clearings greatly influence the water quality. In fact, the water pollution affects not only human health, but also the entire environment, especially aquatic lives that live inside it [3, 4].

For the time being, Malaysia is struggling to develop the infrastructure of the country to meet the vision of 2020. Unluckily, the evolution that had been carried through the infrastructure contributes a bad impact to the environment, particularly in regard to water quality. Water quality index considered as the basic function for the environmental assessment of surface water in relation to pollution load categorization and designation of classes and beneficial uses as provided by Interim National Water Quality Standards (INWQS) in Malaysia [5]. Malaysia has 189 river basins nationwide; in particular, Peninsular Malaysia is characterized by highly diverse ecosystem and support extensive artisanal fisheries. In addition, rivers are considered the main source of drinking water for the zone. Rainfall is the main source of water for all kinds of surface water (i.e., channels, rivers, impoundment, and reservoirs). Those forms contribute 96 % of water supply sources. According to Department of Environment (DOE), out of the 189 rivers 120 rivers are controlled and monitored; around 44.5 % clean rivers, 48.4 % slightly polluted, and 7.1 % are polluted. Forty years ago, the DOE suggested an adoption of WQI to assess and rank the level of water rivers pollution. Thereafter, the DOE approved an approach called “Opinion Poll WQI (OP-WQI)” for calculating the index of water quality of local rivers. The method that used for calculating the WQI in Malaysia involves lengthy calculations, transformations, consuming time, and effort. Therefore, recommending an alternative technique which is direct and quicker with high accuracy of computing the WQI is obligatory.

Most recently, artificial intelligence methods such as artificial neural networks (ANNs) have been successfully employed to solve many prediction problems related to engineering [6–13]. ANNs modeling have the ability to deal with unknown knowledge in order to learn the complex model functions from examples, i.e., by training the input/output data sets. The greatest advantage that ANNs characterized over other modeling approaches is their ability to model the complex pattern, nonlinear processes without any advance knowledge form of the relationship between input and output variables. The last two decades, applications of ANN have been reported in many studies of water resources and engineering. ANN’s techniques have been utilized intensively in the field of water resource for pattern matching, optimization [14], data compression, forecasting, and predictions [15–17]. Several researchers have been used ANN in water resources application; for instance, storm water prediction [18], waste water modeling [19], heavy metal prediction [20], sediment transport modeling [21], stream-flow forecasting [22, 23], water level forecasting [24].

The use of soft computing techniques has been successfully applied in modeling freshwater quality [25–29] and in seawater [30–32]. Muttil and Chau [33] investigated the past, present, and prospect of integrating AI into water modeling. The researchers indicated that the limitations of water quality data and the high cost of water quality monitoring often pose a serious problem for process-based modeling methods. Hence, artificial intelligence gives the optimal solution, as they are computationally very fast and require many fewer input parameters and input conditions than deterministic models. The earliest research has been conducted in water quality modeling by Lek and Guégan [34]. Zou et al. [35] proposed neural network embedded Monte Carlo (NNMC) approach to model the uncertainty in water quality applications. The results indicated that NNMC approach has potential for efficient uncertainty analysis of water quality modeling. Kralisch et al. [36] used an ANN technique for the optimization of watershed management to maintain a reasonable balance between water quality demand and consequent restrictions for the farming industry. In 2004, Maier et al. [37] predicted the optimal alum doses and treated water quality variables. Authors found that the proposed method is capable to produce accurate prediction model. Diamantopoulou et al. [38] used the application ANN to predict stream-flow water quality parameters. The results show a satisfactory prediction modeling of nitrates (NO_3), specific electrical conductivity (EC), and dissolved oxygen parameters. Elhatip and Kömür [39] studied the water quality parameters in reservoir. The results illustrated the capability of ANN to predict the dissolved oxygen of recharge and discharge of reservoir bank. Singh et al. [40] predicted the biochemical oxygen demand (BOD) and dissolved oxygen (DO) in river banks using feed forward neural network algorithm (FFNN). The researchers concluded that FFNN algorithm performed good results in predicting BOD and DO water quality variables in rivers. An assessment for the effecting of physiochemical on the treatment performance of contaminated wetland sediment in replicate integrated constructed wetland was conducted by Dong et al. [41]. An evaluation for wetland surface water quality utilizing fuzzy neural network was accomplished by Wang et al. [42]. Total dissolved solid (TDS) prediction in river runoff using two ANN algorithms, namely multi-layer perceptron (MLP) and radial basis function (RBF), computed by Niroobakhsh [43]. Authors indicated that MLP and RBF are able to simulate water quality variables in river with more than 90 % accuracy.

The aim of this research is to construct two different ANN algorithms, namely back propagation neural network (BPNN) and radial basis function neural network (RBFNN), models to produce an efficient predictive model of WQI. In addition, to explore whether it is possible to predict WQI with rolling out the BOD variable as it requires a high controlled laboratory conditions and needs time that makes BOD test could be costly, furthermore, to establish a neural network model that applicable to predict WQI status of stream-flow river basin in tropical zone, and thus to provide an authoritative alternative to compute WQI instead of manual calculation. In the following section, an introduction to the manual calculation for the WQI is provided. In Sect. 3, materials and methods are presented. In Sect. 4, the application and discussion are described. Finally, the conclusions from this research are presented in Sect. 5.

2 Manual calculation of WQI for Malaysia

Thirty years ago, an opinion poll formula adopted to calculate WQI for Malaysia. The formula was proposed by the Department of Environment (DOE), Malaysia. A panel of professionals in water quality was consulted on the choice of the parameters and the weightage of each parameter. The six parameters that have been chosen by DOE are dissolved oxygen (DO), biochemical oxygen demand (BOD), chemical oxygen demand (COD), suspended solid (SS), ammoniacal nitrogen (AN), and pH value (pH). The WQI approved by the DOE is calculated based on the mentioned six parameters, the formula as follow:

$$\begin{aligned}\mathrm{WQI} = {} & 0.22\mathrm{SL}_{\mathrm{DO}} + 0.19\mathrm{SL}_{\mathrm{BOD}} + 0.16\mathrm{SL}_{\mathrm{COD}} \\ & + 0.16\mathrm{SL}_{\mathrm{SS}} + 0.15\mathrm{SL}_{\mathrm{AN}} + 0.12\mathrm{SL}_{\mathrm{PH}}. \qquad (1)\end{aligned}$$

Apparently, the DO carries the maximum weight value, whereas pH carries the minimum weight value. Finally, the evaluation of WQI formula consists of the subindexes, which are calculated according to the best-fit relations given in Table 1. The main advantage of using AI technique over manual calculation to assess WQI is that the manual calculation consumes a lot of time and converting the raw data to subpollutant index is needed.

3 Materials and methods

3.1 Study area and data collection

Two rivers have been used in this research as a study area Langat River and Klang River, Peninsular Malaysia. Langat River has a total catchment area approximately 1815 km^2. The total length of the river around 141 km, and it is located 40 km east Kuala Lumpur. There are two reservoirs located on this river Langat Reservoir and the Semenyih Reservoir. Climatology, the annual temperature, humidity, and rainfall are considerably high. This fact has a dominant influence on the hydrological and geomorphological process of the river environment. Lately, Langat River water quality has become considerably poor due to rapid urbanization along the river. The rapid growth of urban land use has been at the expense of agricultural land and reclamation of peat and swampy areas. Whereas, Klang River has a total length about 120 km with estimated area catchment around 1290 km^2. Klang River is a significant river in the west coast of Peninsular Malaysia, as it encompasses many cities in that area. Klang River basin is considered the most densely populated area of the country. The watershed has experienced the highest economic growth in the country as it included about 35 % of which is developed for residential, commercial, industrial, and institutional use. It is worth mentioning that the discarding of unprocessed sewage from treatment plants and livestock farms contributes to the further deterioration of the Klang River's water quality, adding to its solid waste and suspended-sediment loading. Geographically, Fig. 1 illustrates the two river location.

The selected data for this research represent the monthly measurement of water quality variables and the evaluated water quality index. The water quality variables are dissolved oxygen (DO), biochemical oxygen demand (BOD), chemical oxygen demand (COD), ammoniacal nitrogen (NH_3-N), suspended solid (SS), and PH used to evaluate the WQI, for Malaysia. The data obtained from several monitoring stations located on Klang Basin and Langat Basin and their branches that pass Selangor and Kuala

Table 1 Subindex calculation formulae for the local WQI [44]

Parameter	Value[a]	Subindex equation
DO (%saturation)[b]	$X \leq 8$	$\mathrm{SL}_{\mathrm{DO}} = 0$
	$8 < x < 92$	$\mathrm{SL}_{\mathrm{DO}} = -0.395 + 0.03X^2 - 0.0002X^3$
	$x \geq 92$	$\mathrm{SL}_{\mathrm{DO}} = 0$
BOD	$X \leq 5$	$\mathrm{SL}_{\mathrm{BOD}} = 100.4 - 4.2X$
	$x > 5$	$\mathrm{SL}_{\mathrm{BOD}} = (108\mathrm{e}^{-0.055X}) - 0.1X$
COD	$X \leq 20$	$\mathrm{SL}_{\mathrm{COD}} = 99.1 - 1.33X$
	$x > 20$	$\mathrm{SL}_{\mathrm{BOD}} = (103\mathrm{e}^{-0.0157X}) - 0.04X$
NH_3-N	$X \leq 0.3$	$\mathrm{SL}_{\mathrm{AN}} = 100.5 - 105X$
	$0.3 < x < 4$	$\mathrm{SL}_{\mathrm{AN}} = (94\mathrm{e}^{-0.573X}) - 5(X - 2)$
	$x \geq 4$	$\mathrm{SL}_{\mathrm{AN}} = 0$
SS	$X \leq 100$	$\mathrm{SL}_{\mathrm{SS}} = (97.5\mathrm{e}^{-0.00676X}) - 0.05X$
	$100 < x < 1000$	$\mathrm{SL}_{\mathrm{SS}} = (71\mathrm{e}^{-0.0016X}) - 0.015X$
	$x \geq 1000$	$\mathrm{SL}_{\mathrm{SS}} = 0$
PH	$X < 5.5$	$\mathrm{SL}_{\mathrm{PH}} = 17.2 - 17.2X - 6.67X^2$
	$5.5 \leq x < 7$	$\mathrm{SL}_{\mathrm{PH}} = -242 + 95.5X - 2.76X^2$
	$7 \leq x < 8.75$	$\mathrm{SL}_{\mathrm{PH}} = -181 + 82.4X - 6.05X^2$
	$x \geq 8.75$	$\mathrm{SL}_{\mathrm{PH}} = 536 + 77X - 2.76X^2$

For DO, X refers to DO percentage saturation and for pH it refers to the pH value

[a] X is the concentration of the indicated parameter in mg/L, except for pH and DO

[b] DO (%saturation) = [DO (mg/L) * 12.795] − 0.05

Lumpur states. The data set consisted of 5233 samples for the time period from January 2001 to October 2010. These data divided into two set: The first set used for training the model that involved the data for the time period from 2001 to 2008 and the rest of data used for testing the model.

3.2 Artificial neural network modeling

3.2.1 *Back propagation neural network (BPNN)*

ANN is inspired by simulating the function of a human brain, owing to its ability to model high complex relationships or to find partners in data; ANN can be widely used to represent a nonlinear mapping between input and output vectors. BPNN is the most broadly neural network algorithm has been employed for prediction of water resources and management problems [45]. Figure 2 illustrates the typical architecture of BPNN: input layer, hidden layer, and output layer. The input layer indicates the correspondence water quality variables that connected to the hidden layer. This connection represented by specific weights to determine the relationship with is donated by W_{kj}. The main duty of the hidden layer nodes is to solve the complexity of the problem being modeled. The hidden layer nodes comprise the activation function that handles nonlinearly transforming of the inputs parameters into an alternative space where the training samples are linearly separable [46]. The connection between the hidden layer and the output layer (WQI targeted) exhibited by W_{lk}. b_k and $b1$ factors denote the bias of the corresponding hidden/output layer neurons, respectively. They imply a percentage of weight to shift the activation function. x_j and $y1$ indicate the input/out variables, respectively. They can be expressed by the following equations:

$$y1 = f1\left[\sum_{K=1}^{K} w_{lk} f2\left(\sum_{J=1}^{J} w_{kj} x_j + b_k\right) + b1\right] \qquad (2)$$

$$f2(p) = \frac{2}{(1 + e^{-2p})} - 1 \qquad (3)$$

where $f1(.)$ and $f2(.)$ represent the linear and tansigmoidal activation functions of the algorithm. In general, the choice of the transfer function in the hidden layer and output layer is very significant task because it can effect essentially on the target of the model. Bayesian Regulation algorithm has been employed to train the ANN modeling because it updates the weight and bias values according to Levenberg–Marquardt optimization and save time, in addition, its ability to reduce the combination of mean-squared errors that caused by using huge number of data.

3.2.2 *Radial basis function (RBF)*

ANNs are flexible processing approaches that have the efficiency to mimic the relationship between the inputs and output data sets, in addition to pattern recognition capability. Radial basis function algorithm is the remarkable efficient approach used in various engineering applications [21–23, 47–49]. Initially, RBF was introduced by Lowe and Broomhead [50]. The architecture of RBF neural network is similar to the BPNN structure, except it has a radial basis

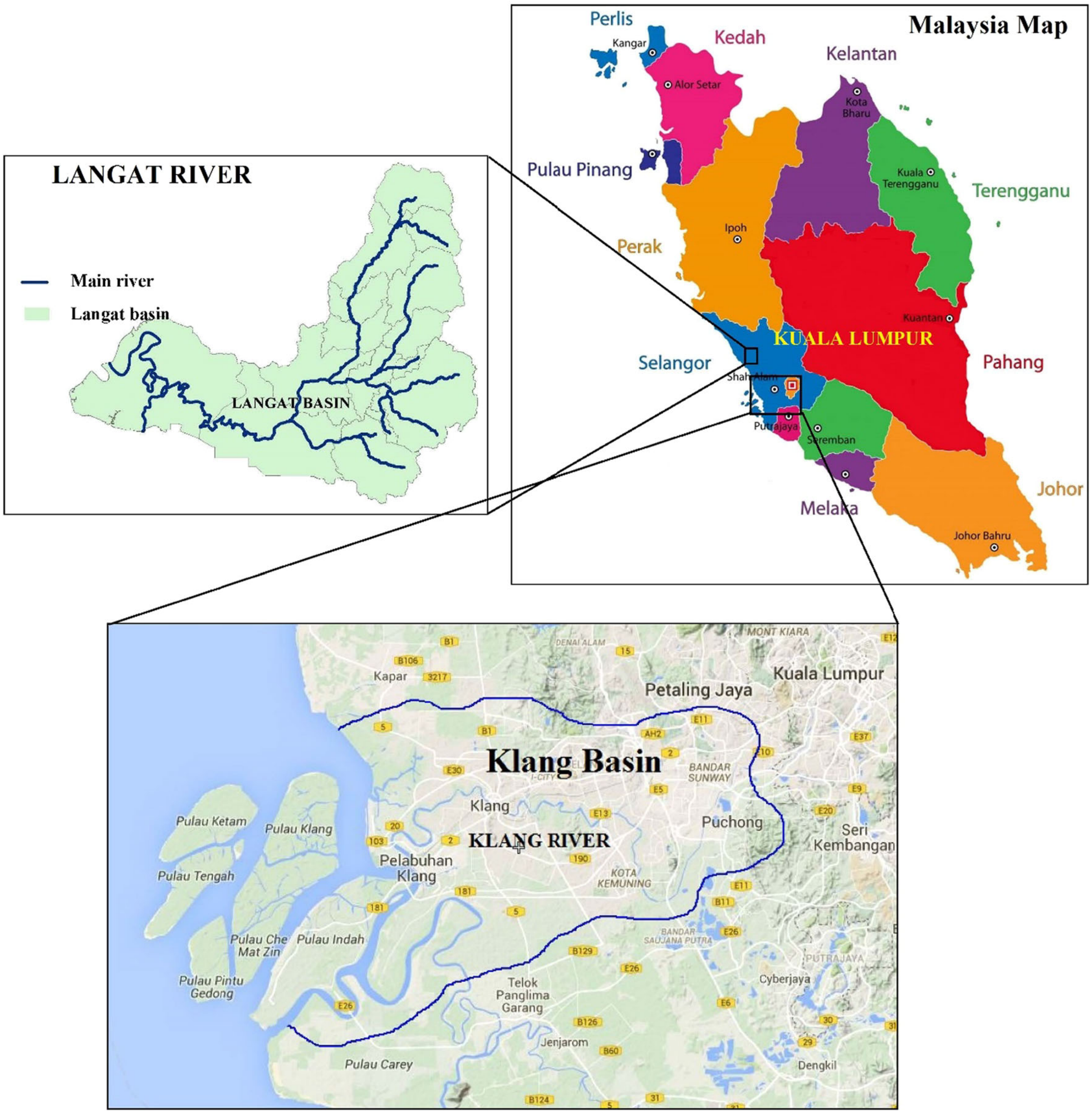

Fig. 1 Location of the two rivers, Langat and Klang River Basins, Malaysia

activation function in the hidden layer that to characterize the partitions of the input space. Here, RBF implies an approximation between the input/output phases with linear combination of the radial basis function. The main advantage of RBF is response by (decreasing or increasing) monotonically with distance from a central point. Figure 3 illustrates the architecture of the RBF neural network algorithm. What is necessary to highlight here, the middle layer connects by linear function with output layer. In addition, the radial basis functions $\varphi 1$, $\varphi 2$,… φN are known as hidden functions while $[\{\varphi_i\,(x)\}N_i = 1]$ is called the hidden space. Gaussian function used in this study as a radial basis function, and the following formula can use to express the one-dimensional representation of Gaussian function:

$$\varphi(x,\mu) = e^{\frac{x-\mu^2}{2d^2}} \tag{4}$$

where μ is the center of the Gaussian function (mean value of x) and d is the distance (radius) from the center of $\varphi(x, \mu)$, which gives a measure of the spread of the Gaussian curve. During the training processes of RBF, the center μ and the spread d are determined. Figure 4 reveals the Gaussian radial basis function manners, with large value of d produce less sensitivity of the used function to the supplied data in the input phase. Radial basis functions number inside the hidden layer subjected to the complexity of the problem nor on the size of the data set, where it can appear in the utilization of multi-layer perceptron algorithm [51–53].

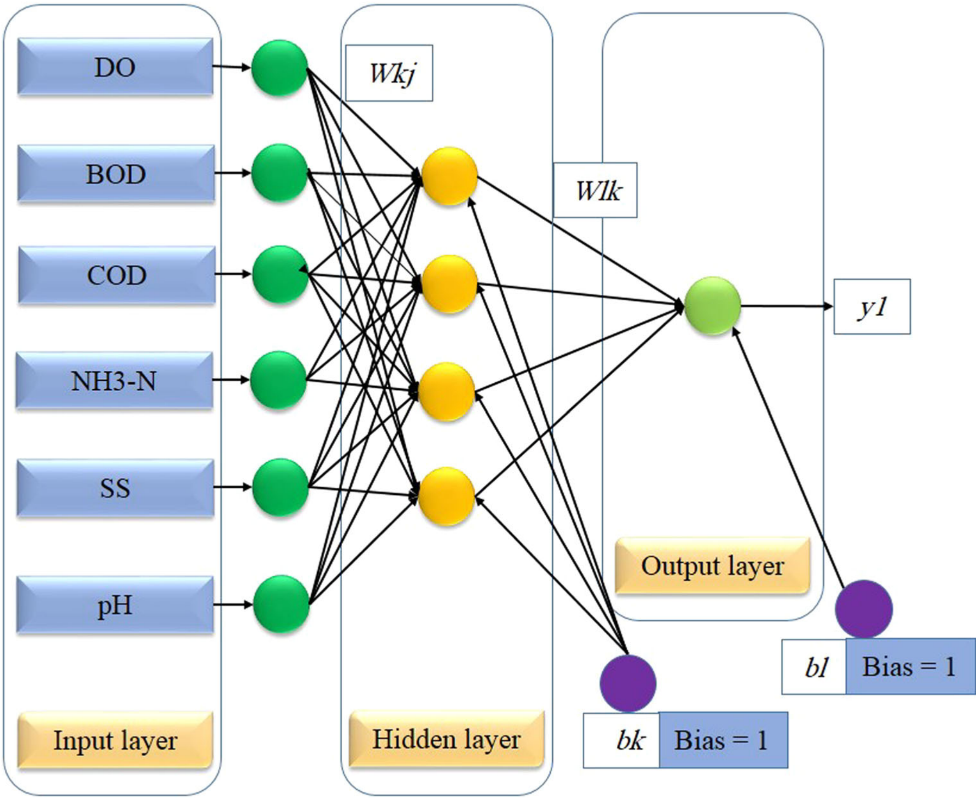

Fig. 2 Three layer of feed forward artificial neural network back propagation configuration

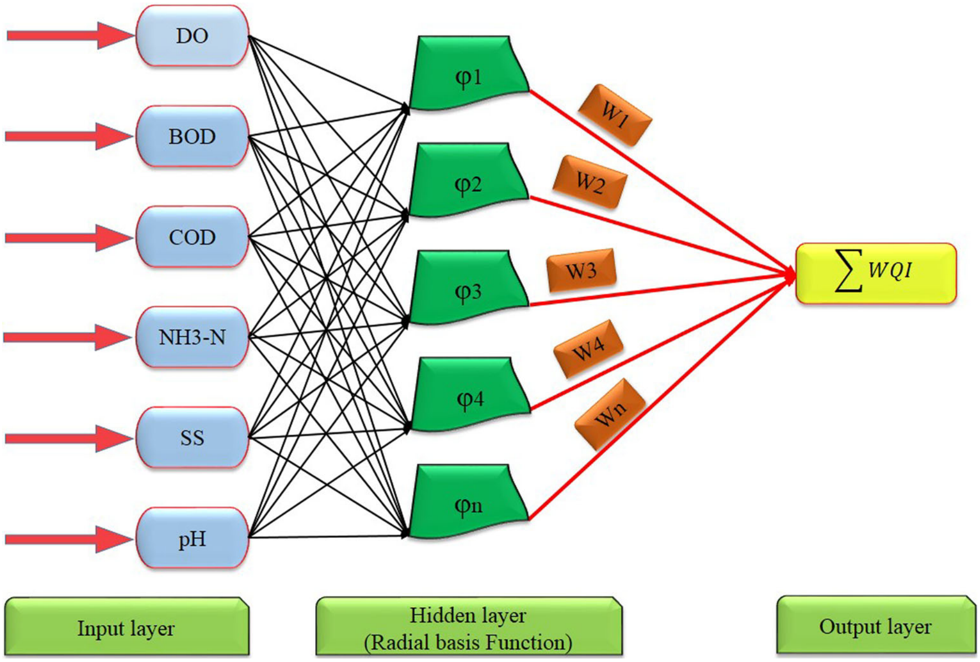

Fig. 3 Structure of radial basis function algorithm

3.2.3 Model configuration development

In this research, water quality index has been predicted for tropical environment, Langat River and Klang River, Malaysia. Water quality parameters (i.e., DO, BOD, COD, NH_3-N, SS, and pH) used as a predictors for WQI, as clearly shown in Figs. 2 and 4. At the initial stage, input of the experiment, input variables, and output variable are normalized linearly in the range of 0.1 and 0.9. The normalization is done according to the following formula:

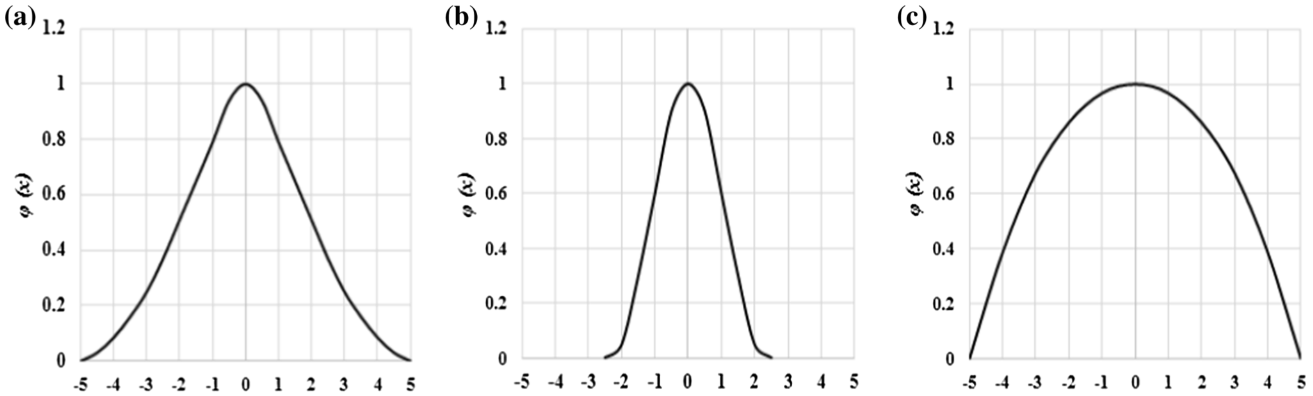

Fig. 4 Radial bases function with different stages of spread value. **a** Normal spread, **b** small spread, **c** large spread

$$X_{\text{new}} = 0.8\frac{X - X_{\min}}{X_{\max} - X_{\min}} + 0.1 \tag{5}$$

where X_{new} is the normalized value of an original parameter, X is the original data point, $X_{\min}$ and $X_{\max}$ are the minimum and maximum values in the data set, respectively. The normalization is done to ensure that the minimum value in the data set is normalized to 0.1, and the maximum value is normalized to 0.9. This pattern of normalizing the data is chosen because it tends to provide a better outcome on the water quality application. In addition, the variables are usually measured in different units, by normalizing the variables and recasting them in dimensionless units, the arbitrary effect of similarity between objects is removed [54].

3.2.4 Model performance criteria

In order to evaluate the prediction modeling accuracy, four criteria were used for comparative evaluation of the performance of the modeling. A multi-criteria approach was adopted for assessing the models developed, in which model performance was evaluated using several statistical error and goodness-of-fit measures, including the root-mean-squared error (RMSE), coefficient of determination (R^2), Nash–Sutcliffe coefficient (NE), and relative error (RE%). The performances criteria formulas are presented as:

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum\nolimits_{t=1}^{n}\left(\text{WQI}_a - \text{WQI}_P\right)^2} \tag{6}$$

$$R^2 = \frac{\sum_{i=1}^{n}\left[\left(\text{WQI}_a - \overline{\text{WQI}}_a\right) * \left(\text{WQI}_p - \overline{\text{WQI}}_p\right)\right]}{\sqrt{\sum_{i=1}^{n}\left(\text{WQI}_a - \overline{\text{WQI}}_a\right)^2 \sum_{i=1}^{n}\left(\text{WQI}_p - \overline{\text{WQI}}_p\right)^2}} \tag{7}$$

$$\text{NE} = 1 - \frac{\sum_{t=1}^{n}\left(\text{WQI}_a - \text{WQI}_p\right)^2}{\sum_{t=1}^{n}\left(\text{WQI}_a - \overline{\text{WQI}}_a\right)^2} \tag{8}$$

$$\%\text{RE} = \frac{\text{WQI}_a - \text{WQI}_p}{\text{WQI}_a} * 100 \tag{9}$$

where WQI_a is the observed value of water quality index, WQI_p is the predicted value of water quality index and N number of samples. Furthermore, sensitivity data analysis has been investigated. Sensitivity analysis indicator is a method to extract the cause and effect relationship between input/output phases. The principal concept of this approach is that the inputs to the network are shifted slightly and the corresponding change in the output is reported either as a percentage or a raw difference.

4 Application and discussion

The current research employed two different ANN algorithms BPNN and RBFNN. Initially, the number of neurons for the input layer fixed six (number of the water quality parameters; DO, BOD, COD, NH_3-N, SS, and pH, for the first scenario) and five neurons for the second scenario with rolling out the BOD parameter. On the other hand, one neuron in the output layer presents the WQI. The monthly data of water quality parameters (2001–2010) were normalized within the range of 0.1–0.9 in order to accelerate the training procedure and achieve minimum (MSE). Then, the optimal data division scheme was conducted using NeuroSolutions 7 software, 80 % for training the network and 20 % for testing the network, based on trial-and-error procedure. However, one of the most critical features of BPNN approach is determining the number of neurons in the hidden layer. An insufficient number of neurons are utilized, and the model will be unable to capture the complexity of the relationship between input and output phases. Thus, in this research, a trial-and-error procedure used to determine the optimal hidden layer neurons. The trial-and-error procedure started with two hidden neurons initially, and the number of hidden neurons was increased up to 10 with a step size of 2 in each trial. The optimal architecture was determined based on three performance indices R^2, RMSE, and NE (see Table 2). In the RBFNN approach, different spread values were utilized to achieve the best performance indices.

4.1 First scenario: six parameters water quality modeling the WQI

In this scenario, prediction model was built based on the six water quality variables (i.e., DO, BOD, COD, NH_3-N, SS, and pH). The best performances evaluation criteria for the first scenario using BPNN algorithm have been presented in Table 2. According to Table 2, it was determined that six neurons in the input layer (number of water quality parameters), eight neurons in the hidden layer, and one neuron in the output layer (WQI as a target) performed the best R^2, RMSE, and NE (0.7472, 0.0699 and 0.7701, respectively), whereas according to Table 3, the first scenario using RBFNN model showed the best performances evaluation criteria of R^2, RMSE, and NE (0.9872, 0.0157, and 0.9871, respectively) with spread value 0.8. In order to validate the functionality of the ANN models, Figs. 5 and 6 indicate the scatterplots between the observed and the predicted value of WQI for the testing data set (2009–2010), for the both scenarios. The BPNN model performed poorly comparing with RBFNN model (see Figs. 5, 6). This observation is clearly explaining the capability of the radial basis function neural network in capturing the nonlinearity between the water quality predictors and the predicted (WQI).

To have more assessment of the performances of the both models, the RE computed for the both models. The (%RE) distribution as computed by the formula 9 is illustrated in Figs. 7 and 8 for BPNN and RBFNN models, respectively. As it can be seen, BPNN model did not perform very well for the first scenario (see Fig. 7a). The percentage of the error residual exceeds the 45 % for more than 14 % of the testing data set and limited between −35 and +15 for more than 65 % of testing data set. The RBFNN model performed very well for the first scenario (see Fig. 7b). Less than 5 % of the testing data set performed error distribution between −20 and +20, while, there are more than 85 % of the testing data set limited the distribution error of the RE between −8 and +7. This is return back to the Gaussian radial basis function, which is capable to mimic the pattern phenomena of the water quality parameters that influenced the WQI.

In the systematic analysis, the influence of each water quality parameter was investigated using the sensitivity analysis procedure. Figure 9 presented the sensitivity analysis for the best ANN algorithm, which is RBFNN modeling. According to Fig. 9a, it can be observed that the DO is the most effective parameter correlated with the predicted WQI, around 28 %. The observation based on this percentage reveals that the sensitivity analysis of this parameter was harmonized with manual calculated WQI. Furthermore, the nature of Langat and Klang rivers is characterized by polluted environment, which indicated that the dissolved oxygen is the most effected parameter in the aquatic system of the rivers. The followed parameters percentages are BOD 21 %, COD 18 %, SS 15 %, NH_3-N 9 %, and pH 9 %. Accordingly, the results modeling of sensitivity analysis briefed the significant of each variable to WQI.

4.2 Second scenario: five parameters water quality modeling the WQI

The main objective of this scenario is to reduce the water quality parameters that needed to predict WQI without much loss of information; this is for the reason that BOD parameter requires a high controlled laboratory conditions, needs time and costly tested parameter. Therefore, the five parameters (DO, COD, NH_3-N, SS, and pH) used as predictors for the WQI. The accuracy of the models examined using the formal quantitative measures of accuracy including R^2, RMSE, and NE. Table 2 presents the performance criteria for the second scenario using BPNN model. The BPNN model performed the best results of R^2, RMSE, and NE (0.7007, 0.0867, and 0.7164, respectively) with five neurons input layer, six neurons hidden layer, and one neuron in the output layer, while RBFNN model performance criteria for the second scenario is tabulated in Table 3. The best evaluation criteria of R^2, RMSE, and NE

Table 2 BPNN model performances assessment for both scenarios

Models	First scenario				Second scenario			
	Architecture	R^2	RMSE	NE	Architecture	R^2	RMSE	NE
M1	6-2-1	0.6853	0.0983	0.7156	5-2-1	0.7001	0.0981	0.7151
M2	6-4-1	0.7409	0.0734	0.7631	5-4-1	0.6976	0.0928	0.7149
M3	6-6-1	0.7309	0.0801	0.7592	5-6-1	0.7007	0.0867	0.7164
M4	6-8-1	0.7472	0.0699	0.7701	5-8-1	0.6593	0.1018	0.6828
M5	6-10-1	0.7267	0.0783	0.7434	5-10-1	0.5109	0.1314	0.5372

Table 3 RBFNN model performances assessment for both scenarios

Models	First scenario				Second scenario			
	Spread values	R^2	RMSE	NE	Spread values	R^2	RMSE	NE
M1	0.2	0.9552	0.0294	0.9547	0.2	0.9228	0.0405	0.9140
M2	0.4	0.9242	0.0399	0.9165	0.4	0.9228	0.0405	0.9140
M3	0.6	0.9852	0.0168	0.9852	0.6	0.9807	0.0194	0.9803
M4	0.8	0.9872	0.0157	0.9871	0.8	0.9705	0.0247	0.9678
M5	1.0	0.9820	0.0186	0.9819	1.0	0.8999	0.0452	0.8926

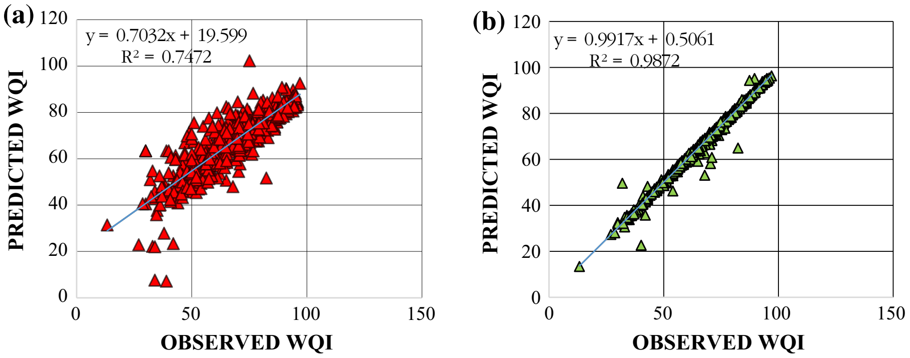

Fig. 5 First scenario scatterplot between the observed and the predicted value of WQI. **a** BPNN model using six variables of water quality, **b** RBFNN model using six variables of water quality

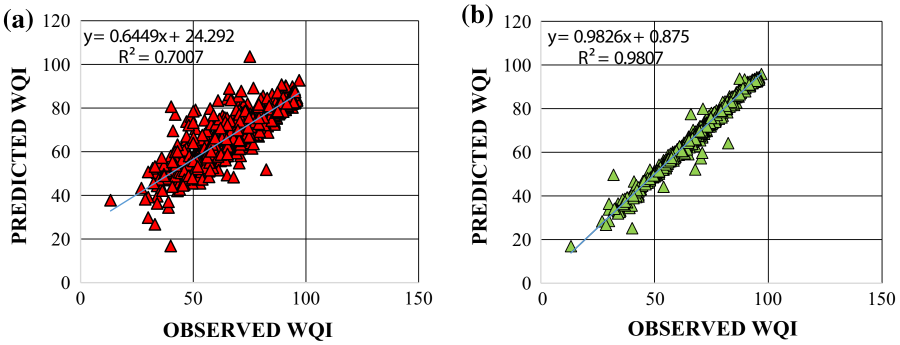

Fig. 6 Second scenario scatterplot between the observed and the predicted value of WQI. **a** BPNN model using five variables of water quality, **b** RBFNN model using five variables of water quality

are 0.9807, 0.0194, and 0.9803, respectively, with spread value 0.6.

Figure 6a shows the correlation of coefficient (R^2), which provides information for linear dependence between observed and predicted WQI using BPNN model. Similar to the first scenario, the scatter plot performed less accuracy comparatively with RBFNN model. However, Fig. 6b indicates that RBFNN model falls close to the ideal line for the scatterplot. Eventually, the result indicated that the RBFNN model is more reliable and suitable in prediction WQI than BPNN model because it could capture the pattern behavior between the predictors and the predictand. Furthermore, it is obvious that RBFNN algorithm has predicted WQI accurately and with a satisfactory acceptable performance even without using BOD parameter. Nevertheless, this study clearly shows that it is possible to reduce the number of the environmental parameters needed to make the water quality prediction without losses much information and still have an acceptable performance.

To further analyze the effectiveness of the modeling, the RE is demonstrated in Fig. 8a, d (BPNN and RBFNN models, respectively). The divergence between the actual and the predicted testing data exceeds (+65 and −65 %) 10 % of the data, while the RE reaches a level of −40 % for more than 38 % of the testing data. On the other hand, RBFNN model accomplished the RE percentage with very well divergence between +5 and −5 % for more than 87 % of the testing data set. Finally, for a better understanding of the above observations, further analysis was accomplished to study the sensitivity analysis of each variable. The sensitivity analysis is the influence of each variable has been investigated. Indeed, this indicator showed up to what extend each input variables correlated with the predicted (target output). The percentage of the

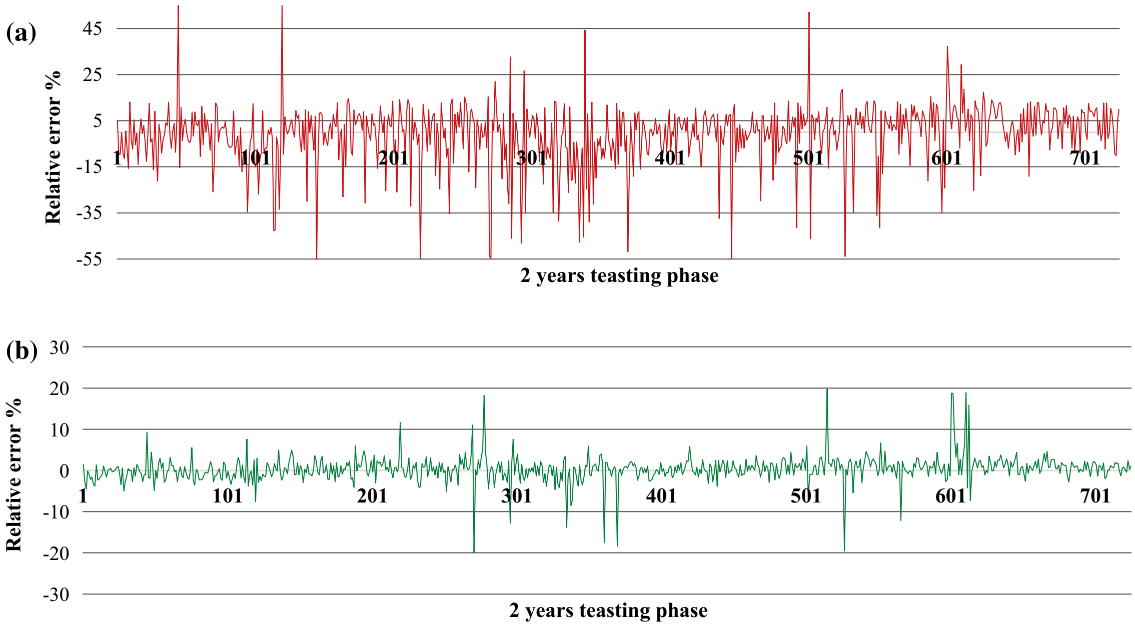

Fig. 7 First scenario relative error distribution of the water quality index modeling, WQI. **a** BPNN algorithm using six variables of water quality, **b** RBFNN algorithm using six variables of water quality

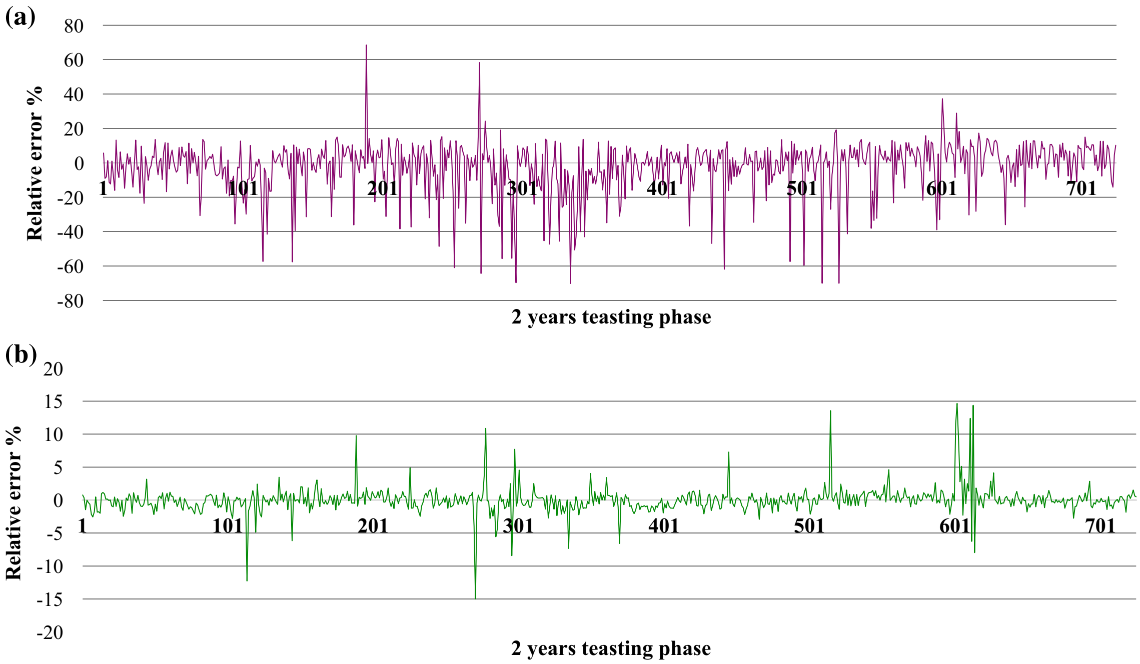

Fig. 8 Second scenario relative error distribution of the water quality index modeling, WQI. **a** BPNN algorithm using five variables of water quality, **b** RBFNN algorithm using five variables of water quality

association of the BOD almost divided equally to the other water quality parameters. DO influenced the WQI prediction by 35 %, COD 23 %, SS 17 %, NH_3-N 13 %, and pH 12 % (see Fig. 9b).

In the light of the application and analysis, machine learning approaches produce a versatile modeling in predicting water quality index. Based on this observation, the integration of soft computing with modeling water quality is a very useful application that could enhance the environmental engineering field in term of online implementation monitoring and accurate prediction. However, better and more proficient use of the database

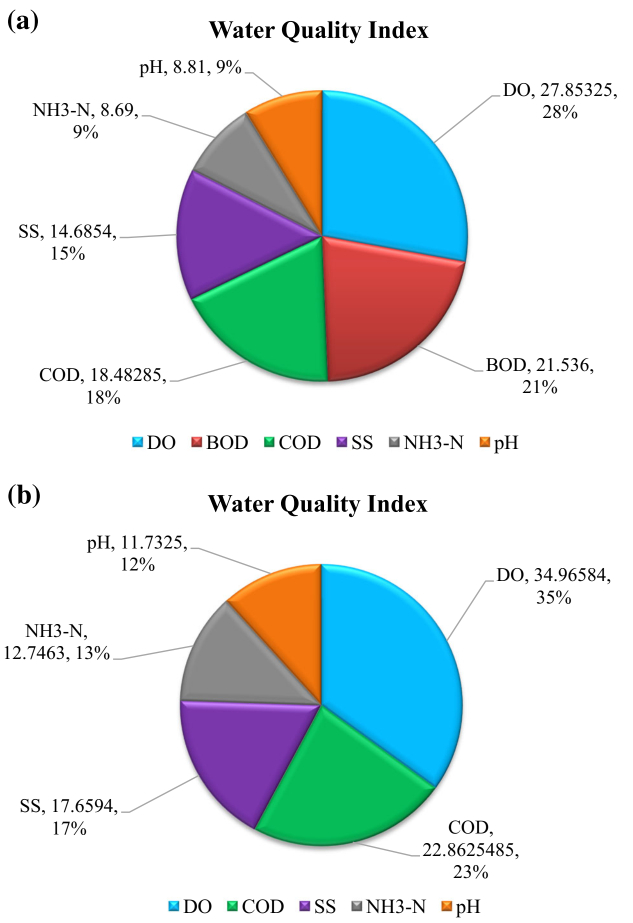

Fig. 9 Sensitivity analysis percentage of each water quality variables on WQI prediction. **a** RBFNN modeling for the first scenario, **b** RBFNN modeling for the second scenario

administration frameworks, graphical presentations, and learning obtaining modules will upgrade the applicability of modeling systems in real practice.

5 Conclusion

This manuscript described the implementation of ANN to predict water quality index in tropical environment based on monthly measurement of water quality variables. The monthly data (i.e., DO, BOD, COD, NH_3-N, SS, and pH) for 10 years' time period (2001–2010) were selected for this analysis. Two different models have been examined, namely feed forward back propagation neural network (BPNN) and radial basis function neural network (RBFNN). Results obtained from this research are positively encouraging with high performances accuracy. Nevertheless, the following conclusions can be drawn.

1. Generally, ANN employed effectively in predicting water quality index in tropical environment, Langat and Klang river basins, Malaysia.
2. The current research emphasizes that ANN modeling, in particular RBFNN model constitutes an effective tool for assessment of river water quality that simplifies the computing of WQI.
3. RBFNN performed very well in recognizing the most associated input variables on prediction WQI, for the both scenarios.
4. It can be seen from the summary of the results that there was a significant improvement in the modeling performance when the RBFNN model was used in the both scenarios, compared to using the BPNN model. This is due to the fact that Gaussian radial basis function, which is capable to mimic the pattern phenomena chaotic disturbances, complex nonlinear dynamics, and randomness of the water quality parameters that influenced the WQI.

References

1. Kotti ME, Vlessidis AG, Thanasoulias NC, Evmiridis NP (2005) Assessment of river water quality in Northwestern Greece. Water Resour Manag 19:77–94. doi:10.1007/s11269-005-0294-z
2. Niemi GJ, DeVore P, Detenbeck N et al (1990) Overview of case studies on recovery of aquatic systems from disturbance. Environ Manage 14:571–587. doi:10.1007/BF02394710
3. Najah A, Elshafie A, Karim OA, Jaffar O (2009) Prediction of Johor River water quality parameters using artificial neural networks. Eur J Sci Res 28:422–435
4. Fahmi M, Nasir M, Samsudin MS et al (2011) River water quality modeling using combined principle component analysis (PCA) and multiple linear regressions (MLR): a case study at Klang River, Malaysia Department of Environmental Sciences, Faculty of Environmental Studies, Department of Environment. World Appl Sci J 14:73–82
5. Zali MA, Retnam A, Juahir H et al (2011) Sensitivity analysis for water quality index (WQI) prediction for Kinta River, Malaysia. World Appl Sci J 14:60–65
6. Behboudian S, Tabesh M, Falahnezhad M, Ghavanini FA (2014) A long-term prediction of domestic water demand using preprocessing in artificial neural network. J Water Supply Res Technol 63:31. doi:10.2166/aqua.2013.085
7. Caselli M, Trizio L, Gennaro G, Ielpo P (2009) A simple feedforward neural network for the PM10 forecasting: comparison with a radial basis function network and a multivariate linear regression model. Water Air Soil Pollut 201:365–377. doi:10.1007/s11270-008-9950-2
8. Albergaria JT, Martins FG, Alvim-Ferraz MCM, Delerue-Matos C (2014) Multiple linear regression and artificial neural networks to predict time and efficiency of soil vapor extraction. Water Air Soil Pollut. doi:10.1007/s11270-014-2058-y
9. Modarres R (2008) Multi-criteria validation of artificial neural network rainfall-runoff modeling. Hydrol Earth Syst Sci Discuss 5:3449–3477. doi:10.5194/hessd-5-3449-2008

10. Shrestha RR, Theobald S, Nestmann F (2005) Simulation of flood flow in a river system using artificial neural networks. Hydrol Earth Syst Sci 9:313–321. doi:10.5194/hess-9-313-2005
11. Mwale FD, Adeloye AJ, Rustum R (2014) Application of self-organising maps and multi-layer perceptron-artificial neural networks for streamflow and water level forecasting in data-poor catchments: the case of the Lower Shire floodplain, Malawi. Hydrol Res 45:838. doi:10.2166/nh.2014.168
12. Cheng J, Li QS (2008) Reliability analysis of structures using artificial neural network based genetic algorithms. Comput Methods Appl Mech Eng 197:3742–3750. doi:10.1016/j.cma.2008.02.026
13. Ahmad A, El-Shafie A, Mohd Razali SF, Mohamad ZS (2014) Reservoir optimization in water resources: a review. Water Resour Manag 28:3391–3405. doi:10.1007/s11269-014-0700-5
14. Hossain MS, El-shafie A (2013) Intelligent systems in optimizing reservoir operation policy: a review. Water Resour Manag 27:3387–3407. doi:10.1007/s11269-013-0353-9
15. Chau KW (2006) A review on integration of artificial intelligence into water quality modelling. Mar Pollut Bull 52:726–733. doi:10.1016/j.marpolbul.2006.04.003
16. Rahim NA, Ahmad Z (2013) Features selection in water quality prediction in neural network using canonical correspondence analysis (CCA). The 6th international conference on process systems engineering (PSE ASIA), pp 25–27
17. Hipni A, El-shafie A, Najah A et al (2013) Daily forecasting of dam water levels: comparing a support vector machine (SVM) model with adaptive neuro fuzzy inference system (ANFIS). Water Resour Manag 27:3803–3823. doi:10.1007/s11269-013-0382-4
18. Kazemi Yazdi S, Scholz M (2010) Assessing storm water detention systems treating road runoff with an artificial neural network predicting fecal indicator organisms. Water Air Soil Pollut 206:35–47. doi:10.1007/s11270-009-0084-y
19. Ye J, Zhang P, Hoffmann E et al (2014) Comparison of response surface methodology and artificial neural network in optimization and prediction of acid activation of bauxsol for phosphorus adsorption. Water Air Soil Pollut. doi:10.1007/s11270-014-2225-1
20. Lesven L, Lourino-Cabana B, Billon G et al (2009) Water-quality diagnosis and metal distribution in a strongly polluted zone of Deûle River (Northern France). Water Air Soil Pollut 198:31–44. doi:10.1007/s11270-008-9823-8
21. Afan HA, El-Shafie A, Yaseen ZM et al (2014) ANN based sediment prediction model utilizing different input scenarios. Water Resour Manag 29:1231–1245. doi:10.1007/s11269-014-0870-1
22. Yaseen ZM, El-Shafie A, Afan HA et al (2015) RBFNN versus FFNN for daily river flow forecasting at Johor River. Neural Comput Appl, Malaysia. doi:10.1007/s00521-015-1952-6
23. El-Shafie A, Abdin AE, Noureldin A, Taha MR (2009) Enhancing inflow forecasting model at Aswan high dam utilizing radial basis neural network and upstream monitoring stations measurements. Water Resour Manag 23:2289–2315. doi:10.1007/s11269-008-9382-1
24. Sulaiman M, El-Shafie A, Karim O, Basri H (2011) Improved water level forecasting performance by using optimal steepness coefficients in an artificial neural network. Water Resour Manag 25:2525–2541
25. Chen Q, Mynett AE (2003) Integration of data mining techniques and heuristic knowledge in fuzzy logic modelling of eutrophication in Taihu Lake. Ecol Model 162:55–67. doi:10.1016/S0304-3800(02)00389-7
26. 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. doi:10.1016/j.jhydrol.2011.05.024
27. Najah A, Karim OA, Jaafar O, El-shafie AH (2011) An application of different artificial intelligences techniques for water quality prediction. Int J Phys Sci 6:5298–5308. doi:10.5897/IJPS11.1180
28. Amiri BJ, Nakane K (2009) Comparative prediction of stream water total nitrogen from land cover using artificial neural network and multiple linear regression approaches. Pol J Environ Stud 18(2):151–160
29. Chebud Y, Naja GM, Rivero RG, Melesse AM (2012) Water quality monitoring using remote sensing and an artificial neural network. Water Air Soil Pollut 223:4875–4887. doi:10.1007/s11270-012-1243-0
30. Lee JHW, Huang Y, Dickman M, Jayawardena AW (2003) Neural network modelling of coastal algal blooms. Ecol Model 159:179–201. doi:10.1016/s0304-3800(02)00281-8
31. Palani S, Liong SY, Tkalich P (2008) An ANN application for water quality forecasting. Mar Pollut Bull 56:1586–1597. doi:10.1016/j.marpolbul.2008.05.021
32. Palani S, Liong S, Tkalich P, Palanichamy J (2009) Development of a neural network model for dissolved oxygen in seawater. Indian J Geo-Mar Sci 38:151–159
33. Muttil N, Chau K-W (2006) Neural network and genetic programming for modelling coastal algal blooms. Int J Environ Pollut 28:223–238. doi:10.1504/IJEP.2006.011208
34. Lek S, Guégan JF (1999) Artificial neural networks as a tool in ecological modelling, an introduction. Ecol Model 120:65–73. doi:10.1016/S0304-3800(99)00092-7
35. Zou R, Lung W-S, Guo H (2002) Neural network embedded Monte Carlo approach for water quality modeling under input information uncertainty. J Comput Civ Eng 16:135–142. doi:10.1061/(ASCE)0887-3801(2002)16:2(135)
36. Kralisch S, Fink M, Flügel W-A, Beckstein C (2003) A neural network approach for the optimisation of watershed management. Environ Model Softw 18:815–823. doi:10.1016/S1364-8152(03)00081-1
37. Maier HR, Morgan N, Chow CWK (2004) Use of artificial neural networks for predicting optimal alum doses and treated water quality parameters. Environ Model Softw 19:485–494. doi:10.1016/S1364-8152(03)00163-4
38. Diamantopoulou MJ, Papamichail DM, Antonopoulos VZ (2005) The use of a neural network technique for the prediction of water quality parameters. Oper Res 5:115–125. doi:10.1007/BF02944165
39. Elhatip H, Kömür MA (2008) Evaluation of water quality parameters for the Mamasin Dam in Aksaray City in the central Anatolian part of Turkey by means of artificial neural networks. Environ Geol 53:1157–1164. doi:10.1007/s00254-007-0705-y
40. 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. doi:10.1016/j.ecolmodel.2009.01.004
41. Dong Y, Scholz M, Harrington R (2012) statistical modeling of contaminants removal in mature integrated constructed wetland sediments. J Environ Eng 138:1009–1017. doi:10.1061/(ASCE)EE.1943-7870.0000572
42. Wang L, Li X, Cui W (2012) Fuzzy neural networks enhanced evaluation of wetland surface water quality. Int J Comput Appl Technol 44:235. doi:10.1504/IJCAT.2012.049087
43. Niroobakhsh M (2012) Prediction of water quality parameter in Jajrood River basin: application of multi layer perceptron (MLP) perceptron and radial basis function networks of artificial neural networks (ANNs). Afr J Agric Res 7:4131–4139. doi:10.5897/AJAR11.1645
44. Department of Environment (2005) Malaysia environmental quality report. Petaling Jaya, Malaysia, 2007

45. Maier HR, Dandy GC (2000) Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications. Environ Model Softw 15:101–124. doi:10.1016/S1364-8152(99)00007-9
46. Brown M, Harris CJ (1995) A perspective and critique of adaptive neurofuzzy systems used for modelling and control applications. Int J Neural Syst 6:197–220
47. Lin G-F, Chen L-H (2004) A non-linear rainfall-runoff model using radial basis function network. J Hydrol 289:1–8. doi:10.1016/j.jhydrol.2003.10.015
48. Ma L, Xin K, Liu S (2008) Using radial basis function neural networks to calibrate water quality model. World Acad Sci Eng Technol Int J Environ Chem Ecol Geol Geophys Eng 2(2):9–17
49. Beckert A, Wendland H (2001) Multivariate interpolation for fluid–structure-interaction problems using radial basis functions. Aerosp Sci Technol 5:125–134. doi:10.1016/S1270-9638(00)01087-7
50. Lowe D, Broomhead D (1988) Multivariable functional interpolation and adaptive networks. Complex Syst 2:321–355
51. Bishop CM (1995) Neural networks for pattern recognition. J Am Stat Assoc. doi:10.2307/2965437
52. Ripley BD (1996) Pattern recognition and neural networks. Cambridge University Press, Cambridge
53. Haykin S (1994) Neural networks: a comprehensive foundation. Macmillan, Englewood Clifs, NJ
54. El-shafie A, Mukhlisin M, Najah AA, Taha MR (2011) Performance of artificial neural network and regression techniques for rainfall-runoff prediction. Int J Phys Sci 6:1997–2003. doi:10.5897/IJPS11.314