Machine Learning Methods for Better Water Quality Prediction

Ali Najah Ahmed[1], Faridah Binti Othman [2], Haitham Abdulmohsin Afan[2,], Rusul Khaleel Ibrahim[2,*], Chow Ming Fai[1], Md Shabbir Hossain[3], Mohammad Ehteram [4] and Ahmed Elshafie[2]

[1] Institute of Energy Infrastructure (IEI), Universiti Tenaga Nasional, 43000 Selangor, Malaysia.
[2] Department of Civil Engineering, Faculty of Engineering, University Malaya, Malaysia.
[3] Department of civil engineering, Heriot-Watt University, 62200 Putrajaya, Malaysia.
[4] Department of Water Engineering and Hydraulic Structures, Faculty of Civil Engineering, Semnan University, Semnan 35131-19111, Iran.
* Correspondence: ruaw119.j@gmail.com

Abstract

In any aquatic system analysis, the modelling water quality parameters are of considerable significance. The traditional modelling methodologies are dependent on datasets that involve large amount of unknown or unspecified input data and generally consist of time-consuming processes. The implementation of artificial intelligence (AI) leads to a flexible mathematical structure that has the capability to identify non-linear and complex relationships between input and output data. There has been a major degradation of the Johor River Basin because of several developmental and human activities. Therefore, setting up of a water quality prediction model for better water resource management is of critical importance and will serve as a powerful tool. The different modelling approaches that have been implemented include: Adaptive Neuro-Fuzzy Inference System (ANFIS), Radial Basis Function Neural Networks (RBF-ANN), and Multi-Layer Perceptron Neural Networks (MLP-ANN). However, data obtained from monitoring stations and experiments are possibly polluted by noise signals as a result of random and systematic errors. Due to the presence of noise in the data, it is relatively difficult to make an accurate prediction. Hence, a Neuro-Fuzzy Inference System (WDT-ANFIS) based augmented wavelet de-noising technique has been recommended that depends on historical data of the water quality parameter. In the domain of interests, the water quality parameters primarily include ammoniacal nitrogen (AN), suspended solid (SS) and pH. In order to evaluate the impacts on the model, three evaluation techniques or assessment processes have been used. The first assessment process is dependent on the partitioning of the neural network connection weights that ascertains the significance of every input parameter in the network. On the other hand, the second and third assessment processes ascertain the most effectual input that has the potential to construct the models using a single and a combination of parameters, respectively. During these processes, two scenarios were introduced: Scenario 1 and Scenario 2. Scenario 1 constructs a prediction model for water quality parameters at every station, while Scenario 2 develops a prediction model on the basis of the value of the same parameter at the previous station (upstream). Both the scenarios are based on the value of the twelve input parameters. The field data from 2009–2010 was used to validate WDT-ANFIS. The WDT-ANFIS model exhibited a significant improvement in predicting accuracy for all the water quality parameters and outperformed all the recommended models. Also, the performance of Scenario 2 was observed to be more adequate than Scenario 1, with substantial improvement in the range of 0.5% to 5% for all the water quality parameters at all stations. On validating the recommended model, it was found that the model satisfactorily predicted all the water quality parameters (R^2 values equal or bigger than 0.9).

Keywords: water quality parameters; machine learning; WDT-ANFIS.

1. Introduction

Rivers are considered as one of the most critical sources of water for irrigation purposes, industrial needs and other uses. The dynamic nature of the river systems and their easy accessibility for waste disposal make the river systems most vulnerable to the adverse effects of environmental pollution. The term "water quality" refers to the state or condition of water, which takes into account the physical, chemical, and biological properties of the water. In conducting the study of any aquatic system, modelling the water quality parameters is of utmost significance. Evaluation and prediction of the surface water quality is necessary for effective management of river basins so that sufficient measures can be adopted to ensure that the pollution levels remain within permissible limits. Accurate prediction of future phenomena in relation to the water quality is the essence of optimal water resources management. The conventional process-based modelling methods offer comparatively accurate predictions for water quality parameters. However, these models have limitations as they depend on data sets that require a substantial amount of processing time and a huge amount of input data that is often unknown.

Nearly 60% of the major rivers in Malaysia are used for agricultural, household and industrial applications (DID, 2000). As per Rosnani Ibrahim (Ibrahim, 2001), the major sources of pollution that affect these rivers are dumping of sewage, waste releases from medium and small-sized industries not having proper waste matter treatment equipment, clearing of land and groundwork activities. On the basis of the records of 1999, 50 catchments (that is 42% of river) were contaminated with SS (suspended solids) caused by badly planned and unregulated earth clearing attempts and 33 catchments (that is, 28% of river) were polluted with AN (ammoniacal nitrogen) from activities related to cattle breeding and household sewage dumping.

Johor River is regarded as somewhat polluted as per DOE (Department of Environment)(DOE, 2007) because of the developmental activities alongside the bank of the river. Moreover, the river continues to be chocked and dumped by waste and litter due to lack of enforcement by the local administration. These pollutants ultimately end up in the Joho River tributaries, rich areas for nourishment and breeding of poultry and fish. Consequently, several statistical frameworks and computer simulations must be introduced as powerful and critical tools for planning and monitoring the maintenance of the water bodies.

Growing concerns regarding environment, along with scarce funding, are giving rise to a growing interest in cost-effective and judicious strategies for the management of water quality. Since the quality of water directly affects the health of the humans, quality improvement of the water accessible for human use will play a significant role in decreasing health related hazards.

The project of water pollution regulation is based on the management of water quality. It estimates the kind of water quality from the present water quality condition, as well as from the rules of disposal of the pollutants into the river. Moreover, many models for water quality, like stochastic and deterministic models, have been created so as to provide best processes to conserve the quality of water (Hull et al., 2008). Nevertheless, getting efficient and precise water quality model in complex water resources is still difficult because of the variations and complications in the actual world, the ambiguities in the framework and variables of the model, and the deviations in the field data. Thus, conventional methods for data processing are not sufficiently efficient anymore for solving issues related to the water quality. Additional efforts are required to improve the consistency of the findings of the model.

Deterministic models try to represent all the chemical and physical processes included in statistical terms, with variables acquired either from past data or obtained empirically, or computed by experience or examination. Generally, the differential equations are simplified so as to find solutions suitable for the model. Solution of the involved equations may need suppositions and simplifications which are derived from the performance of the model, and usually practical experience is necessitated from the user prior to achievement of optimal outcomes.

Statistical models attempt to seek general rules from the experimental data, which can be done by obtaining information from the field data. Statistical modelling and assessment involve a meticulous selection of techniques for analysis, and validation of suppositions as well as data. A majority of such models are quite complex and involve a substantial field data amount to conduct the analysis. Moreover, several statistical-based models of water quality, which assume the association among the prediction and the response variables, are distributed normally and linear in nature. Nevertheless, since the quality of water can be impacted by several parameters, conventional techniques for data processing are not sufficiently efficient anymore for solving this issue, and as such parameters show a complex non-linear relation to the water quality prediction parameters. Thus, using statistical techniques generally does not have high accuracy.

Of late, the AI (Artificial Intelligence) approach has been recognised as an effective alternative method for modelling of complicated non-linear systems. Generally, such models do not take into account the internal process but develop models through the inputs and outputs correlation. Presently, AI is used exhaustively for estimating several water-related regions (Muttil and Chau, 2006).

Recently, AI has offered the techniques for operation optimisation and selection of equipment, and problem solving that involve large quantities of data that cannot be processed by humans for the purpose of decision making. For this purpose, AI methods are proficient to replicate this behaviour and balance the deficiency. Thus, the growth of technology of efficient parallel computing and growing computing power have facilitated the researchers to employ the AI approaches (for instance, ANN (Artificial Neural Network) and ANFIS (Adaptive Neuro-Fuzzy Inference System)) for field data modelling solutions. The neuro-fuzzy technique has been used effectively in certain fields of water bodies engineering like the rainfall-runoff model (Chang and Chen, 2001)and basin operation (Chang and Chang, 2006; Chang et al., 2005). ANFIS has been known to enhance the accuracy of day-to-day estimation of evaporation (Kişi, 2006), reservoir water level prediction (Chang & Chang, 2006) and prediction of the river flow (Firat and Güngör, 2007).

The data obtained from experimentation and examination may be corrupted by signals of noise because of objective and/or subjective errors. For instance, experimental faults may be caused by measuring, recording, reading and external situations. As this noise can possibly distort the model outcomes, it is essential to eliminate them (i.e. signal de-noising) prior to the use of this data. The noisy signals can be de-noised by applying a series of linear filters (Bell and Martin, 2004). Nonetheless, these filters are more suitable for linear systems rather than the non-linear ones. Moreover, the FAT (Fourier analysis technique) is a standard tool for de-noising, though it is only favourable for de-noising signals or data involving stable noises. In addition, as there are unstable noises in real situations, it cannot be applied effectively. Thus, to solve the issues of conventional de-noising methods, more complex methods, like the WDT (wavelet de-noising technique), have been recommended. Above all, WDT is effective for de-noising multi-dimensional temporal or spatial signals having stable or unstable noises. Also, it has been extensively applied to industrial systems for information finding and patterns recognition (Avci, 2007; Tirtom et al., 2008). Nonetheless, some of these investigations were employed for water quality monitoring, where its data was utilised for estimation of parameters (Dohan and Whitfield, 1997).

In Malaysia WQIP requires extensive calculations and transformations. Two studies have been proposed to use Artificial Intelligence techniques (AI) in Malaysia in order to develop an accurate predictive model to WQP. However, many studies show that AI needs pre-processing tool to enhance the accuracy of the model practically in dealing with measured water quality data which is often contain noise (Han et al. 2011, Xu and Liu 2013).

The main objective of this investigation is to evolve a computationally proficient and robust method for the estimation of water quality variables decreasing the labour and cost for measurement of those parameters. This study focuses on the Malaysian Johor River situated in Johor State where the water quality dynamics are significantly altered. This research has many primary objectives, as follows:

- To evaluate and assess the correlation among the parameters of water quality on the basis of the experimental data using ANN (Artificial Neural Network).
- To propose various ANN approaches, like MLP (Multi-Layer Perceptron) Neural Network and RBF (Radial Basis Function) Neural Network so as to confirm the effectiveness of these techniques in the estimation of the parameters of water quality.
- To get familiar with the correctness of the ANFIS (Adaptive Neuro-Fuzzy Inference System) in the prediction of the parameters of water quality.
- To develop an augmented WDT-ANFIS (wavelet de-noising technique with the Neuro-Fuzzy Inference System).
- To examine the effectiveness of the suggested model for spatial position by presenting two different situations: the first situation (Scenario 1) is designed to set the model prediction at each station pertaining to the water parameters by considering the 13 input parameters from the same station. Where for Scenario 2, the input parameters for this scenario based on the measured water quality parameters from the same station and the predicted parameter from upstream station.

- To validate the augmented WDT-ANFIS (wavelet de-noising technique with the Neuro-Fuzzy Inference System) based on the experimental data for the duration 2009-2010.

3. Case Study: Johor River Basin

Johor state is regarded as the third largest region in Malaysia with an area of 19.984 km^2. It comprises of eight districts namely are Kota Tinggi, Muar, Pontian, Johor Bahru, Segamat Kluang, and lastly Batu Pahat which is considered as the second largest districts in Johor with an area of 187,702.06 hectares. Johor state has five principal rivers which are Sungai Muar, Sungai Johor, Sungai Endau, Sungai Batu Pahat and Sungai Sedilfi. This research sheds the light solely on Sungai Johor river. The Johor river basin is located in the southeast of Peninsular Malaysia. At an altitude of 1010 m, the Johor river orginates from the Gunung Belumut and from Bukit Gemuruh at an altitude of 109 m un the north. The river has irregular shape, its drainage area is around 2636 km2 and its length is approximately 122.7 Km. The river flows southeast into the Johor straits. An average annual precipitation of 2470 mm added to the river while during the period of 1963-1992, the annual mean discharge at Rantau Panjang was found to be 37.5 m3/s. The Johor river and its tributaries play a significant role as water suppliers for the state of Johor as well as for Singapore. Many factors contribute to the deterioration of the water quality of Johor River, mainly include the release of different kinds of pollutants at levels exceeding the allowed limits with the absence of local authorities' enforcement. These pollutants travel through Johor River and ultimately end in the estuaries of the rivers which are known to be a natural feeding area for poultries and fishes and a natural environment that provide spawning. Figure 1 depicts the location map of the surveying area which compromises of four monitoring stations on Johor River.

Figure 1.

3. Methodology

3.1 Multi-Layer Perceptron Neural Network (MLP-ANN)

A feed-forward network is the multi-layer perceptron neural network (MLPNN) that includes many layers of neurons, where one neuron's output is propagated to the other neuron's input that is in the next layer. Figure 2 presents the multi-layer perceptron neural network. In MLPNN, the input layer's nodes only propagate the input values of the first hidden layer's nodes. In the hidden layers, each node's input-output relationship can be presented as follows:

$$y = f\left(\sum_j w_j x_j + b\right) \qquad (1)$$

where, x_j signifies the output from the previous layer's j node, w_j denotes the connection weight between the current node and j node, b represents the current node's bias, and f defines a non-linear transfer function usually of the sigmoid form as shown in Equation (3.4):

$$f(z) = \frac{1}{1+\exp(z)} \qquad (2)$$

where, z denotes the weighted sum pertaining to the input to the neuron and $f(z)$ signifies the neuron output. The output nodes' input-output relationship is comparable to the one defined by Equation (3.4), with the exception of the case where the network is employed for function approximation, and the type of function f could vary (e.g. linear function).

Figure 2.

The units define a MLPNN architecture, which allows computation of a non-linear function in terms of the scalar product pertaining to the weight vector and input vector. Overall, the MLPNN models' performance relies on the network's inherent architecture. Apart from the number of hidden layers as well as the number of neurons pertaining to each layer, it also includes the computation type applied to each neuron.

3.2 ADAPTIVE NEURO-FUZZY INFERENCE SYSTEM (ANFIS)

Jang (Jang, 1993) first put forward the Adaptive Neuro-Fuzzy Inference System (ANFIS) that allowed realising a highly non-linear mapping and compared with common linear methods, it is considered to be superior in yielding non-linear time series (Jang, 1993). The ANFIS architecture was employed throughout this research for the first-order Sugeno fuzzy model (Sugeno and Kang, 1988). ANFIS can be defined as a multi-layer feed-forward network that employs neural network learning algorithms as well as fuzzy reasoning to aid in mapping input space with that of the output space (Chang and Chang, 2006). Considering that for a first-order Sugeno fuzzy model, the fuzzy inference system has one output, f, and two inputs, x and y, a common rule set that includes two fuzzy 'if.then' rules can be defined as follows:

$$\text{Rule 1: If x is } A_1 \text{ and y is } B_1\text{, then } f_1 = p_1\, x+q_1\, y+r_1 \tag{3}$$
$$\text{Rule 2: If x is A2 and y is } B_2\text{, then } f_2 = p_2\, x+q_2\, y+r_2 \tag{4}$$

where, A_1, A_2 and B_1, B_2 signify the membership functions (mfs) pertaining to inputs x and y, respectively; p_i, q_i and r_i (i = 1 or 2) represent the linear parameters pertaining to the first-order Sugeno fuzzy model's consequent part. Figure 3(a) represents the fuzzy reasoning mechanism pertaining to this Sugeno model that also allows deriving the output function (f) from that of inputs x and y. Figure 3(b) presents the corresponding equivalent ANFIS architecture, in which similar functions are associated with the same layer's nodes. ANFIS comprises five layers as stated below:

Figure 3.

3.3 WAVELET DE-NOISING

The next logical step is characterised by wavelet analysis post the short-time Fourier transforms (STFT). This is with regards to the windowing technique that includes variable-sized regions. With the help of wavelet transform (WT), long time intervals can be employed in those areas where more precise low frequency information is needed, as well as for shorter regions in which high frequency information is needed. Overall, the key benefit provided by the wavelets is allowing conducting local analysis for localised area pertaining to a larger signal. The discrete-time WT pertaining to a time domain signal $x[k]$ can be expressed as follows (Dohan and Whitfield, 1997):

$$DWT(m,n) = 1/_{\sqrt{2^m}} \sum_k x[k]\psi[2^{-m}n - k] \quad (5)$$

Here, (n) defines the mother wavelet, while m represents the scaling and k denotes the shifting indices. The *DWT* logarithmic frequency coverage is provided through scaling, as opposed to the uniform frequency coverage of STFT. This analysis technique includes segmenting a signal into components at various frequency levels, which are linked by the powers of two (a dyadic scale). The filtering approach that is applied to multi-resolution WT involves formation of a series of half-band filters that segment a spectrum into low and high frequency bands. The formulation is based on a wavelet function or high-pass (UP) filter as well as a scaling function or low-pass (LP) filter. Wavelet multi-resolution analysis (WMRA) allows constructing a pyramidal structure that needs an iterative application of wavelet functions and scaling to high-pass and low-pass filters, respectively. At the beginning, these filters are first applied to the entire signal band under high frequency (small-scale values) and then the signal band is decreased at every stage gradually. As presented in Figure 4, the detail coefficients of Dl, D2 and D3 define the high-frequency band outputs, while the approximation coefficients of Al, A2 and A3 define the low-frequency band outputs.

Figure 4.

Numerous factors need to be accounted when wavelets are employed to de-noise the WQP data. Examples of such choices include the level of decomposition, wavelet and thresholding methods to be employed. MATLAB provides various families of wavelets such as Morlet, Meyer, Mexican hat, Coiflets, Haar, Symlets, Daubechies and Spline biorthogonal wavelets, and also offers additional documentation regarding these wavelet families

("Wavelet Toolbox - MATLAB," n.d.). Only orthogonal wavelets need to be accounted to get perfect reconstruction results. Certain advantages are associated with the orthogonal wavelet transform. It can be characterised as being relatively concise, permitting perfect reconstruction of the original signal and relatively easy to calculate. The two common employed approaches for thresholding a signal include hard thresholding and soft thresholding, which are employed in the MATLAB wavelet toolbox. Although the easiest method is hard thresholding, better results are achieved through soft thresholding versus hard thresholding. Thus, this study uses soft thresholding. Four threshold selection rules can be used with the wavelet toolbox, which employ statistical regression pertaining to the noisy coefficients over time that allows getting a non-parametric estimation regarding the reconstructed signal absent noise. This study examined just Sqtwolog, wherein a fixed form of threshold is employed, leading to minimax performance that is multiplied by a factor proportional of signal length's logarithm. In this research, in terms of the decomposition level, we can conclude that a level 4 decomposition offered reasonable results post applying the trial-and-error method to all modules.

3.4 Model Performance Evaluation

It is necessary to clearly recognise the criteria that are associated with judging the model's performance. The criteria employed to assess the performance of the model in this study were clearly mentioned in the literature. Dogan et al. (Dogan et al., 2009) employed the Average Absolute Relative Error (AARE), which not only provides the performance index with regards to predicting water quality parameters but also demonstrates the prediction errors distribution. To examine the performance of the model, Singh et al. (2009) employed the bias statistical index. The bias signifies the mean of all the individual errors as well as allows determining if the dependent variable is underestimated or overestimated by the model. In this study, correlation coefficient as well as Root Mean Square Error (RMSE) was employed to examine the model's performance (Soyupak et al., 2003; Zhao et al., 2007).

Usually, the model performance is assessed through coefficient of determination, as put forward by Nash and Sutcliffe (1970), while MSE is employed to check the level of fitness between the network output and desired output.

In this research work, the models' performances were assessed based on three statistical indexes. As mentioned by Nash and Sutcliffe (1970) coefficient of efficiency (CE) is commonly employed to assess the performance of the model.

$$CE = 1 - \frac{\sum_{i=1}^{n}(X_m - X_p)^2}{\sum_{i=1}^{n}(X_m - \overline{X}_m)^2} \qquad (6)$$

where n represents the number of observations, X_m and X_p define the measured and predicted parameters, respectively, and $\overline{X}_m$ signifies the average of measured parameter.

Mean square error (MSE) is employed to see the level of fitness between network output and the desired output. Better performances are guaranteed with smaller MSE values. It is defined as follows:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(X_m - X_p)^2 \qquad (7)$$

More commonly, the coefficient of correlation (CC) is employed to examine the linear relationship between the measured and predicted dissolved oxygen. This can be expressed as follows:

$$CC = \frac{\sum_{i=1}^{n}(X_m - \overline{X}_m)(X_p - \overline{X_p})}{\sqrt{\sum_{i=1}^{n}(X_m - \overline{X}_m)^2 \sum_{i=1}^{n}(X_p - \overline{X_p})^2}} \qquad (8)$$

Further, for visual comparison of the predicted and measured values, the Scatter plot was employed (Kuo et al., 2007).

3.5 Input Variables and Data Processing

One of the key functions of ANN is to identify the model input parameters that could impact the output parameters considerably. As indicated above, the selection of input parameters depends on a priori knowledge regarding causal variables as well as statistical analysis pertaining to the potential outputs and inputs. In the literature, different input parameters were employed to develop the model to determine water quality parameters, as presented in Table 1.

Table 1.

On the basis of the literature, the following water quality parameters were chosen for ANN modelling: temperature (Temp), electrical conductivity (COND), salinity (SAL), nitrate (NO3), turbidity (TURB), phosphate (PO4), chloride (CI), potassium (K), sodium (Na), magnesium (Mg), iron (Fe) and Escherichia coli (E-coli). The basic statistical parameters, i.e. mean, minimum, maximum, standard deviation (S.D.), and coefficient of variation (CV) of the input and output parameters deployed in this study are depicted in Table 2 and Table 3.

Table 2.

Based on the concentration levels of both output and input parameters, large changes between the samples were seen, along with a high coefficient of variation (i.e. 254.94% for AN and 325.96% for E. coli). The coefficient of variation (CV) can be defined as a measure of statistical dispersion pertaining to the data. For a given data set, it is the mean normalised standard deviation (CV %) that can be computed as (standard deviation/mean) × 100. The existence of large disparity in the parameters' concentrations can be attributed to the types (non-point and point) and nature of sources that have been distributed in the river basin's wide geographical area. During the course, the river flows through different townships, and many tributaries and wastewater drains pouring large quantities of untreated wastewater into the river's main channel. A coefficient of variation in the range of 3.08% and 325.96% was seen with the parameters. Such variability that exists amongst the samples could be due to large geographical variations in climate as well as seasonal effects pertaining to the study region. For the various sampling sites, a spatial and significant variation was seen in terms of Johor River's turbidity, which varied from 0.2 to 343 NTU. It was higher, which could because of the mixing of industrial effluents and domestic sewerage water in Johor River. The rise in turbidity near downstream sites can be attributed to settling factors and flow turbulences. At downstream sites, the observed trend of turbidity, i.e. SN02, SN03 and SN04, was seen to support the above-mentioned hypothesis. Comparable patterns pertaining to spatial variations in turbidity were reported by (Khadse et al., 2007) when investigating Kanhan River's water quality. Amongst the sampling sites, the conductivity of the Johor River water was found to be considerably different, in which the mean ranged from 54 to 64 μS, although least significant difference was between SN01 and SN03. The high conductivity at SN04 and SN02 sites signify sewerage mixing into the river water. The dilution of industrial and urban runoffs could be attributed to the lower conductivity seen in the downstream water. Nitrate is considered to be a crucial parameter of river water that could be an indicator for the pollution status and anthropogenic load in river water.

The mean of nitrate ranged from 0.66 to 163.5 mg/l for Johor River. At the site wherein urban runoff mixing was noticed, NO3 was seen to be the maximum. It is interesting to note that in the downstream non-point pollution sites, lower NO3 was seen. The concentration of chloride in water was deemed not to be harmful. A higher concentration of chloride found in freshwater signified that pollutants are present. Moreover, in Johor River, the chloride level fell in the range of 5.27 to 7.37 mg/l. Nonetheless, at various sampling sites, a clear trend was not seen with chloride concentration in terms of the non-point or point pollution sites. The mixing of industrial effluents or urban wastewater in the river water is signified by higher levels of chloride content at SN04.

Table 3.

pH of water indicates alkaline and acidic conditions. DOE (DOE, 2007) suggested that pH for water in the range of 6.5–8.5 can be employed for any purposes in that respect; the ranges showed that Johor River had moderately alkaline water. The change in mean pH ranged from 6.22 to 6.36 at various locations. At some sites, higher pH could be a result of carbonate and bicarbonates of magnesium and calcium in water. The key source pertaining to such chemicals include industrial wastewater or urban runoff. SS further signifies the river water’s salinity behaviour. The mean SS content pertaining to river water was found in the range of 72.61 to 91.01 mg/l. The chemical and biological oxygen demand increase in tandem with higher SS level in the water system, which ultimately results in depletion of the dissolved oxygen level in water. In water, SS stems from natural sources, industrial wastewater, urban runoff, sewage and chemicals employed in the water treatment process.

For the current neural network modelling, the second assessment of selecting the input parameters is done by considering a statistical correlation analysis pertaining to the field data. Calculation of the correlation coefficient existing between the input and output parameters was done and listed in Table 4.

Based on the table, pH was clearly seen to be inversely associated with water temperature (r = -0.306) as well as potassium (r = -0.425). We performed an experiment by taking water quality variables that were accounted along with the parameters mentioned above pertaining to various models to realise the optimal predictive model as well as reduce the monitoring cost by accounting for fewer input parameters.

Table 4.

3.6 Stopping Criteria

Normally, there is a gradual decrease in the training error of AI since the training process is on-going. Nonetheless, this minimisation of training error does not guarantee enhancement of generalisation ability, which gained our interest. It is not necessary that AI showing good performance with the training set will do the same with the testing data. Therefore, it is also sometime important to stop the training phase at the right time before over-fitting occurs. When a generalisation characteristic is lost by the neural network, an over-fitting issue follows. However, relations between the training inputs as well as their associated outputs to similar hidden patterns pertaining to the unobserved data cannot be generalised. Thus, this occurs as a result of a difficult question that asks how long a network needs to be trained. The issue of over-fitting is usually solved by employing techniques like weight elimination, weight decay and early stopping. Stopping criteria is the most commonly employed method to address this issue. As noted by numerous researchers (e.g. Singh et al. (Singh et al., 2009);

Palani et al. (Palani et al., 2008)), two frequently employed stopping criteria include stopping post a specific number of runs via the complete training data (it needs to be noted that an epoch is defined as each run that passes through the complete training data) and stopping on reaching some low level by the target error.

3.6. Different Scenarios

Two different scenarios have been proposed in this study. The concept behind the development of these both scenarios is based on the spatial pattern of the input-output structure of the model. Mainly, the reason behind proposing these scenarios is to examine the model performance considering the spatial dimension of the model input. Keeping in mind that the model output in both scenarios is the prediction values of the AN, pH and SS, the input patterns has been changed in terms of the number of the inputs and location of the monitored data. In order to clarify the structure and show the difference between these two scenarios, an example for the structure of both scenarios to predict the AN parameter will be presented. For scenario I, to predict AN parameter at certain station, different twelve input parameters were used that have been acquired at the same station. While, the structure of scenario II is developed as, in addition to the same twelve water quality parameters used as inputs in scenario I, the value of AN parameter that has been acquired from the upstream station will be added.

The prediction procedure can be defined as an operation that allows offering water quality parameter patterns for the future. This research employs the WDT-ANFIS along with its stochastic and non-linear modelling capabilities to design a prediction model that mirrored the water quality parameter patterns pertaining to Johor River with regards to the 12 input parameters (Scenario 1) cited earlier, which is represented as follows:

$$WQIP_N = f_{WDT-ANFIS}(Temp_N + COND_N + SAL_N + TUR_N + NO_{3N} + CI_N + PO_{4N} + Fe_N + K_N + Mg_N + Na_N + E-coli_N) \quad (9)$$

$$N = 1,2,3,4$$

Where $\mathrm{WQIP_N}$ signifies the water quality index parameters pertaining to station *N*, and $f_{WDT-ANFIS}(.)$ defines the non-linear function predictor built via the WDT-ANFIS network. Thus, at each station, four models were built for predicting the parameters for water quality. A majority of the recent studies were aimed at predicting the concentrations pertaining to the parameters of water quality at every station. Usually, discharge via the local area from the upstream station causes an impact on the water pollution pertaining to a downstream station (Zaqoot et al., 2009). Therefore, in the put forward model, it was important to consider the impact cast by water parameters at the upstream station. Thus, the second scenario (Scenario 2) was designed to set the model prediction at each station pertaining to the water parameters by considering the 13 input parameters. At the previous station (upstream), the predicted WQIP could be represented by following Eq. (10). Repetition of this procedure involving the predicted WQIP is done for the fourth and third stations at downstream. Figure 5 presents a schematic representation pertaining to the put forward networks for Scenario 2.

$$WQIP_{N+1} = f_{WDT-ANFIS}(Temp_N + COND_N + SAL_N + TUR_N + NO_{3N} + CI_N + PO_{4N} + Fe_N + K_N + Mg_N + Na_N + E-coli_N + WQIP_{pN}) \quad (10)$$

Figure 5.

7. Results and Discussion

7.1 MLP-ANN Training

The construction of an ANN model normally includes three steps. The training stage is the first step, in which the network is exposed to a training set pertaining to the input-output patterns. The second step involves the validation stage, in which the network's performance is evaluated when patterns are not 'observed' by the network in the training stage. The third step includes the testing stage, in which the network's performance is evaluated when the unknown patterns were not 'observed' during the stages of validating and training (Bowden et al., 2005). Designing of three MLP-ANN architectures was done (one for each parameter). The Levenberg-Marquardt back propagation algorithm (LMA) is employed by all three networks in the entire training procedure. This study employed three activation functions, namely tan-sigmoidal (Tansig), log-sigmoidal (logsig) function and linear transfer function (purelin). After initialising the network weights and biases during the training process, iterative adjustments of the weights and biases pertaining to the network were carried out to decrease the network performance function pertaining to mean square error (MSE) – the average squared error between the target outputs and the network outputs.

We introduced different values of learning rate (lr) to the networks in a bid to achieve the optimum result pertaining to this study. For back propagation learning algorithm, the learning rate is important as it helps determine the level of weight changes. However, since the learning process tends to slow down when smaller learning rate values are employed for training, it is not a favoured choice. Employing larger learning rates values for training could lead to network oscillation in the weight space. One approach to enhance the gradient descent method is by introducing an additional momentum parameter (mc) that facilitates larger learning rates leading to faster convergence while decreasing the oscillation tendency (Rumelhart et al., 1986). The momentum term is introduced so that the next weight changes are similarly aligned to the same direction as the previous one, which allows minimising the oscillation impact of larger learning rates. Although there are certain systematic approaches to simultaneously choose the learning rate and momentum, the best values pertaining to these learning parameters are normally selected based on experimentation. Since any value falling between 0 and 1 can be accounted by the learning rate and the momentum, it becomes almost impossible to perform an exhaustive search to detect the best combinations pertaining to these training parameters. In this research paper, we evaluated different momentum and learning rates pertaining to both networks; in real practice, 0.9 and 0.95 were selected as momentum and optimum learning rate pertaining to SS, AN and pH models, respectively.

7.2 Optimisations of the Neurons Number

The number of neurons in the hidden layer is the key characteristic pertaining to AI technique. The network fails to model the complex data that could lead to poor fitting if the number of neurons employed is insufficient. On the flip side, the training time could become unreasonably long as well as the network may also over fit the data if there are too many neurons employed. In this paper, to investigate the best performance, various MLP-ANN architectures were employed. In fact, a formal and/or mathematical approach does not exist, which allows determination of appropriate ‘optimal set’ pertaining to neural network’s key parameters. Thus, the trial-and-error method was selected to perform this task. Randomisation of the hidden layer’s neurons was done from N=1 to 20 neurons. In the hidden layer, the best numbers of nodes are those that provide the lowest error (Lek et al., 1996). Based on two performance indices, determination of the optimum number of neurons was done. The root-mean-square error (RMSE) value pertaining to the prediction error is the first index, while the value of the maximum error is the second index. To get both indices, the ANN model was evaluated by considering the WQP data between 1998 and 2007. When building such a predicting model that employs the neural network, the model could do well during the training period and could give a higher level of error when assessment was done during either the testing or validation period. Based on this study, these performance indices were employed to ensure that the put forward model would offer consistent accuracy levels during all periods. As the performance indicator for the put forward model, the key benefit of using these two statistical indices is to ensure that the highest error falls within the acceptable error range for the forecasting model when the performance is being evaluated. This is done when RMSE is employed and making sure that the summation of the error distribution is not high in the validation period. Consequently, employing both indices ensures consistent level of errors and offers high potential to maintain the same error level while evaluating the model for unseen data during the testing period.

When the number of hidden neurons to the network is varied, it has a clear impact to a considerable degree on the prediction performance. It clearly demonstrates that there is a rise in prediction performance with increase in the number of hidden neurons (from 1 to 18), along with subsequent decrease in RMSE and maximum error pertaining to all parameters. However, a drop in prediction performance occurred when hidden neurons were added further (19 to 20) to the network. For instance, it can be seen that the best combination pertaining to the put forward statistical indices to examine the predicting model for the pH was when 18 neurons with RMSE 0.15 were associated with the ANN architecture and a maximum error as 3.22%. The best combination pertaining to the put forward statistical indices to examine the predicting model for the SS was when 17 neurons with RMSE 0.30 were associated with the ANN architecture and a maximum error of 3.46%. Table 5 lists out the optimal numbers of neurons pertaining to the remaining parameters.

Table 5.

7.3 WATER QUALITY PREDICTION MODEL OF MLP-ANN

The MLP-ANN model for the estimation of the 6 parameters of water quality (as the output), which are SS, AN and pH, was evaluated in this section. Figure 6 depicts the measured and estimated parameters of water quality for the most excellent network, which provided the most precise estimation. On the whole, the predictive capability of this model was fairly good for each of the parameters of the water quality in the training duration, though less accurate when the validation and testing stages were carried out. The findings showed that it was challenging to develop a consistent model using the MLP-ANN models due to high variations and intrinsic non-linear correlation among the parameters of the water quality because of the probabilistic nature and chemical procedure. Additionally, the MLP-ANN models encountered delayed convergence during the training because of the necessity of comparatively a huge amount of hidden neurons. Also, several researchers observed that these models failed to acquire values lying outside the scope of values included in the calibration data of MLP-ANN (boundary values) (Campolo et al., 1999; DAWSON and WILBY, 1998; Hsu et al., 1995; Karunanithi et al., 1994; MINNS and HALL, 1996). This constraint, arising chiefly due to the application of a logistic function to translate the output of the model, makes these models inappropriate for several applications.

Alternatively, the RBF-ANN (Radial Basis Function Network) is commonly employed for strict interpolation issues in space with multiple dimensions, which has equivalent abilities as the MLP-ANN in solving problems related to function estimations (Park and Sandberg, 1993). There are chiefly 2 benefits of the RBF-ANN: (a) network training in shorter duration in comparison to MLP-ANN , and (b) best solution estimation without managing the local minimums. In addition, RBF-ANN works as a local network in contrast to the feed-forward networks which are global mapping networks. Also, RBF-ANN employs one processing units set, and every unit is most accessible to a local area of the input region.

Due to this, RBFNs are employed more recently as a substitute NN model in function estimation applications and prediction of time series (Sheta and De Jong, 2001; Yu et al., 2008). Thus, the following section describes the attempt to get familiar with RBF-ANN suitability to be used as a model for predicting the parameters of water quality.

Figure 6.

7.4 SENSITIVITY ANALYSIS

To assess the input variables, impact on the model, 3 assessment methods were used. First method was based on dividing the NN connection weights so as to establish the relative significance of every input variable in the network (Stern and Garson, 1999). In this research, the recommended network comprises 12 environmental variables. Presuming the connection weights from the input nodes to the hidden nodes exhibit the relative predictive significance of the independent parameter, the significance of every input parameter can be articulated as follows:

$$Ij = \frac{\sum_{m=1}^{m=Nh}\left(\left(\left|W_{jm}^{ih}\right| / \sum_{k=1}^{Ni}\left|W_{km}^{ih}\right|\right)\times\left|W_{mn}^{ho}\right|\right)}{\sum_{k=1}^{k=Ni}\left\{\sum_{m=1}^{m=Nh}\left(\left(\left|W_{jm}^{ih}\right| / \sum_{k=1}^{Ni}\left|W_{km}^{ih}\right|\right)\times\left|W_{mn}^{ho}\right|\right)\right\}} \quad (11)$$

Where Ij represents the relative significance of jth input variable on the output variable, Ni and Nh denote the quantities of input and hidden neurons, correspondingly, and W represents the connection weight. Also, the superscripts 'i', 'h' and 'o' signify the input, hidden and output levels, correspondingly, while the subscripts 'k', 'm' and 'n' signify the input, hidden and output neurons, correspondingly. The first method of evaluation was to assess the relative significance of every input variable as calculated by Eq. (11) and illustrated in Figure 7. The relative significance demonstrates the importance of a variable in comparison to the other variables belonging to the model. Even though the network did not essentially signify physical sense using weights, it indicates that all the variables had intense effects on the estimation of all output variables, in which the estimator contribution varied from 5 to 14%. Apparently, the most useful inputs were considered to be those that involved oxygen containing nitrate (NO3) and phosphate (PO4). Conversely, pH and Temp were discovered to be the least useful parameters. Additionally, MG proved to be providing the greatest contribution for the recommended model for AN. For pH, it was apparent that the most useful input was Temp.

Figure 7. Relative importance of each input parameter.

7.5 WATER QUALITY PREDICTION MODEL OF ANFIS

As a matter of fact, among the difficulties in ANFIS-based modelling is establishing its variables for optimal learning (i.e. the membership function number and step size's initial value) before training, in a way that the optimal training is achieved. Two techniques have been proposed by several researchers for establishing these variables in ANFIS: optimisation techniques (Hassanain et al., 2004) and the trial-and-error approach (Kim et al., 2002). While determining the variables for optimal learning could be ensured by the optimisation algorithms (i.e. derivative based or derivative free optimisation), this alternative has a downside of being computationally costly. Conversely, the trial-and-error technique has been confirmed to be effective in case the target root mean square error can be realised. This technique is also advantageous as it yields a knowledge rule-base having a lower possibility of surpassing the data set of training in comparison to the optimisation technique. Thus, this research did not include the optimisation technique and established the variables for optimal learning of ANFIS through the trial-and-error technique.

For every parameter related to the water quality, this study employed the architectures proposed in the preceding section, in which 12 inputs were utilised to estimate the WQIP. It is noteworthy that there is no systematic technique to establish the optimal quantity of MFs. The optimal quantity of MFs is generally established inductively and validated empirically. Thus, the quantity of MFs was selected using the trial-and-error method. Meanwhile, it is to be observed that this study had tested 4 kinds of membership functions: (a) triangular, (b) gaussian, (c) trapezoidal, and (d) bell-shaped, to compose the fuzzy numbers. Following several trials, the outcome revealed a distributed membership function having bell-shaped nature in comparison to others which had acquired the minimal relative error. Table 6 demonstrates the kinds and quantity of MFs that were implemented in this study to develop the modules.

Table 6.

For demonstrating the performance of the suggested ANFIS model, an evaluation of predicted against observed parameters of water quality during training, validation and experimentation phases is displayed in the Figure 8. It is apparent that the suggested ANFIS model procedure provided the estimated variables that mimicked the dynamics (pattern) in the noted values besides those boundary values measured during this time.

Figure 8.

7.6 WATER QUALITY PREDICTION MODEL OF WDT-ANFIS

The above findings were obtained with the general assumption that the mined data must be precise and reliable. Nevertheless, the data acquired from the study, test, and simulation procedures may be corrupted by noise because of objective and/or subjective errors (Li and Shue, 2004). For instance, the errors arising in the experiment may be caused by measuring, recording, reading, or external scenarios; the errors from simulation might cover uncertainties of the model and parameters, as well as computational errors. As these noisy signals possibly distort the data mining outcomes, it is necessary to eliminate them (i.e. signal de-noising process) before the use of any initial data. Thus, an augmented WDT-ANFIS based on historical information for WQPP will be presented.

Training and cross-validation processes of the model of WDT-ANFIS were carried out to reduce the Root Mean Square Error among the output as well as predicted responses. The WDT-ANFIS model outperformed the ANFIS model and provided improvement in estimation accuracy of all the variables, while the ANFIS model performed inefficiently. As the noise intensity increased, it was obvious that WQP possibly had more accurate estimation values due to de-noising of data. This suggests the WDT superiority in data cleaning. Despite the occurrence of errors during stages of training, validation and experimentation, which were regarded as considerably high in comparison to the training and cross-validation stages, it had obtained a high precision for all variables. The findings displayed in Figure 9 demonstrate that the WDT-ANFIS model could be regarded as a suitable technique for modelling for estimation like WQP.

Figure 9.

7.7 COMPARATIVE ANALYSIS

The models introduced in prior discussion were all compared for the purpose of providing precise predictions for each water-quality parameter at Johor River. Similar findings were achieved in determining models for predicting suspended solids concentrations (SS), wherein WDT-ANFIS forecast SS with comparatively less accuracy, in which errors for most records were below 10%. Peak SS values were more closely approximated using WDT-ANFIS in comparison to that attained using other techniques, as depicted in Figure 10. The numbers of inaccurate SS forecasts decreased meaningfully using WDT-ANFIS. The use of physics-based distributed processing in complex computer software is frequently problematic, owing to the usage of idealised sedimentation components or the requirement of large volumes of detailed temporal and spatial data on the environment which is not always available (Cigizoglu, 2004). It should be noted that AI approaches to determining suspended-sediment data estimations remain sparse in the relevant literature (Abrahart and White, 2001).

The success attained in modelling dynamic systems implies that this strategy may well provide an efficient and productive means for simulating complex suspended-sediment processes in rivers, under conditions where precise knowledge of internal sub-processes is not necessary. Each proposed model in this study was constructed on the assumption that land cover/use would remain unchanged during this research. However, land cover/use remains an important factor in the production and transport of sediments, along with other factors. More precise predictions of suspended sediments may be attained by including variables that represent land cover/use status into the scheme. We are planning such analytical studies soon enough. In conclusion, this research establishes WDT as an appropriate method, along with classical ANFIS, for modelling suspended sediments in river environments. It is therefore worth considering the use of WDT-ANFIS approaches in such analysis, given the findings of studies regarding the physics embedded in ANFIS structures.

Figure 10.

With regards to pH, Figure 11 depicts comparisons between ANFIS and other models' performances, based on the test data set. In the figure, it is clear that ANFIS performance exceeds that of the two ANN methods. Furthermore, the effort reveals the challenges in devising reliable schemes based on MLP-ANN RBF-ANN models, as a result of the high variances as well as the inherent non-linear associations among the water-quality parameters, as a result of the stochastic quality and chemical-based process. Furthermore, as depicted in Figure 10, the findings show that WDT-ANFIS-based modules outperform ANFIS and also have the ability to improve predictive accuracy for pH, albeit for MAE with comparatively lesser accuracy, whereby errors for most records were below 7%. Otherwise, inefficient executions were observed based on the ANFIS module, wherein most errors were above 15%. Clearly, given increases in noise intensities, WQP offers more precise predictions from data de-noised with WDT than data without such de-noising. This suggests the advantage of using WDT to clean the data.

It is fact that the training process for big data using any of AI models is both time-consuming and computation- and memory-intensive especially when several number of model' inputs variables is used. The computer specification that have been used to run models are Intel Processor Core i7 (12M Cache, up to 4.60 GHz) and Ram 16 Gb. It is fact that in our study the data used is not big data to be considered as problem to the computational memory. However, due to the fact that the number of the model' input variables is relatively big (twelve or thirteen based on the structure of scenario I and scenario II, respectively), the training process is slightly time-consuming to achieve the performance goal. Table 7 summarize the training time for each models in seconds where it is noticeable that the ANFIS and WDT-ANFIS models consuming more time than ANN models (MLP and RBF) but it is still minimal.

Figure 11.

Table 7

7.8 SCENARIOS

The comparatively low correlation among forecast and observed values during test phases was perhaps a result of the non-homogenous nature of water-quality parameters. Moreover, Ying et al. (Zhao et al., 2007) demonstrated that the selection of influential factors (namely, input parameters) has a critical role as these factors greatly affect forecasts. Clearly, the low correlations in this research can be attributed to the realisation that its input parameters had not included every relevant parameter. Furthermore, pollution levels at downstream stations were associated with discharge from upstream stations. To overcome this difficulty, the researchers applied another approach (i.e. Scenario 2), such that higher levels of accuracy could be attained. This strategy is associated with the prediction of each water-quality parameter, given the actual values measured at upstream stations as model inputs, as described by Eq. (12). For a most appropriate analysis, the researchers implemented an accuracy improvement (AI) index for the correlational coefficient statistical index, in order to determine the significance of Scenario 2 as against Scenario 1, described as follows:

$$AI(\%) = \left(\frac{CC_{Scen2} - CC_{Scen1}}{CC_{Scen2}}\right) * 100 \tag{12}$$

Wherein CCScen2 denotes the coefficient of correlation for Scenario 2, whereas CCScen1 denotes a similar statistical index for Scenario 1. From Table 8, it is clear that Scenario 2 is more satisfactory than Scenario 1, with meaningful improvements observed in every station, which ranged from 0.5% to 5%. Predictive accuracy was meaningfully enhanced after introducing Scenario 2 for every station. As in the case for pH, Scenario 2 showed more satisfactory performance than Scenario 2, with meaningful improvements observed in AI, which ranged from 3% in Station 2 to 5% in Station 3.

Conversely, less improvement was gained with AN, wherein AI was equal to 0.5 in Stations 1 and 3. Even though it is clear that Scenario 2 was less efficient with AN, accuracy does increase by 2% once it is applied to Station 3. Furthermore, the findings indicate that Scenario 2 not only showed improved accuracy for certain parameters, but this particular model had the ability to capture temporal patterns in water-quality parameters. This enabled the scheme to apply meaningful improvements to station scenarios.

Table 8.

7.9 MODEL VALIDATION

Models must be verified whenever resulting outputs and observed values are near enough to satisfy all validation criteria (Palani et al., 2008). To investigate the effectiveness of this proposed scheme, validation of the enhanced wavelet de-noising method using the Neuro-Fuzzy Inference System (WDT-ANFIS), in accordance with field measurements collected from 2009 to 2010, is therefore applied. The scatter plots among the forecast and observed values for all 5 selected parameters for water quality are depicted in Figure 12. Clearly, the majority of forecast water-quality parameters had closely approximated actual observations. As well, R^2 must be as near 1 as possible, with values that exceed 0.9 implying very satisfactory model execution, values from 0.6 to 0.9 implying fairly good execution, and values below 0.5 indicating unsatisfactory execution. Based on these criteria, the WDT-ANFIS model's ability to predict both pH and SS concentrations is very satisfactory (in that R^2 values are at least 0.9) for every station but for AN, wherein models showed merely decent performances (in that R^2 values were below 0.9) for Station 3. Based on these findings, WDT-ANFIS can be said to demonstrate good predictive performance. For predictions of water-quality parameters using AI, other researchers have advanced network modelling strategies that apply differing types of AI as well as input datasets. Moatar et al. (Moatar et al., 1999) applied solar radiation and discharge levels in predictions of pH, with an R^2 value equal to 0.86. For predictions of AN, WDT-ANFIS predictive performance in this research managed better in comparison (R^2 ranging from 0.88 to 0.96) with ANN predictive performance. Cigizoglu (Cigizoglu, 2004) utilised ANN models that were trained and then tested with daily flows, for predicting SS concentrations a day ahead, with R^2 values ranging from 0.75 to 0.81 (with upstream flows as inputs). A comparable prediction for SS was similarly claimed by Zhu et al. (Zhao et al., 2007). For predictions of SS, the WDT-ANFIS predictive performance in this research managed better in comparison (R^2 ranging from 0.91 to 0.95) to previous studies. The proposed scheme demonstrated efficiency in its predictions of the concentrations of water-quality parameters for the Johor River, which corresponds to the findings of other research. The findings also show that the proposed scheme is a useful alternative that offers a comparatively fast algorithm, featuring decent theoretical properties for predicting water-quality parameters, which could be extended to predictions of other water-quality parameters.

Figure 12.

8. CONCLUSION

The study proposes the use of enhanced Wavelet De-noising Techniques using Neuro-Fuzzy Inference Systems (WDT-ANFIS) according to historical water-quality parametric data. The effectiveness of each model was examined in order to predict key parameters that could be affected as a result of urbanisation surrounding rivers. This area of research accords with the available secondary data for each water-quality parameter of Johor River. The parameters comprise ammoniacal nitrogen (AN), suspended solid (SS), and pH. Dual scenarios were presented: the first (Scenario 1) was designed to confirm prediction models for water-quality parameters at each stations according to 12 input parameters, whereas the second (Scenario 2) is designed to confirm prediction models for water-quality parameters according to 12 input parameters, as well as the parametric values from prior upstream stations. In evaluating the impact of input parameters on this scheme, validation of enhanced Wavelet De-noising Techniques using Neuro-Fuzzy Inference Systems (WDT-ANFIS), in accordance with measurements taken from 2009 to 2010, was thereby employed. The findings showed the challenge of determining reliable schemes based on MLP-ANN models, from the high variances as well as inherent non-linear associations among the water-quality parameters that emerge as a result of the stochastic quality and chemical-based process. Furthermore, MLP-ANN was subject to slow convergence during training, as a result of the requirement for comparatively large numbers of hidden neurons. In the example of RBF-ANN, its predictive capability for water-quality parameters in training phases was decent, but showed less precision during validation and test phases. The findings indicated that ANFIS determined solutions faster than alternative MLP-ANN and RBF-ANN methods and is the most precise and reliable method for processing large volumes of non-linear as well as non-parametric data. Of note is the performance of the WDT-ANFIS scheme, which exceeded that of ANFIS and improved predictive accuracy for every quality parameter, in that this model achieves higher prediction accuracy overall. Generally, WDT-ANFIS can therefore be seen as having the best network architecture, since it outperformed ANFIS. The findings indicate that WDT-ANFIS not only offered a means to improve accuracy but it also features the ability to capture temporal patterns in water quality. This enables it to provide meaningful improvements in the generation of forecasts. Consequently, the ANFIS model appears more capable at capturing the more complex and dynamic processes that are hidden within the data for WQP, following enhancement with WDT. In comparisons between Scenarios 1 and 2, Scenario 2 achieved higher accuracy in terms of simulating the patterns and magnitudes for every water-quality parameter, at every station. The suggested WDT-ANFIS model in Scenario 2 gave predictions for water-quality parameters that ably mimicked patterns (dynamics) in recorded values, aside from extreme outliers observed within this period. Furthermore, validation of WDT-ANFIS, according to measurements collected from 2009 to 2010, demonstrated that WDT-ANFIS performed well in predicting both pH and SS concentrations (with R^2 values of at least 0.9) for every station but for AN, wherein models still showed decent performances (with R^2 values lower than 0.9) for Station 3. Since forecasts of water quality are readily influenced by external environments, the acquired model would at times generate findings that deviated much from the observed values. In general, the methodology of the proposed models development for water quality has proved its effectiveness. However, it should be highlighted that there are no structured methods today to identify which network structure that can best in predicting water quality parameters. Moreover, the optimal selection of the hyper parameters still requires to be achieved by augmenting the AI model with other advanced meta-heuristic optimization algorithms. Overall, this study integrates several analytical and modelling techniques that could become useful to institutions that are committed to river basin management within Malaysia. Furthermore, the approach utilised in this research could lay ground for better decision-making that assists policy makers in maintaining and improving river basin management.

Acknowledgments: The authors would like to appreciate the technical and financial support received from research grant coded J510050822 by Innovation & Research Management Center (iRMC), Universiti Tenaga Nasional (UNITEN) and from research grant coded UMRG RP025A-18SUS funded by the University of Malaya

Conflicts of Interest: The authors declare no conflict of interest.

References

Abrahart, R.J., White, S.M., 2001. Modelling sediment transfer in Malawi: comparing backpropagation neural network solutions against a multiple linear regression benchmark using small data sets. Phys. Chem. Earth, Part B Hydrol. Ocean. Atmos. 26, 19–24. https://doi.org/10.1016/s1464-1909(01)85008-5

Avci, E., 2007. An expert system based on Wavelet Neural Network-Adaptive Norm Entropy for scale invariant texture classification. Expert Syst. Appl. 32, 919–926. https://doi.org/10.1016/j.eswa.2006.01.025

Bell, W.R., Martin, D.E.K., 2004. Computation of asymmetric signal extraction filters and mean squared error for ARIMA component models. J. Time Ser. Anal. 25, 603–623. https://doi.org/10.1111/j.1467-9892.2004.01920.x

Bowden, G.J., Dandy, G.C., Maier, H.R., 2005. Input determination for neural network models in water resources applications. Part 1—background and methodology. J. Hydrol. 301, 75–92. https://doi.org/10.1016/j.jhydrol.2004.06.021

Campolo, M., Andreussi, P., Soldati, A., 1999. River flood forecasting with a neural network model. Water Resour. Res. 35, 1191–1197. https://doi.org/10.1029/1998wr900086

Chang, F.-J., Chang, Y.-T., 2006. Adaptive neuro-fuzzy inference system for prediction of water level in reservoir. Adv. Water Resour. 29, 1–10. https://doi.org/10.1016/j.advwatres.2005.04.015

Chang, F.-J., Chen, Y.-C., 2001. A counterpropagation fuzzy-neural network modeling approach to real time streamflow prediction. J. Hydrol. 245, 153–164. https://doi.org/10.1016/S0022-1694(01)00350-X

Chang, Y.-T., Chang, L.-C., Chang, F.-J., 2005. Intelligent control for modeling of real-time reservoir operation, part II: artificial neural network with operating rule curves. Hydrol. Process. 19, 1431–1444. https://doi.org/10.1002/hyp.5582

Cigizoglu, H.K., 2004. Estimation and forecasting of daily suspended sediment data by multi-layer perceptrons. Adv. Water Resour. 27, 185–195. https://doi.org/10.1016/j.advwatres.2003.10.003

DAWSON, C.W., WILBY, R., 1998. An artificial neural network approach to rainfall-runoff modelling. Hydrol. Sci. J. 43, 47–66. https://doi.org/10.1080/02626669809492102

DID, 2000. Urban Stormwater Management Manual for Malaysia.

DOE, 2007. Malaysia Environmental Quality Report 2007. Malaysia Environ. Qual. Rep. 1–86. https://doi.org/10.1007/s13398-014-0173-7.2

Dogan, E., Sengorur, B., Koklu, R., 2009. Modeling biological oxygen demand of the Melen River in Turkey using an artificial neural network technique. J. Environ. Manage. 90, 1229–1235. https://doi.org/10.1016/j.jenvman.2008.06.004

Dohan, K., Whitfield, P.H., 1997. Identification and characterization of water quality transients using wavelet analysis. I. Wavelet analysis methodology. Water Sci. Technol. 36, 325–335. https://doi.org/10.2166/wst.1997.0229

Firat, M., Güngör, M., 2007. River flow estimation using adaptive neuro fuzzy inference system. Math. Comput. Simul. 75, 87–96.

Hsu, K., Gupta, H.V., Sorooshian, S., 1995. Artificial Neural Network Modeling of the Rainfall-Runoff Process. Water Resour. Res. 31, 2517–2530.

Hull, V., Parrella, L., Falcucci, M., 2008. Modelling dissolved oxygen dynamics in coastal lagoons. Ecol. Modell. 211, 468–480. https://doi.org/10.1016/j.ecolmodel.2007.09.023

Ibrahim, R., 2001. River Water quality Status In Malaysia, in: National Conference On Sustainable River Basin Management In Malaysia.

Jang, J.-S.R., 1993. ANFIS: adaptive-network-based fuzzy inference system. Syst. Man Cybern. IEEE Trans. 23, 665–685.

Karunanithi, N., Grenney, W.J., Whitley, D., Bovee, K., 1994. Neural Networks for River Flow Prediction. J. Comput. Civ. Eng. 8, 201–220. https://doi.org/10.1061/(asce)0887-3801(1994)8:2(201)

Khadse, G.K., Patni, P.M., Kelkar, P.S., Devotta, S., 2007. Qualitative evaluation of Kanhan river and its tributaries flowing over central Indian plateau. Environ. Monit. Assess. 147, 83–92. https://doi.org/10.1007/s10661-007-0100-x

Kim, B., Park, J.H., Kim, B.-S., 2002. Fuzzy logic model of Langmuir probe discharge data. Comput. Chem. 26, 573–581. https://doi.org/10.1016/s0097-8485(02)00021-9

Kişi, Ö., 2006. Daily pan evaporation modelling using a neuro-fuzzy computing technique. J. Hydrol. 329, 636–646. https://doi.org/10.1016/j.jhydrol.2006.03.015

Koklu, R., 2006. Dissolved oxygen estimation using artificial neural network for water quality control, Fresenius Environmental Bulletin.

Kuo, J.-T., Hsieh, M.-H., Lung, W.-S., She, N., 2007. Using artificial neural network for reservoir eutrophication prediction. Ecol. Modell. 200, 171–177.

https://doi.org/10.1016/j.ecolmodel.2006.06.018

Lek, S., Delacoste, M., Baran, P., Dimopoulos, I., Lauga, J., Aulagnier, S., 1996. Application of neural networks to modelling nonlinear relationships in ecology. Ecol. Modell. 90, 39–52. https://doi.org/10.1016/0304-3800(95)00142-5

Li, S.-T., Shue, L.-Y., 2004. Data mining to aid policy making in air pollution management. Expert Syst. Appl. 27, 331–340. https://doi.org/10.1016/j.eswa.2004.05.015

MINNS, A.W., HALL, M.J., 1996. Artificial neural networks as rainfall-runoff models. Hydrol. Sci. J. 41, 399–417. https://doi.org/10.1080/02626669609491511

Moatar, F., Fessant, F., Poirel, A., 1999. pH modelling by neural networks. Application of control and validation data series in the Middle Loire river. Ecol. Modell. 120, 141–156. https://doi.org/10.1016/s0304-3800(99)00098-8

Muttil, N., Chau, K.W., 2006. Neural network and genetic programming for modelling coastal algal blooms. Int. J. Environ. Pollut. 28, 223. https://doi.org/10.1504/ijep.2006.011208

Palani, S., Liong, S.-Y., Tkalich, P., 2008. An ANN application for water quality forecasting. Mar. Pollut. Bull. 56, 1586–1597. https://doi.org/10.1016/j.marpolbul.2008.05.021

Park, J., Sandberg, I.W., 1993. Approximation and Radial-Basis-Function Networks. Neural Comput. 5, 305–316. https://doi.org/10.1162/neco.1993.5.2.305

Rumelhart, D.E., Hinton, G.E., Williams, R.J., 1986. Learning representations by back-propagating errors. Nature 323, 533–536. https://doi.org/10.1038/323533a0

Sheta, A.F., De Jong, K., 2001. Time-series forecasting using GA-tuned radial basis functions. Inf. Sci. (Ny). 133, 221–228. https://doi.org/10.1016/s0020-0255(01)00086-x

Singh, K.P., Basant, A., Malik, A., Jain, G., 2009. Artificial neural network modeling of the river water quality—A case study. Ecol. Modell. 220, 888–895. https://doi.org/10.1016/j.ecolmodel.2009.01.004

Soyupak, S., Karaer, F., Gürbüz, H., Kivrak, E., Sentürk, E., Yazici, A., 2003. A neural network-based approach for calculating dissolved oxygen profiles in reservoirs. Neural Comput. Appl. 12, 166–172. https://doi.org/10.1007/s00521-003-0378-8

Stern, C., Garson, G.D., 1999. Neural Networks. An Introductory Guide for Social Scientists. Contemp. Sociol. 28, 753. https://doi.org/10.2307/2655607

Sugeno, M., Kang, G.T., 1988. Structure identification of fuzzy model. Fuzzy Sets Syst. 28, 15–33. https://doi.org/10.1016/0165-0114(88)90113-3

Tirtom, H., Engin, M., Engin, E.Z., 2008. Enhancement of time-frequency properties of ECG for detecting micropotentials by wavelet transform based method. Expert Syst. Appl. 34, 746–753. https://doi.org/10.1016/j.eswa.2006.10.009

Wavelet Toolbox - MATLAB [WWW Document], n.d.

Yu, L., Lai, K.K., Wang, S., 2008. Multistage RBF neural network ensemble learning for exchange rates forecasting. Neurocomputing 71, 3295–3302. https://doi.org/10.1016/j.neucom.2008.04.029

Zaqoot, H.A., Ansari, A.K., Unar, M.A., Khan, S.H., 2009. Prediction of dissolved oxygen in the Mediterranean Sea along Gaza, Palestine – an artificial neural network approach. Water Sci. Technol. 60, 3051–3059. https://doi.org/10.2166/wst.2009.730

Zhao, Y., Nan, J., Cui, F., Guo, L., 2007. Water quality forecast through application of BP neural network at Yuqiao reservoir. J. Zhejiang Univ. A 8, 1482–1487. https://doi.org/10.1631/jzus.2007.a1482

Figures

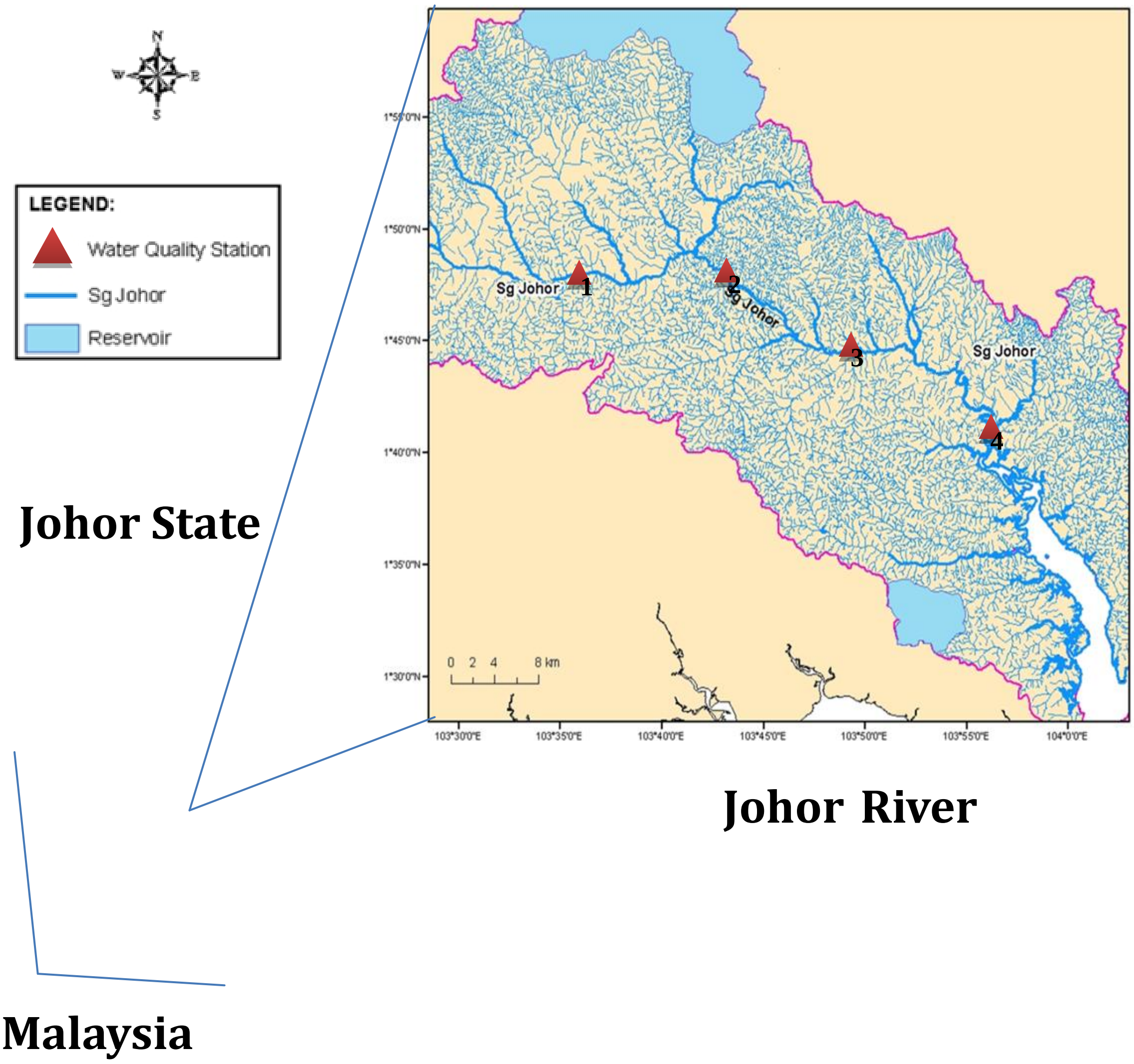


Figure 1. A map showing the geographical setting of the survey area with four field monitoring stations on the main stream

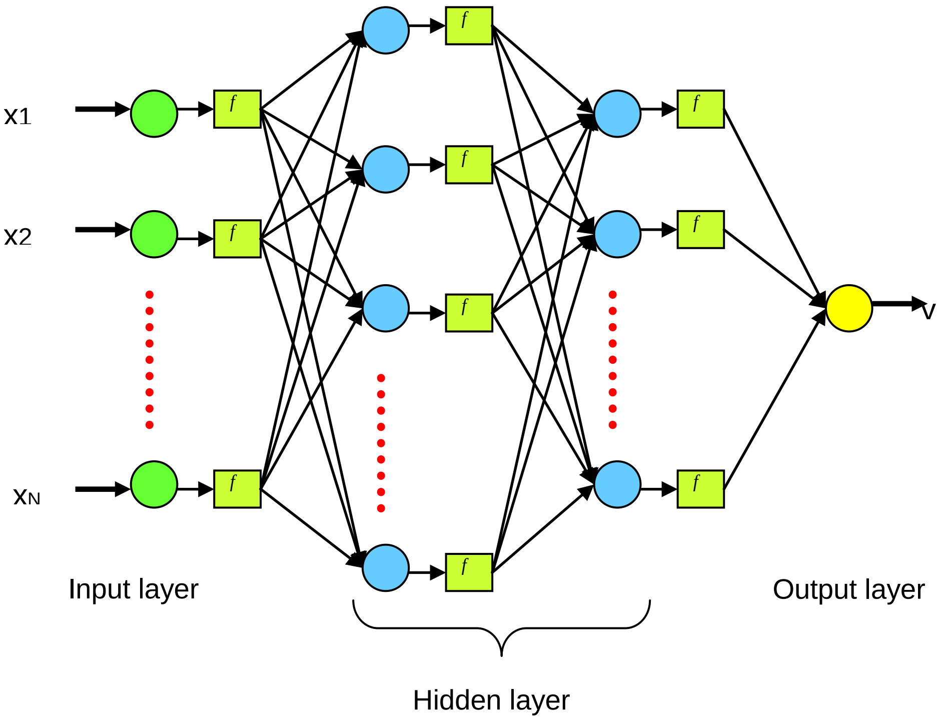
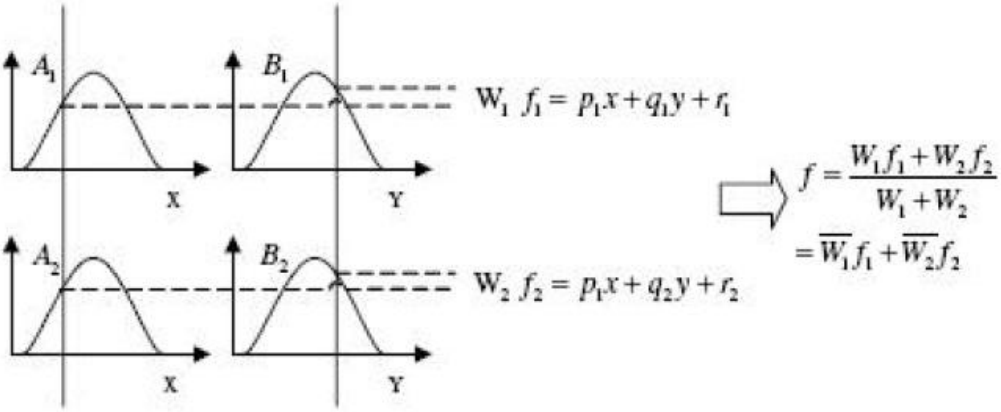


Figure 2. A multi-layer perceptron neural network architecture.



(a)

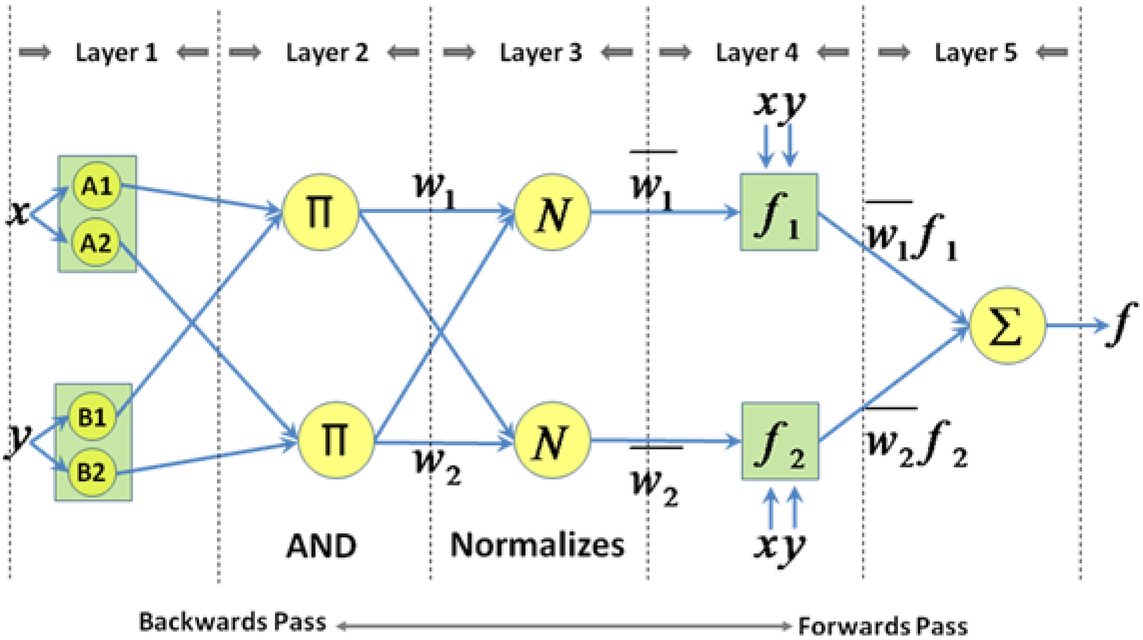


Figure 3. (a) A two-input first-order Sugeno fuzzy model with two rules; (b) An equivalent ANFIS structure.

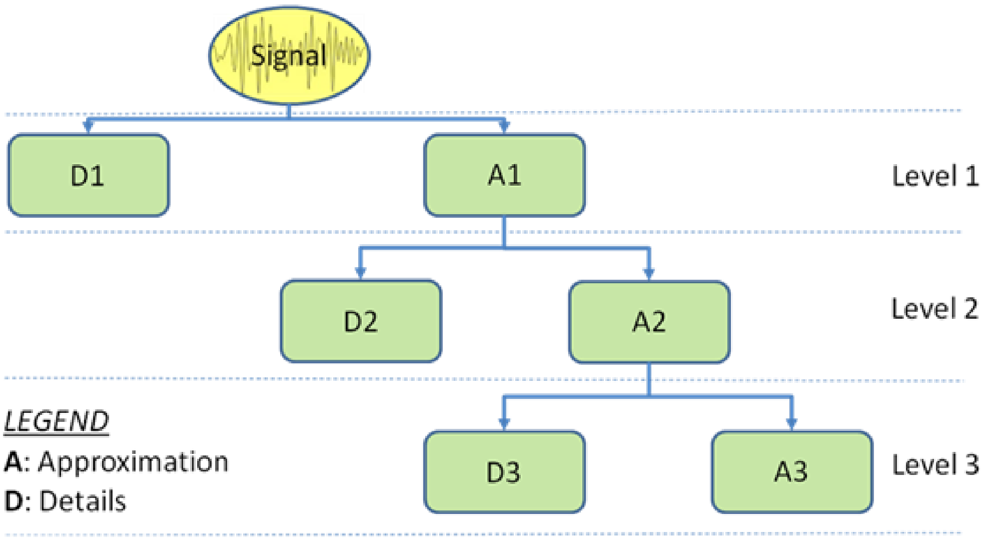


Figure 4. A schematic representation of the pyramid structure representing the WMRA.

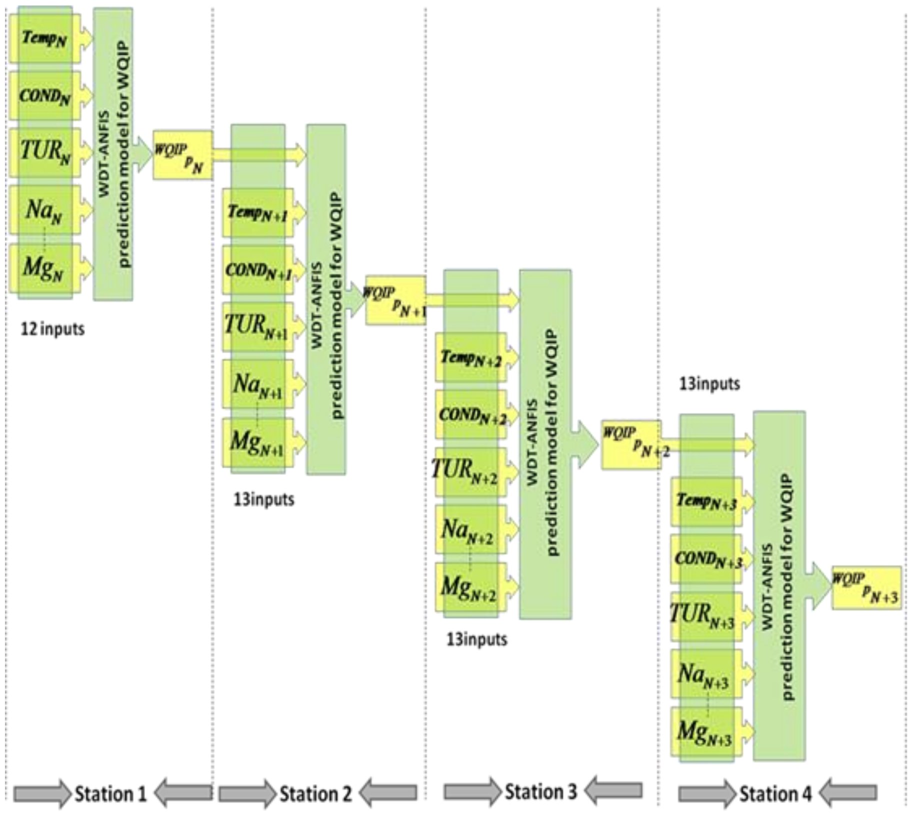


Figure 5. Schematic representation of the proposed networks for Scenario 2.

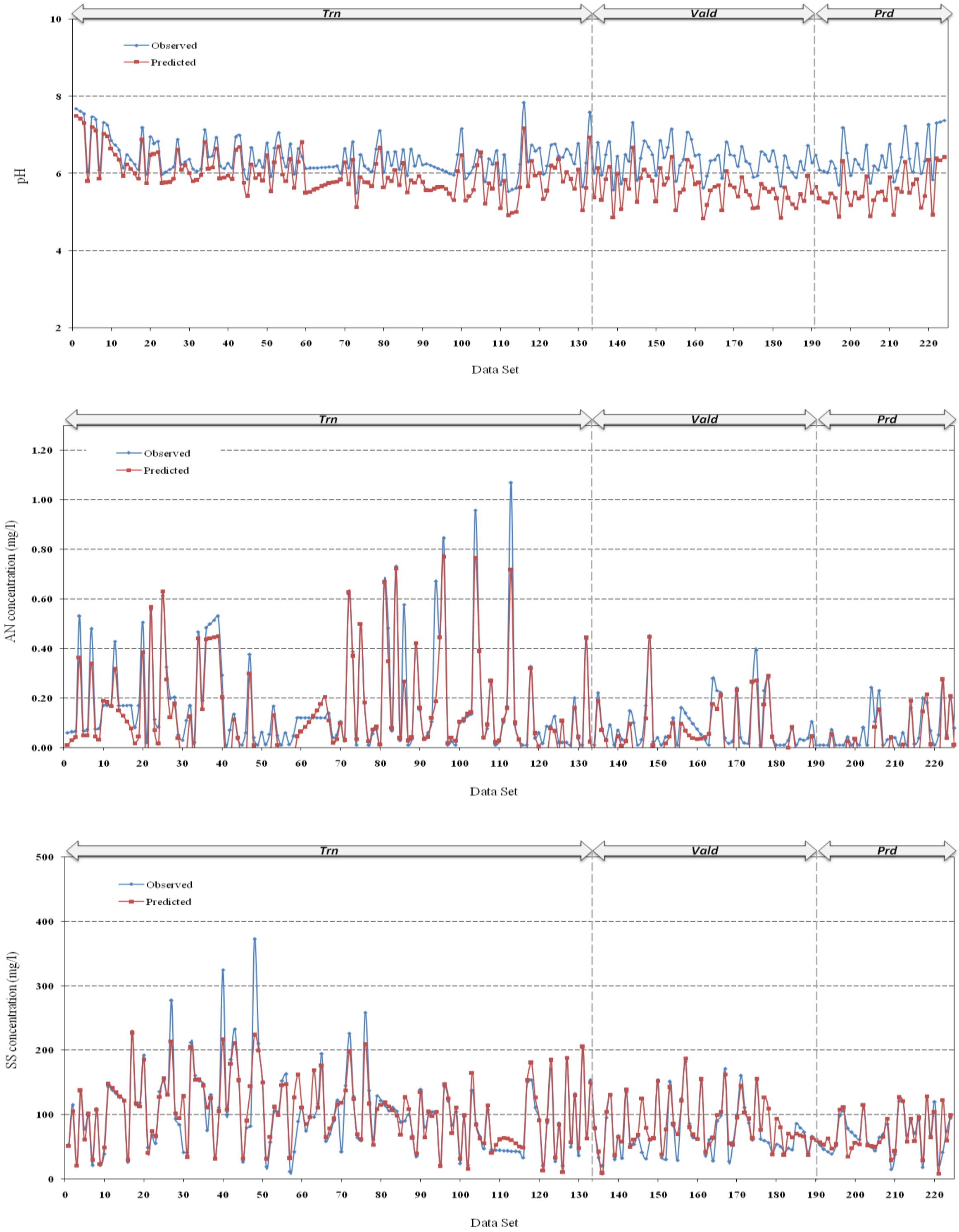
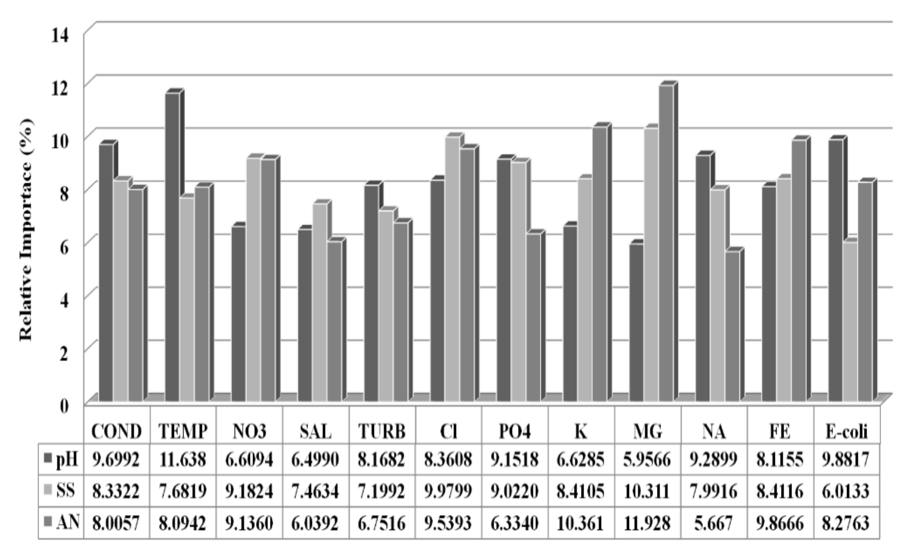


Figure 6. Performance of the MLP-ANN model: A comparison between the predicted and observed values.



	COND	TEMP	NO3	SAL	TURB	Cl	PO4	K	MG	NA	FE	E-coli
pH	9.6992	11.638	6.6094	6.4990	8.1682	8.3608	9.1518	6.6285	5.9566	9.2899	8.1155	9.8817
SS	8.3322	7.6819	9.1824	7.4634	7.1992	9.9799	9.0220	8.4105	10.311	7.9916	8.4116	6.0133
AN	8.0057	8.0942	9.1360	6.0392	6.7516	9.5393	6.3340	10.361	11.928	5.667	9.8666	8.2763

Figure 7. Relative importance of each input parameter.

Figure 8. Performance of the ANFIS model: A comparison between the predicted and observed values.

Figure 9. Performance of the WDT-ANFIS model: A comparison between the predicted and observed values.

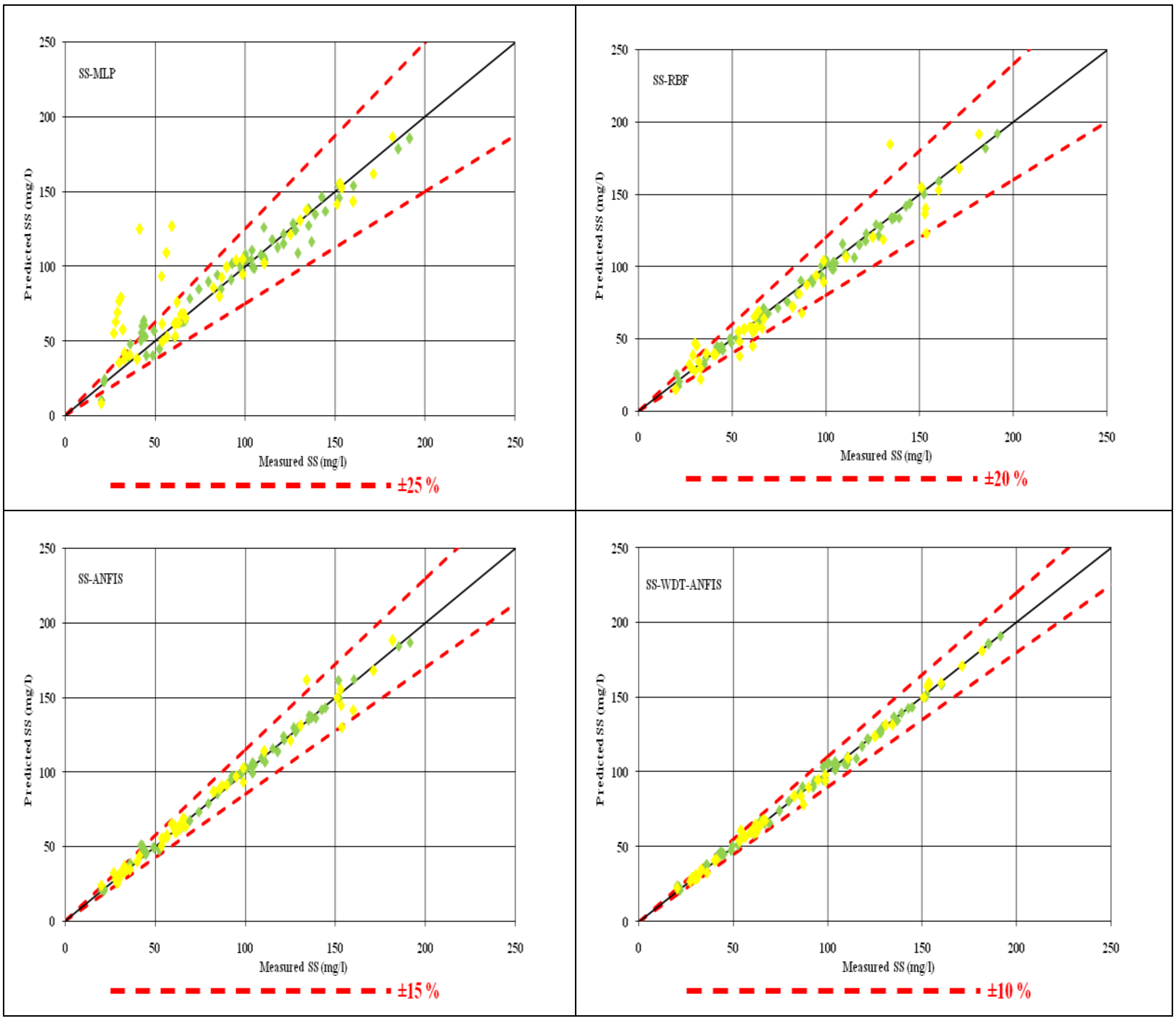


Figure 10. Comparison between the predicted SS versus the observed SS utilizing different techniques.

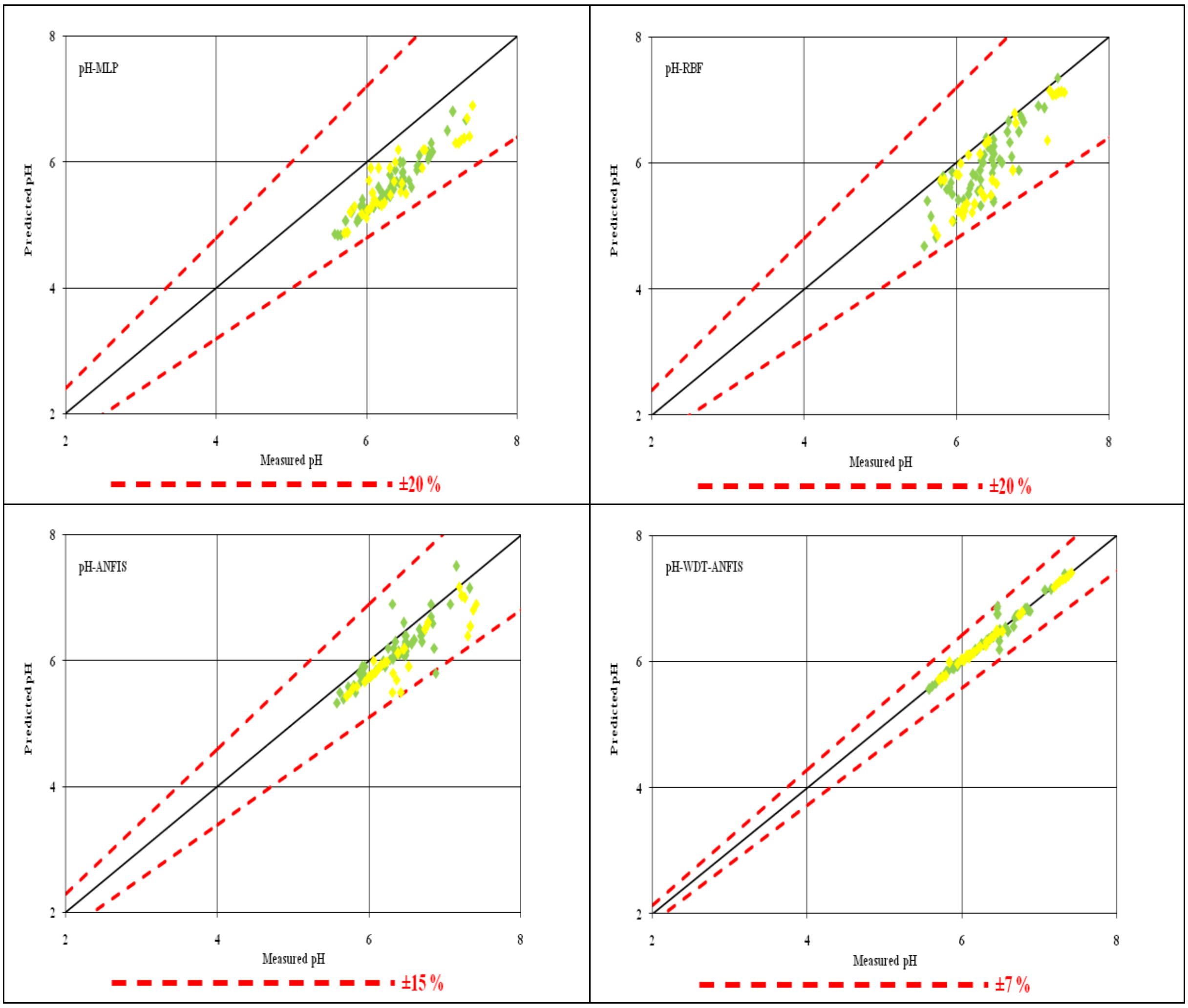


Figure 11. Comparison between the predicted pH versus the observed pH utilising different techniques.

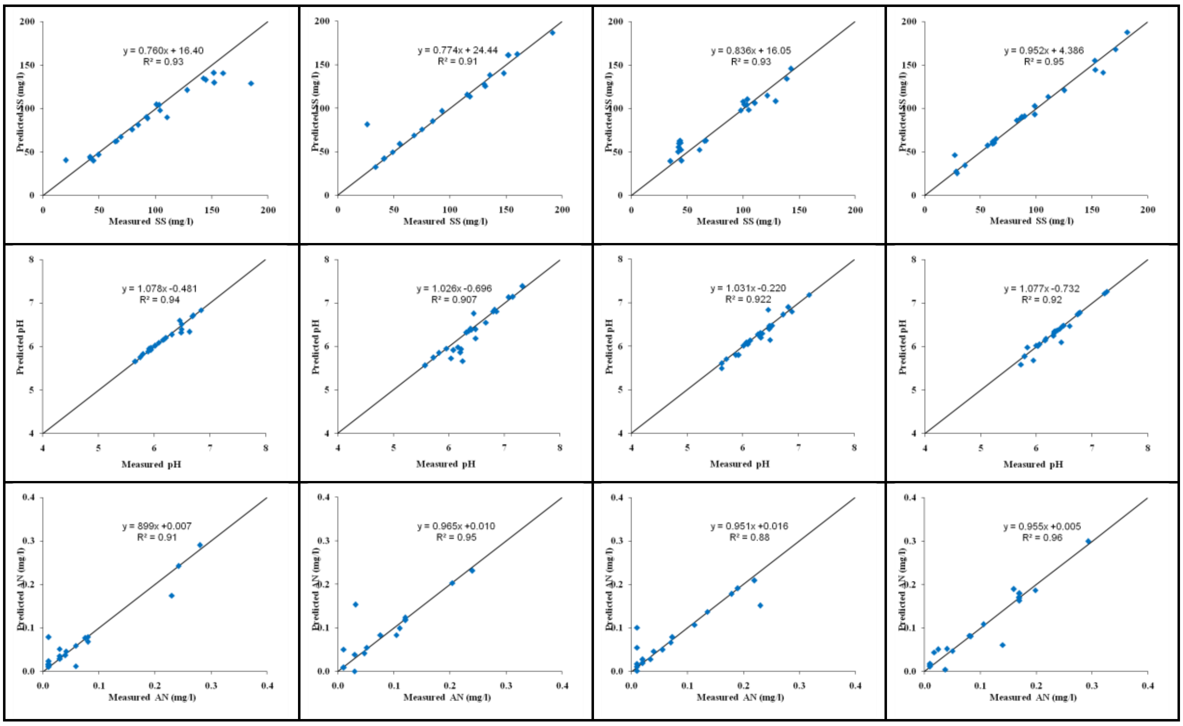


Figure 12. WDT-ANFIS model verification for each water quality parameter at each station.

Tables

Table 1. Input parameters used in previous studies for the ANN model.

Author(s) and year	Input variable	Location(s)
Rabia (Koklu, 2006)	BOD, Temp, Water discharge, NO2-N, NO3-N	N/A
Kuo *et al*. (Kuo et al., 2007)	pH, Chl-a, NH4N, No3N, temp, month	Te-Chi Reservoir, Taiwan
Ying *et al*. (Zhao et al., 2007)	Turbidity, Temp, pH, Hardness, Alkalinity, Chloride, NH_4-N, NO_2-N	Yuqiao reservoir, China
Palani *et al*. (Palani et al., 2008)	DO, Chl-a, temp	Singapore coastal, Singapore
Zaqoot *et al*. (Zaqoot et al., 2009)	Conductivity, Turbidity, Temp, PH, Wind speed	Mediterranean Sea along Gaza, Palestine
Singh *et al*. (Singh et al., 2009)	pH, TS, T-AlK, T-Hard, CL, PO4, K, Na, NH4N, No_3N, COD	Gomti, India

Table 2. Basic statistical analysis for input parameters.

	Unit	**Mean**	**Minimum**	**Maximum**	**SD**	**CV**
			SN01			
TEMP	o C	27.03	24.08	30.33	0.83	3.08
COND	µS	55.42	32.00	92.00	13.82	24.93
SAL	ppt	0.64	0.01	2.93	0.36	56.00
TUR	NTU	0.03	0.01	0.20	0.05	152.38
NO3	mg/l	163.50	15.50	775.00	130.61	79.88
CL	mg/l	5.27	1.00	18.00	2.49	47.16
PO4	mg/l	0.04	0.01	1.08	0.12	283.32
FE	mg/l	4.61	1.00	10.30	1.74	37.63
K	mg/l	0.87	0.10	2.40	0.44	50.59
MG	mg/l	3.13	1.22	11.54	1.42	45.18
NA	mg/l	0.87	0.08	2.32	0.44	51.20
E-COLI	cfu/100ml	3844.98	40.00	48000.00	6377.64	165.87
			SN02			
TEMP	o C	27.16	24.08	29.82	1.11	4.10
COND	µS	62.64	28.00	300.00	38.78	61.91
SAL	ppt	0.02	0.01	0.07	0.01	54.16

TUR	NTU	127.79	30.70	370.00	77.64	60.76
NO3	mg/l	0.73	0.12	5.55	0.69	93.53
CL	mg/l	5.66	1.00	24.00	3.28	57.89
PO4	mg/l	0.07	0.01	0.66	0.12	159.91
FE	mg/l	0.82	0.09	2.02	0.48	58.85
K	mg/l	4.63	0.90	7.80	1.56	33.76
MG	mg/l	0.80	0.10	1.40	0.33	40.69
NA	mg/l	3.27	1.40	26.70	3.33	101.77
E-COLI	cfu/100ml	2564.82	20.00	22000.00	3802.25	148.25
			SN03			
TEMP	o C	26.14	23	31.93	1.38	5.07
COND	µS	54.16	26.07	373.00	45.62	84.24
SAL	ppt	9.56	0.01	61.00	20.43	213.64
TUR	NTU	113.33	0.01	820.00	139.73	123.29
NO3	mg/l	11.55	0.00	133.00	27.26	236.03
CL	mg/l	5.43	0.06	20.00	2.78	51.13
PO4	mg/l	0.09	0.00	1.02	0.22	233.34
FE	mg/l	1.21	0.15	5.60	1.35	111.53
K	mg/l	3.87	0.40	7.00	1.66	42.84
MG	mg/l	1.03	0.20	5.20	0.82	79.40
NA	mg/l	3.23	1.00	20.80	2.69	83.17
E-COLI	cfu/100ml	3498.07	0.00	86000.00	11402.45	325.96
			SN04			
TEMP	o C	27.43	24.58	29.78	1.10	4.02
COND	µS	64.54	37.80	186.00	28.93	44.82
SAL	ppt	0.02	0.01	0.07	0.01	64.09
TUR	NTU	104.31	2.00	343.00	77.09	73.90
NO3	mg/l	0.66	0.06	3.22	0.40	61.13
CL	mg/l	7.32	2.00	28.00	5.60	76.50
PO4	mg/l	0.08	0.01	0.99	0.21	249.18
FE	mg/l	0.68	0.03	2.02	0.48	71.03
K	mg/l	4.03	0.40	6.40	1.22	30.30
MG	mg/l	0.94	0.20	2.90	0.54	57.05
NA	mg/l	4.15	1.60	24.00	3.79	91.28
E-COLI	cfu/100ml	4950.04	0.00	41000.00	7419.36	149.88

Table 3. Basic statistical analysis for three water quality parameters.

	Unit	Mean	Minimum	Maximum	SD	CV
			SN01			
PH	-	6.39	5.49	7.83	0.45	7.07
SS	mg/l	91.01	11.00	372.00	56.26	61.81
NH3-NL	mg/l	0.14	0.01	1.07	0.18	129.30
SN02						
PH	-	6.22	5.43	7.28	0.36	5.77
SS	mg/l	73.44	7.00	274.00	50.16	68.30
NH3-NL	mg/l	0.10	0.01	0.45	0.11	103.81
			SN03			
PH	-	6.36	5.67	8.41	0.48	7.59
SS	mg/l	72.61	1.00	574.00	83.44	114.91
NH3-NL	mg/l	0.15	0.01	2.46	0.38	254.94
			SN04			
PH	-	6.29	5.59	8.09	0.41	6.56
SS	mg/l	47.98	1.00	146.00	32.05	66.80
NH3-NL	mg/l	0.15	0.01	0.83	0.20	131.79

Table 4. Correlation coefficient between WQP and the input parameters.

	PH	**SS**	**NH3-NL**	**PH**	**SS**	**NH3-NL**	**PH**	**SS**	**NH3-NL**	**PH**	**SS**	**NH3-NL**
		SN01			SN02			SN03			SN04	
TEMP	0.316	-0.171	-0.137	-0.425	0.361	0.014	-0.022	0.090	0.083	-0.295	0.154	-0.076
COND	-0.029	0.301	0.208	-0.113	0.061	0.144	0.216	0.002	-0.069	-0.290	0.083	0.094
NO3	0.228	0.131	0.383	-0.364	-0.101	0.067	-0.183	-0.279	0.201	-0.264	-0.196	0.054
SAL	0.202	-0.043	0.393	0.835	-0.118	-0.115	0.844	-0.071	-0.028	0.757	-0.147	-0.073
TURB	-0.167	0.766	0.137	0.071	0.061	0.000	-0.079	-0.200	0.191	-0.008	0.131	0.221
Cl	-0.114	0.354	0.411	-0.063	0.287	0.084	0.146	-0.076	-0.316	-0.302	0.067	0.245
PO4	0.181	-0.148	0.065	0.025	0.121	-0.083	0.077	-0.114	0.454	0.088	0.052	0.569
K	-0.306	0.184	0.253	-0.005	0.014	-0.108	-0.012	0.039	0.018	0.325	0.013	-0.248
MG	0.038	0.191	0.376	0.247	-0.023	0.152	0.115	-0.104	-0.192	0.020	-0.074	0.142
NA	0.127	0.088	0.400	0.106	0.283	0.077	-0.027	0.104	0.269	-0.268	0.176	0.025
FE	0.023	-0.080	-0.038	-0.165	0.143	-0.001	0.152	-0.045	0.017	-0.345	-0.024	0.106
E-coli	-0.085	0.315	0.007	0.142	0.024	0.014	0.223	-0.095	0.036	-0.042	0.143	0.367

Table 5. ANN architecture for each parameter.

Parameter	No. of neuron	RMSE	Maximum error (%)	TFHL	TFOL	TA
pH	18	0.15	3.22	TS	PL	LMA
SS	17	0.30	3.46	LS	PL	LMA
AN	17	0.26	3.12	TS	PL	LMA

TFHL: Transfer function between input layer and hidden layer; TFOL: Transfer function between hidden layer and output layer; TA: Training algorithm; LS: Log sigmoid; TS: Tan sigmoid; PL: Pure-line; LMA: Levenberg–Marquardt algorithm.

Table 6. The number and types of MFs for each module.

Parameter	AFNIS Module	
	MFs (Type)	MFs (Number)
PH	gbellmf	3 4
SS	gbellmf	4
NH3-NL	gbellmf	3 4 4

Table 7. The running time (seconds) of training process for each model

Model	MLP	RBF	ANFIS	WDT-ANFIS
pH	51	44	67	78
SS	53	46	71	81
AN	49	43	64	75

Table 8. A summary of correlation coefficients for Scenario 1, Scenario 2 and the AI %.

Model	SNO2		SNO3		SNO4		AI (%)		
	Scen1	Scen2	Scen1	Scen2	Scen1	Scen2	SNO2	SNO3	SNO4
pH	0.95	0.98	0.94	0.98	0.93	0.98	3.1	4.1	5.1
SS	0.96	0.97	0.97	0.98	0.97	0.98	1.1	1	1
AN	0.96	0.97	0.96	0.97	0.95	0.97	0.5	0.5	2

1. Introduction

Rivers are considered as one of the most critical sources of water for irrigation purposes, industrial needs and other uses. The dynamic nature of the river systems and their easy accessibility for waste disposal make the river systems most vulnerable to the adverse effects of environmental pollution. The term "water quality" refers to the state or condition of water, which takes into account the physical, chemical, and biological properties of the water. In conducting the study of any aquatic system, modelling the water quality parameters is of utmost significance. Evaluation and prediction of the surface water quality is necessary for effective management of river basins so that sufficient measures can be adopted to ensure that the pollution levels remain within permissible limits. Accurate prediction of future phenomena in relation to the water quality is the essence of optimal water resources management. The conventional process-based modelling methods offer comparatively accurate predictions for water quality parameters. However, these models have limitations as they depend on data sets that require a substantial amount of processing time and a huge amount of input data that is often unknown.

Nearly 60% of the major rivers in Malaysia are used for agricultural, household and industrial applications (DID, 2000). As per Rosnani Ibrahim (Ibrahim, 2001), the major sources of pollution that affect these rivers are dumping of sewage, waste releases from medium and small-sized industries not having proper waste matter treatment equipment, clearing of land and groundwork activities. On the basis of the records of 1999, 50 catchments (that is 42% of river) were contaminated with SS (suspended solids) caused by badly planned and unregulated earth clearing attempts and 33 catchments (that is, 28% of river) were polluted with AN (ammoniacal nitrogen) from activities related to cattle breeding and household sewage dumping.

Johor River is regarded as somewhat polluted as per DOE (Department of Environment)(DOE, 2007) because of the developmental activities alongside the bank of the river. Moreover, the river continues to be chocked and dumped by waste and litter due to lack of enforcement by the local administration. These pollutants ultimately end up in the Joho River tributaries, rich areas for nourishment and breeding of poultry and fish. Consequently, several statistical frameworks and computer simulations must be introduced as powerful and critical tools for planning and monitoring the maintenance of the water bodies.

Growing concerns regarding environment, along with scarce funding, are giving rise to a growing interest in cost-effective and judicious strategies for the management of water quality. Since the quality of water directly affects the health of the humans, quality improvement of the water accessible for human use will play a significant role in decreasing health related hazards.

The project of water pollution regulation is based on the management of water quality. It estimates the kind of water quality from the present water quality condition, as well as from the rules of disposal of the pollutants into the river. Moreover, many models for water quality, like stochastic and deterministic models, have been created so as to provide best processes to conserve the quality of water (Hull et al., 2008). Nevertheless, getting efficient and precise water quality model in complex water resources is still difficult because of the variations and complications in the actual world, the ambiguities in the framework and variables of the model, and the deviations in the field data. Thus, conventional methods for data processing are not sufficiently efficient anymore for solving issues related to the water quality. Additional efforts are required to improve the consistency of the findings of the model.

Deterministic models try to represent all the chemical and physical processes included in statistical terms, with variables acquired either from past data or obtained empirically, or computed by experience or examination. Generally, the differential equations are simplified so as to find solutions suitable for the model. Solution of the involved equations may need suppositions and simplifications which are derived from the performance of the model, and usually practical experience is necessitated from the user prior to achievement of optimal outcomes.

Statistical models attempt to seek general rules from the experimental data, which can be done by obtaining information from the field data. Statistical modelling and assessment involve a meticulous selection of techniques for analysis, and validation of suppositions as well as data. A majority of such models are quite complex and involve a substantial field data amount to conduct the analysis. Moreover, several statistical-based models of water quality, which assume the association among the prediction and the response variables, are distributed normally and linear in nature. Nevertheless, since the quality of water can be impacted by several parameters, conventional techniques for data processing are not sufficiently efficient anymore for solving this issue, and as such parameters show a complex non-linear relation to the water quality prediction parameters. Thus, using statistical techniques generally does not have high accuracy.

Of late, the AI (Artificial Intelligence) approach has been recognised as an effective alternative method for modelling of complicated non-linear systems. Generally, such models do not take into account the internal process but develop models through the inputs and outputs correlation. Presently, AI is used exhaustively for estimating several water-related regions (Muttil and Chau, 2006).

Recently, AI has offered the techniques for operation optimisation and selection of equipment, and problem solving that involve large quantities of data that cannot be processed by humans for the purpose of decision making. For this purpose, AI methods are proficient to replicate this behaviour and balance the deficiency. Thus, the growth of technology of efficient parallel computing and growing computing power have facilitated the researchers to employ the AI approaches (for instance, ANN (Artificial Neural Network) and ANFIS (Adaptive Neuro-Fuzzy Inference System)) for field data modelling solutions. The neuro-fuzzy technique has been used effectively in certain fields of water bodies engineering like the rainfall-runoff model (Chang and Chen, 2001)and basin operation (Chang and Chang, 2006; Chang et al., 2005). ANFIS has been known to enhance the accuracy of day-to-day estimation of evaporation (Kişi, 2006), reservoir water level prediction (Chang & Chang, 2006) and prediction of the river flow (Firat and Güngör, 2007).

The data obtained from experimentation and examination may be corrupted by signals of noise because of objective and/or subjective errors. For instance, experimental faults may be caused by measuring, recording, reading and external situations. As this noise can possibly distort the model outcomes, it is essential to eliminate them (i.e. signal de-noising) prior to the use of this data. The noisy signals can be de-noised by applying a series of linear filters (Bell and Martin, 2004). Nonetheless, these filters are more suitable for linear systems rather than the non-linear ones. Moreover, the FAT (Fourier analysis technique) is a standard tool for de-noising, though it is only favourable for de-noising signals or data involving stable noises. In addition, as there are unstable noises in real situations, it cannot be applied effectively. Thus, to solve the issues of conventional de-noising methods, more complex methods, like the WDT (wavelet de-noising technique), have been recommended. Above all, WDT is effective for de-noising multi-dimensional temporal or spatial signals having stable or unstable noises. Also, it has been extensively applied to industrial systems for information finding and patterns recognition (Avci, 2007; Tirtom et al., 2008). Nonetheless, some of these investigations were employed for water quality monitoring, where its data was utilised for estimation of parameters (Dohan and Whitfield, 1997).

In Malaysia WQIP requires extensive calculations and transformations. Two studies have been proposed to use Artificial Intelligence techniques (AI) in Malaysia in order to develop an accurate predictive model to WQP. However, many studies show that AI needs pre-processing tool to enhance the accuracy of the model practically in dealing with measured water quality data which is often contain noise (Han et al. 2011, Xu and Liu 2013).

The main objective of this investigation is to evolve a computationally proficient and robust method for the estimation of water quality variables decreasing the labour and cost for measurement of those parameters. This study focuses on the Malaysian Johor River situated in Johor State where the water quality dynamics are significantly altered. This research has many primary objectives, as follows:

- To evaluate and assess the correlation among the parameters of water quality on the basis of the experimental data using ANN (Artificial Neural Network).
- To propose various ANN approaches, like MLP (Multi-Layer Perceptron) Neural Network and RBF (Radial Basis Function) Neural Network so as to confirm the effectiveness of these techniques in the estimation of the parameters of water quality.
- To get familiar with the correctness of the ANFIS (Adaptive Neuro-Fuzzy Inference System) in the prediction of the parameters of water quality.
- To develop an augmented WDT-ANFIS (wavelet de-noising technique with the Neuro-Fuzzy Inference System).
- To examine the effectiveness of the suggested model for spatial position by presenting two different situations: the first situation (Scenario 1) is designed to set the model prediction at each station pertaining to the water parameters by considering the 13 input parameters from the same station. Where for Scenario 2, the input parameters for this scenario based on the measured water quality parameters from the same station and the predicted parameter from upstream station.

- To validate the augmented WDT-ANFIS (wavelet de-noising technique with the Neuro-Fuzzy Inference System) based on the experimental data for the duration 2009-2010.

3. Case Study: Johor River Basin

Johor state is regarded as the third largest region in Malaysia with an area of 19.984 km^2. It comprises of eight districts namely are Kota Tinggi, Muar, Pontian, Johor Bahru, Segamat Kluang, and lastly Batu Pahat which is considered as the second largest districts in Johor with an area of 187,702.06 hectares. Johor state has five principal rivers which are Sungai Muar, Sungai Johor, Sungai Endau, Sungai Batu Pahat and Sungai Sedilfi. This research sheds the light solely on Sungai Johor river. The Johor river basin is located in the southeast of Peninsular Malaysia. At an altitude of 1010 m, the Johor river orginates from the Gunung Belumut and from Bukit Gemuruh at an altitude of 109 m un the north. The river has irregular shape, its drainage area is around 2636 km2 and its length is approximately 122.7 Km. The river flows southeast into the Johor straits. An average annual precipitation of 2470 mm added to the river while during the period of 1963-1992, the annual mean discharge at Rantau Panjang was found to be 37.5 m3/s. The Johor river and its tributaries play a significant role as water suppliers for the state of Johor as well as for Singapore. Many factors contribute to the deterioration of the water quality of Johor River, mainly include the release of different kinds of pollutants at levels exceeding the allowed limits with the absence of local authorities' enforcement. These pollutants travel through Johor River and ultimately end in the estuaries of the rivers which are known to be a natural feeding area for poultries and fishes and a natural environment that provide spawning. Figure 1 depicts the location map of the surveying area which compromises of four monitoring stations on Johor River.

Figure 1.

3. Methodology

3.1 Multi-Layer Perceptron Neural Network (MLP-ANN)

A feed-forward network is the multi-layer perceptron neural network (MLPNN) that includes many layers of neurons, where one neuron's output is propagated to the other neuron's input that is in the next layer. Figure 2 presents the multi-layer perceptron neural network. In MLPNN, the input layer's nodes only propagate the input values of the first hidden layer's nodes. In the hidden layers, each node's input-output relationship can be presented as follows:

$$y = f\left(\sum_j w_j x_j + b\right) \qquad (1)$$

where, x_j signifies the output from the previous layer's j node, w_j denotes the connection weight between the current node and j node, b represents the current node's bias, and f defines a non-linear transfer function usually of the sigmoid form as shown in Equation (3.4):

$$f(z) = \frac{1}{1+\exp(z)} \qquad (2)$$

where, z denotes the weighted sum pertaining to the input to the neuron and $f(z)$ signifies the neuron output. The output nodes' input-output relationship is comparable to the one defined by Equation (3.4), with the exception of the case where the network is employed for function approximation, and the type of function f could vary (e.g. linear function).

Figure 2.

The units define a MLPNN architecture, which allows computation of a non-linear function in terms of the scalar product pertaining to the weight vector and input vector. Overall, the MLPNN models' performance relies on the network's inherent architecture. Apart from the number of hidden layers as well as the number of neurons pertaining to each layer, it also includes the computation type applied to each neuron.

3.2 ADAPTIVE NEURO-FUZZY INFERENCE SYSTEM (ANFIS)

Jang (Jang, 1993) first put forward the Adaptive Neuro-Fuzzy Inference System (ANFIS) that allowed realising a highly non-linear mapping and compared with common linear methods, it is considered to be superior in yielding non-linear time series (Jang, 1993). The ANFIS architecture was employed throughout this research for the first-order Sugeno fuzzy model (Sugeno and Kang, 1988). ANFIS can be defined as a multi-layer feed-forward network that employs neural network learning algorithms as well as fuzzy reasoning to aid in mapping input space with that of the output space (Chang and Chang, 2006). Considering that for a first-order Sugeno fuzzy model, the fuzzy inference system has one output, f, and two inputs, x and y, a common rule set that includes two fuzzy ‘if.then’ rules can be defined as follows:

$$\text{Rule 1: If } x \text{ is } A_1 \text{ and } y \text{ is } B_1, \text{ then } f_1 = p_1 x + q_1 y + r_1 \tag{3}$$
$$\text{Rule 2: If } x \text{ is } A2 \text{ and } y \text{ is } B_2, \text{ then } f_2 = p_2 x + q_2 y + r_2 \tag{4}$$

where, A_1, A_2 and B_1, B_2 signify the membership functions (mfs) pertaining to inputs x and y, respectively; p_i, q_i and r_i (i = 1 or 2) represent the linear parameters pertaining to the first-order Sugeno fuzzy model’s consequent part. Figure 3(a) represents the fuzzy reasoning mechanism pertaining to this Sugeno model that also allows deriving the output function (f) from that of inputs x and y. Figure 3(b) presents the corresponding equivalent ANFIS architecture, in which similar functions are associated with the same layer’s nodes. ANFIS comprises five layers as stated below:

Figure 3.

3.3 WAVELET DE-NOISING

The next logical step is characterised by wavelet analysis post the short-time Fourier transforms (STFT). This is with regards to the windowing technique that includes variable-sized regions. With the help of wavelet transform (WT), long time intervals can be employed in those areas where more precise low frequency information is needed, as well as for shorter regions in which high frequency information is needed. Overall, the key benefit provided by the wavelets is allowing conducting local analysis for localised area pertaining to a larger signal. The discrete-time WT pertaining to a time domain signal $x[k]$ can be expressed as follows (Dohan and Whitfield, 1997):

$$DWT(m,n) = 1/_{\sqrt{2^m}} \sum_k x[k]\psi[2^{-m}n - k] \qquad (5)$$

Here, (n) defines the mother wavelet, while m represents the scaling and k denotes the shifting indices. The *DWT* logarithmic frequency coverage is provided through scaling, as opposed to the uniform frequency coverage of STFT. This analysis technique includes segmenting a signal into components at various frequency levels, which are linked by the powers of two (a dyadic scale). The filtering approach that is applied to multi-resolution WT involves formation of a series of half-band filters that segment a spectrum into low and high frequency bands. The formulation is based on a wavelet function or high-pass (UP) filter as well as a scaling function or low-pass (LP) filter. Wavelet multi-resolution analysis (WMRA) allows constructing a pyramidal structure that needs an iterative application of wavelet functions and scaling to high-pass and low-pass filters, respectively. At the beginning, these filters are first applied to the entire signal band under high frequency (small-scale values) and then the signal band is decreased at every stage gradually. As presented in Figure 4, the detail coefficients of Dl, D2 and D3 define the high-frequency band outputs, while the approximation coefficients of Al, A2 and A3 define the low-frequency band outputs.

Figure 4.

Numerous factors need to be accounted when wavelets are employed to de-noise the WQP data. Examples of such choices include the level of decomposition, wavelet and thresholding methods to be employed. MATLAB provides various families of wavelets such as Morlet, Meyer, Mexican hat, Coiflets, Haar, Symlets, Daubechies and Spline biorthogonal wavelets, and also offers additional documentation regarding these wavelet families

("Wavelet Toolbox - MATLAB," n.d.). Only orthogonal wavelets need to be accounted to get perfect reconstruction results. Certain advantages are associated with the orthogonal wavelet transform. It can be characterised as being relatively concise, permitting perfect reconstruction of the original signal and relatively easy to calculate. The two common employed approaches for thresholding a signal include hard thresholding and soft thresholding, which are employed in the MATLAB wavelet toolbox. Although the easiest method is hard thresholding, better results are achieved through soft thresholding versus hard thresholding. Thus, this study uses soft thresholding. Four threshold selection rules can be used with the wavelet toolbox, which employ statistical regression pertaining to the noisy coefficients over time that allows getting a non-parametric estimation regarding the reconstructed signal absent noise. This study examined just Sqtwolog, wherein a fixed form of threshold is employed, leading to minimax performance that is multiplied by a factor proportional of signal length's logarithm. In this research, in terms of the decomposition level, we can conclude that a level 4 decomposition offered reasonable results post applying the trial-and-error method to all modules.

3.4 Model Performance Evaluation

It is necessary to clearly recognise the criteria that are associated with judging the model's performance. The criteria employed to assess the performance of the model in this study were clearly mentioned in the literature. Dogan et al. (Dogan et al., 2009) employed the Average Absolute Relative Error (AARE), which not only provides the performance index with regards to predicting water quality parameters but also demonstrates the prediction errors distribution. To examine the performance of the model, Singh et al. (2009) employed the bias statistical index. The bias signifies the mean of all the individual errors as well as allows determining if the dependent variable is underestimated or overestimated by the model. In this study, correlation coefficient as well as Root Mean Square Error (RMSE) was employed to examine the model's performance (Soyupak et al., 2003; Zhao et al., 2007).

Usually, the model performance is assessed through coefficient of determination, as put forward by Nash and Sutcliffe (1970), while MSE is employed to check the level of fitness between the network output and desired output.

In this research work, the models' performances were assessed based on three statistical indexes. As mentioned by Nash and Sutcliffe (1970) coefficient of efficiency (CE) is commonly employed to assess the performance of the model.

$$CE = 1 - \frac{\sum_{i=1}^{n}(X_m - X_p)^2}{\sum_{i=1}^{n}(X_m - \overline{X}_m)^2} \qquad (6)$$

where n represents the number of observations, X_m and X_p define the measured and predicted parameters, respectively, and $\overline{X}_m$ signifies the average of measured parameter.

Mean square error (MSE) is employed to see the level of fitness between network output and the desired output. Better performances are guaranteed with smaller MSE values. It is defined as follows:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(X_m - X_p)^2 \qquad (7)$$

More commonly, the coefficient of correlation (CC) is employed to examine the linear relationship between the measured and predicted dissolved oxygen. This can be expressed as follows:

$$CC = \frac{\sum_{i=1}^{n}(X_m - \overline{X}_m)(X_p - \overline{X_p})}{\sqrt{\sum_{i=1}^{n}(X_m - \overline{X}_m)^2 \sum_{i=1}^{n}(X_p - \overline{X_p})^2}} \quad (8)$$

Further, for visual comparison of the predicted and measured values, the Scatter plot was employed (Kuo et al., 2007).

3.5 Input Variables and Data Processing

One of the key functions of ANN is to identify the model input parameters that could impact the output parameters considerably. As indicated above, the selection of input parameters depends on a priori knowledge regarding causal variables as well as statistical analysis pertaining to the potential outputs and inputs. In the literature, different input parameters were employed to develop the model to determine water quality parameters, as presented in Table 1.

Table 1.

On the basis of the literature, the following water quality parameters were chosen for ANN modelling: temperature (Temp), electrical conductivity (COND), salinity (SAL), nitrate (NO3), turbidity (TURB), phosphate (PO4), chloride (CI), potassium (K), sodium (Na), magnesium (Mg), iron (Fe) and Escherichia coli (E-coli). The basic statistical parameters, i.e. mean, minimum, maximum, standard deviation (S.D.), and coefficient of variation (CV) of the input and output parameters deployed in this study are depicted in Table 2 and Table 3.

Table 2.

Based on the concentration levels of both output and input parameters, large changes between the samples were seen, along with a high coefficient of variation (i.e. 254.94% for AN and 325.96% for E. coli). The coefficient of variation (CV) can be defined as a measure of statistical dispersion pertaining to the data. For a given data set, it is the mean normalised standard deviation (CV %) that can be computed as (standard deviation/mean) × 100. The existence of large disparity in the parameters' concentrations can be attributed to the types (non-point and point) and nature of sources that have been distributed in the river basin's wide geographical area. During the course, the river flows through different townships, and many tributaries and wastewater drains pouring large quantities of untreated wastewater into the river's main channel. A coefficient of variation in the range of 3.08% and 325.96% was seen with the parameters. Such variability that exists amongst the samples could be due to large geographical variations in climate as well as seasonal effects pertaining to the study region. For the various sampling sites, a spatial and significant variation was seen in terms of Johor River's turbidity, which varied from 0.2 to 343 NTU. It was higher, which could because of the mixing of industrial effluents and domestic sewerage water in Johor River. The rise in turbidity near downstream sites can be attributed to settling factors and flow turbulences. At downstream sites, the observed trend of turbidity, i.e. SN02, SN03 and SN04, was seen to support the above-mentioned hypothesis. Comparable patterns pertaining to spatial variations in turbidity were reported by (Khadse et al., 2007) when investigating Kanhan River's water quality. Amongst the sampling sites, the conductivity of the Johor River water was found to be considerably different, in which the mean ranged from 54 to 64 μS, although least significant difference was between SN01 and SN03. The high conductivity at SN04 and SN02 sites signify sewerage mixing into the river water. The dilution of industrial and urban runoffs could be attributed to the lower conductivity seen in the downstream water. Nitrate is considered to be a crucial parameter of river water that could be an indicator for the pollution status and anthropogenic load in river water.

The mean of nitrate ranged from 0.66 to 163.5 mg/l for Johor River. At the site wherein urban runoff mixing was noticed, NO3 was seen to be the maximum. It is interesting to note that in the downstream non-point pollution sites, lower NO3 was seen. The concentration of chloride in water was deemed not to be harmful. A higher concentration of chloride found in freshwater signified that pollutants are present. Moreover, in Johor River, the chloride level fell in the range of 5.27 to 7.37 mg/l. Nonetheless, at various sampling sites, a clear trend was not seen with chloride concentration in terms of the non-point or point pollution sites. The mixing of industrial effluents or urban wastewater in the river water is signified by higher levels of chloride content at SN04.

Table 3.

pH of water indicates alkaline and acidic conditions. DOE (DOE, 2007) suggested that pH for water in the range of 6.5–8.5 can be employed for any purposes in that respect; the ranges showed that Johor River had moderately alkaline water. The change in mean pH ranged from 6.22 to 6.36 at various locations. At some sites, higher pH could be a result of carbonate and bicarbonates of magnesium and calcium in water. The key source pertaining to such chemicals include industrial wastewater or urban runoff. SS further signifies the river water's salinity behaviour. The mean SS content pertaining to river water was found in the range of 72.61 to 91.01 mg/l. The chemical and biological oxygen demand increase in tandem with higher SS level in the water system, which ultimately results in depletion of the dissolved oxygen level in water. In water, SS stems from natural sources, industrial wastewater, urban runoff, sewage and chemicals employed in the water treatment process.

For the current neural network modelling, the second assessment of selecting the input parameters is done by considering a statistical correlation analysis pertaining to the field data. Calculation of the correlation coefficient existing between the input and output parameters was done and listed in Table 4.

Based on the table, pH was clearly seen to be inversely associated with water temperature (r = -0.306) as well as potassium (r = -0.425). We performed an experiment by taking water quality variables that were accounted along with the parameters mentioned above pertaining to various models to realise the optimal predictive model as well as reduce the monitoring cost by accounting for fewer input parameters.

Table 4.

3.6 Stopping Criteria

Normally, there is a gradual decrease in the training error of AI since the training process is on-going. Nonetheless, this minimisation of training error does not guarantee enhancement of generalisation ability, which gained our interest. It is not necessary that AI showing good performance with the training set will do the same with the testing data. Therefore, it is also sometime important to stop the training phase at the right time before over-fitting occurs. When a generalisation characteristic is lost by the neural network, an over-fitting issue follows. However, relations between the training inputs as well as their associated outputs to similar hidden patterns pertaining to the unobserved data cannot be generalised. Thus, this occurs as a result of a difficult question that asks how long a network needs to be trained. The issue of over-fitting is usually solved by employing techniques like weight elimination, weight decay and early stopping. Stopping criteria is the most commonly employed method to address this issue. As noted by numerous researchers (e.g. Singh et al. (Singh et al., 2009);

Palani et al. (Palani et al., 2008)), two frequently employed stopping criteria include stopping post a specific number of runs via the complete training data (it needs to be noted that an epoch is defined as each run that passes through the complete training data) and stopping on reaching some low level by the target error.

3.6. Different Scenarios

Two different scenarios have been proposed in this study. The concept behind the development of these both scenarios is based on the spatial pattern of the input-output structure of the model. Mainly, the reason behind proposing these scenarios is to examine the model performance considering the spatial dimension of the model input. Keeping in mind that the model output in both scenarios is the prediction values of the AN, pH and SS, the input patterns has been changed in terms of the number of the inputs and location of the monitored data. In order to clarify the structure and show the difference between these two scenarios, an example for the structure of both scenarios to predict the AN parameter will be presented. For scenario I, to predict AN parameter at certain station, different twelve input parameters were used that have been acquired at the same station. While, the structure of scenario II is developed as, in addition to the same twelve water quality parameters used as inputs in scenario I, the value of AN parameter that has been acquired from the upstream station will be added.

The prediction procedure can be defined as an operation that allows offering water quality parameter patterns for the future. This research employs the WDT-ANFIS along with its stochastic and non-linear modelling capabilities to design a prediction model that mirrored the water quality parameter patterns pertaining to Johor River with regards to the 12 input parameters (Scenario 1) cited earlier, which is represented as follows:

$$WQIP_N = f_{WDT-ANFIS}(Temp_N + COND_N + SAL_N + TUR_N + NO_{3N} + CI_N + PO_{4N} + Fe_N + K_N + Mg_N + Na_N + E-coli_N) \quad (9)$$

$$N = 1,2,3,4$$

Where $WQIP_N$ signifies the water quality index parameters pertaining to station *N*, and $f_{WDT\text{-}ANFIS}(.)$ defines the non-linear function predictor built via the WDT-ANFIS network. Thus, at each station, four models were built for predicting the parameters for water quality. A majority of the recent studies were aimed at predicting the concentrations pertaining to the parameters of water quality at every station. Usually, discharge via the local area from the upstream station causes an impact on the water pollution pertaining to a downstream station (Zaqoot et al., 2009). Therefore, in the put forward model, it was important to consider the impact cast by water parameters at the upstream station. Thus, the second scenario (Scenario 2) was designed to set the model prediction at each station pertaining to the water parameters by considering the 13 input parameters. At the previous station (upstream), the predicted WQIP could be represented by following Eq. (10). Repetition of this procedure involving the predicted WQIP is done for the fourth and third stations at downstream. Figure 5 presents a schematic representation pertaining to the put forward networks for Scenario 2.

$$WQIP_{N+1} = f_{WDT-ANFIS}(Temp_N + COND_N + SAL_N + TUR_N + NO_{3N} + CI_N + PO_{4N} + Fe_N + K_N + Mg_N + Na_N + E-coli_N + WQIP_{pN}) \quad (10)$$

Figure 5.

7. Results and Discussion

7.1 MLP-ANN Training

The construction of an ANN model normally includes three steps. The training stage is the first step, in which the network is exposed to a training set pertaining to the input-output patterns. The second step involves the validation stage, in which the network's performance is evaluated when patterns are not 'observed' by the network in the training stage. The third step includes the testing stage, in which the network's performance is evaluated when the unknown patterns were not 'observed' during the stages of validating and training (Bowden et al., 2005). Designing of three MLP-ANN architectures was done (one for each parameter). The Levenberg-Marquardt back propagation algorithm (LMA) is employed by all three networks in the entire training procedure. This study employed three activation functions, namely tan-sigmoidal (Tansig), log-sigmoidal (logsig) function and linear transfer function (purelin). After initialising the network weights and biases during the training process, iterative adjustments of the weights and biases pertaining to the network were carried out to decrease the network performance function pertaining to mean square error (MSE) – the average squared error between the target outputs and the network outputs.

We introduced different values of learning rate (lr) to the networks in a bid to achieve the optimum result pertaining to this study. For back propagation learning algorithm, the learning rate is important as it helps determine the level of weight changes. However, since the learning process tends to slow down when smaller learning rate values are employed for training, it is not a favoured choice. Employing larger learning rates values for training could lead to network oscillation in the weight space. One approach to enhance the gradient descent method is by introducing an additional momentum parameter (mc) that facilitates larger learning rates leading to faster convergence while decreasing the oscillation tendency (Rumelhart et al., 1986). The momentum term is introduced so that the next weight changes are similarly aligned to the same direction as the previous one, which allows minimising the oscillation impact of larger learning rates. Although there are certain systematic approaches to simultaneously choose the learning rate and momentum, the best values pertaining to these learning parameters are normally selected based on experimentation. Since any value falling between 0 and 1 can be accounted by the learning rate and the momentum, it becomes almost impossible to perform an exhaustive search to detect the best combinations pertaining to these training parameters. In this research paper, we evaluated different momentum and learning rates pertaining to both networks; in real practice, 0.9 and 0.95 were selected as momentum and optimum learning rate pertaining to SS, AN and pH models, respectively.

7.2 Optimisations of the Neurons Number

The number of neurons in the hidden layer is the key characteristic pertaining to AI technique. The network fails to model the complex data that could lead to poor fitting if the number of neurons employed is insufficient. On the flip side, the training time could become unreasonably long as well as the network may also over fit the data if there are too many neurons employed. In this paper, to investigate the best performance, various MLP-ANN architectures were employed. In fact, a formal and/or mathematical approach does not exist, which allows determination of appropriate ‘optimal set’ pertaining to neural network’s key parameters. Thus, the trial-and-error method was selected to perform this task. Randomisation of the hidden layer’s neurons was done from N=1 to 20 neurons. In the hidden layer, the best numbers of nodes are those that provide the lowest error (Lek et al., 1996). Based on two performance indices, determination of the optimum number of neurons was done. The root-mean-square error (RMSE) value pertaining to the prediction error is the first index, while the value of the maximum error is the second index. To get both indices, the ANN model was evaluated by considering the WQP data between 1998 and 2007. When building such a predicting model that employs the neural network, the model could do well during the training period and could give a higher level of error when assessment was done during either the testing or validation period. Based on this study, these performance indices were employed to ensure that the put forward model would offer consistent accuracy levels during all periods. As the performance indicator for the put forward model, the key benefit of using these two statistical indices is to ensure that the highest error falls within the acceptable error range for the forecasting model when the performance is being evaluated. This is done when RMSE is employed and making sure that the summation of the error distribution is not high in the validation period. Consequently, employing both indices ensures consistent level of errors and offers high potential to maintain the same error level while evaluating the model for unseen data during the testing period.

When the number of hidden neurons to the network is varied, it has a clear impact to a considerable degree on the prediction performance. It clearly demonstrates that there is a rise in prediction performance with increase in the number of hidden neurons (from 1 to 18), along with subsequent decrease in RMSE and maximum error pertaining to all parameters. However, a drop in prediction performance occurred when hidden neurons were added further (19 to 20) to the network. For instance, it can be seen that the best combination pertaining to the put forward statistical indices to examine the predicting model for the pH was when 18 neurons with RMSE 0.15 were associated with the ANN architecture and a maximum error as 3.22%. The best combination pertaining to the put forward statistical indices to examine the predicting model for the SS was when 17 neurons with RMSE 0.30 were associated with the ANN architecture and a maximum error of 3.46%. Table 5 lists out the optimal numbers of neurons pertaining to the remaining parameters.

Table 5.

7.3 WATER QUALITY PREDICTION MODEL OF MLP-ANN

The MLP-ANN model for the estimation of the 6 parameters of water quality (as the output), which are SS, AN and pH, was evaluated in this section. Figure 6 depicts the measured and estimated parameters of water quality for the most excellent network, which provided the most precise estimation. On the whole, the predictive capability of this model was fairly good for each of the parameters of the water quality in the training duration, though less accurate when the validation and testing stages were carried out. The findings showed that it was challenging to develop a consistent model using the MLP-ANN models due to high variations and intrinsic non-linear correlation among the parameters of the water quality because of the probabilistic nature and chemical procedure. Additionally, the MLP-ANN models encountered delayed convergence during the training because of the necessity of comparatively a huge amount of hidden neurons. Also, several researchers observed that these models failed to acquire values lying outside the scope of values included in the calibration data of MLP-ANN (boundary values) (Campolo et al., 1999; DAWSON and WILBY, 1998; Hsu et al., 1995; Karunanithi et al., 1994; MINNS and HALL, 1996). This constraint, arising chiefly due to the application of a logistic function to translate the output of the model, makes these models inappropriate for several applications.

Alternatively, the RBF-ANN (Radial Basis Function Network) is commonly employed for strict interpolation issues in space with multiple dimensions, which has equivalent abilities as the MLP-ANN in solving problems related to function estimations (Park and Sandberg, 1993). There are chiefly 2 benefits of the RBF-ANN: (a) network training in shorter duration in comparison to MLP-ANN , and (b) best solution estimation without managing the local minimums. In addition, RBF-ANN works as a local network in contrast to the feed-forward networks which are global mapping networks. Also, RBF-ANN employs one processing units set, and every unit is most accessible to a local area of the input region.

Due to this, RBFNs are employed more recently as a substitute NN model in function estimation applications and prediction of time series (Sheta and De Jong, 2001; Yu et al., 2008). Thus, the following section describes the attempt to get familiar with RBF-ANN suitability to be used as a model for predicting the parameters of water quality.

Figure 6.

7.4 SENSITIVITY ANALYSIS

To assess the input variables, impact on the model, 3 assessment methods were used. First method was based on dividing the NN connection weights so as to establish the relative significance of every input variable in the network (Stern and Garson, 1999). In this research, the recommended network comprises 12 environmental variables. Presuming the connection weights from the input nodes to the hidden nodes exhibit the relative predictive significance of the independent parameter, the significance of every input parameter can be articulated as follows:

$$Ij = \frac{\sum_{m=1}^{m=Nh}\left(\left(\left|W_{jm}^{ih}\right| / \sum_{k=1}^{Ni}\left|W_{km}^{ih}\right|\right)\times\left|W_{mn}^{ho}\right|\right)}{\sum_{k=1}^{k=Ni}\left\{\sum_{m=1}^{m=Nh}\left(\left(\left|W_{jm}^{ih}\right| / \sum_{k=1}^{Ni}\left|W_{km}^{ih}\right|\right)\times\left|W_{mn}^{ho}\right|\right)\right\}} \quad (11)$$

Where Ij represents the relative significance of jth input variable on the output variable, Ni and Nh denote the quantities of input and hidden neurons, correspondingly, and W represents the connection weight. Also, the superscripts ‘i’, ‘h’ and ‘o’ signify the input, hidden and output levels, correspondingly, while the subscripts ‘k’, ‘m’ and ‘n’ signify the input, hidden and output neurons, correspondingly. The first method of evaluation was to assess the relative significance of every input variable as calculated by Eq. (11) and illustrated in Figure 7. The relative significance demonstrates the importance of a variable in comparison to the other variables belonging to the model. Even though the network did not essentially signify physical sense using weights, it indicates that all the variables had intense effects on the estimation of all output variables, in which the estimator contribution varied from 5 to 14%. Apparently, the most useful inputs were considered to be those that involved oxygen containing nitrate (NO3) and phosphate (PO4). Conversely, pH and Temp were discovered to be the least useful parameters. Additionally, MG proved to be providing the greatest contribution for the recommended model for AN. For pH, it was apparent that the most useful input was Temp.

Figure 7. Relative importance of each input parameter.

7.5 WATER QUALITY PREDICTION MODEL OF ANFIS

As a matter of fact, among the difficulties in ANFIS-based modelling is establishing its variables for optimal learning (i.e. the membership function number and step size's initial value) before training, in a way that the optimal training is achieved. Two techniques have been proposed by several researchers for establishing these variables in ANFIS: optimisation techniques (Hassanain et al., 2004) and the trial-and-error approach (Kim et al., 2002). While determining the variables for optimal learning could be ensured by the optimisation algorithms (i.e. derivative based or derivative free optimisation), this alternative has a downside of being computationally costly. Conversely, the trial-and-error technique has been confirmed to be effective in case the target root mean square error can be realised. This technique is also advantageous as it yields a knowledge rule-base having a lower possibility of surpassing the data set of training in comparison to the optimisation technique. Thus, this research did not include the optimisation technique and established the variables for optimal learning of ANFIS through the trial-and-error technique.

For every parameter related to the water quality, this study employed the architectures proposed in the preceding section, in which 12 inputs were utilised to estimate the WQIP. It is noteworthy that there is no systematic technique to establish the optimal quantity of MFs. The optimal quantity of MFs is generally established inductively and validated empirically. Thus, the quantity of MFs was selected using the trial-and-error method. Meanwhile, it is to be observed that this study had tested 4 kinds of membership functions: (a) triangular, (b) gaussian, (c) trapezoidal, and (d) bell-shaped, to compose the fuzzy numbers. Following several trials, the outcome revealed a distributed membership function having bell-shaped nature in comparison to others which had acquired the minimal relative error. Table 6 demonstrates the kinds and quantity of MFs that were implemented in this study to develop the modules.

Table 6.

For demonstrating the performance of the suggested ANFIS model, an evaluation of predicted against observed parameters of water quality during training, validation and experimentation phases is displayed in the Figure 8. It is apparent that the suggested ANFIS model procedure provided the estimated variables that mimicked the dynamics (pattern) in the noted values besides those boundary values measured during this time.

Figure 8.

7.6 WATER QUALITY PREDICTION MODEL OF WDT-ANFIS

The above findings were obtained with the general assumption that the mined data must be precise and reliable. Nevertheless, the data acquired from the study, test, and simulation procedures may be corrupted by noise because of objective and/or subjective errors (Li and Shue, 2004). For instance, the errors arising in the experiment may be caused by measuring, recording, reading, or external scenarios; the errors from simulation might cover uncertainties of the model and parameters, as well as computational errors. As these noisy signals possibly distort the data mining outcomes, it is necessary to eliminate them (i.e. signal de-noising process) before the use of any initial data. Thus, an augmented WDT-ANFIS based on historical information for WQPP will be presented.

Training and cross-validation processes of the model of WDT-ANFIS were carried out to reduce the Root Mean Square Error among the output as well as predicted responses. The WDT-ANFIS model outperformed the ANFIS model and provided improvement in estimation accuracy of all the variables, while the ANFIS model performed inefficiently. As the noise intensity increased, it was obvious that WQP possibly had more accurate estimation values due to de-noising of data. This suggests the WDT superiority in data cleaning. Despite the occurrence of errors during stages of training, validation and experimentation, which were regarded as considerably high in comparison to the training and cross-validation stages, it had obtained a high precision for all variables. The findings displayed in Figure 9 demonstrate that the WDT-ANFIS model could be regarded as a suitable technique for modelling for estimation like WQP.

Figure 9.

7.7 COMPARATIVE ANALYSIS

The models introduced in prior discussion were all compared for the purpose of providing precise predictions for each water-quality parameter at Johor River. Similar findings were achieved in determining models for predicting suspended solids concentrations (SS), wherein WDT-ANFIS forecast SS with comparatively less accuracy, in which errors for most records were below 10%. Peak SS values were more closely approximated using WDT-ANFIS in comparison to that attained using other techniques, as depicted in Figure 10. The numbers of inaccurate SS forecasts decreased meaningfully using WDT-ANFIS. The use of physics-based distributed processing in complex computer software is frequently problematic, owing to the usage of idealised sedimentation components or the requirement of large volumes of detailed temporal and spatial data on the environment which is not always available (Cigizoglu, 2004). It should be noted that AI approaches to determining suspended-sediment data estimations remain sparse in the relevant literature (Abrahart and White, 2001).

The success attained in modelling dynamic systems implies that this strategy may well provide an efficient and productive means for simulating complex suspended-sediment processes in rivers, under conditions where precise knowledge of internal sub-processes is not necessary. Each proposed model in this study was constructed on the assumption that land cover/use would remain unchanged during this research. However, land cover/use remains an important factor in the production and transport of sediments, along with other factors. More precise predictions of suspended sediments may be attained by including variables that represent land cover/use status into the scheme. We are planning such analytical studies soon enough. In conclusion, this research establishes WDT as an appropriate method, along with classical ANFIS, for modelling suspended sediments in river environments. It is therefore worth considering the use of WDT-ANFIS approaches in such analysis, given the findings of studies regarding the physics embedded in ANFIS structures.

Figure 10.

With regards to pH, Figure 11 depicts comparisons between ANFIS and other models' performances, based on the test data set. In the figure, it is clear that ANFIS performance exceeds that of the two ANN methods. Furthermore, the effort reveals the challenges in devising reliable schemes based on MLP-ANN RBF-ANN models, as a result of the high variances as well as the inherent non-linear associations among the water-quality parameters, as a result of the stochastic quality and chemical-based process. Furthermore, as depicted in Figure 10, the findings show that WDT-ANFIS-based modules outperform ANFIS and also have the ability to improve predictive accuracy for pH, albeit for MAE with comparatively lesser accuracy, whereby errors for most records were below 7%. Otherwise, inefficient executions were observed based on the ANFIS module, wherein most errors were above 15%. Clearly, given increases in noise intensities, WQP offers more precise predictions from data de-noised with WDT than data without such de-noising. This suggests the advantage of using WDT to clean the data.

It is fact that the training process for big data using any of AI models is both time-consuming and computation- and memory-intensive especially when several number of model' inputs variables is used. The computer specification that have been used to run models are Intel Processor Core i7 (12M Cache, up to 4.60 GHz) and Ram 16 Gb. It is fact that in our study the data used is not big data to be considered as problem to the computational memory. However, due to the fact that the number of the model' input variables is relatively big (twelve or thirteen based on the structure of scenario I and scenario II, respectively), the training process is slightly time-consuming to achieve the performance goal. Table 7 summarize the training time for each models in seconds where it is noticeable that the ANFIS and WDT-ANFIS models consuming more time than ANN models (MLP and RBF) but it is still minimal.

Figure 11.

Table 7

7.8 SCENARIOS

The comparatively low correlation among forecast and observed values during test phases was perhaps a result of the non-homogenous nature of water-quality parameters. Moreover, Ying et al. (Zhao et al., 2007) demonstrated that the selection of influential factors (namely, input parameters) has a critical role as these factors greatly affect forecasts. Clearly, the low correlations in this research can be attributed to the realisation that its input parameters had not included every relevant parameter. Furthermore, pollution levels at downstream stations were associated with discharge from upstream stations. To overcome this difficulty, the researchers applied another approach (i.e. Scenario 2), such that higher levels of accuracy could be attained. This strategy is associated with the prediction of each water-quality parameter, given the actual values measured at upstream stations as model inputs, as described by Eq. (12). For a most appropriate analysis, the researchers implemented an accuracy improvement (AI) index for the correlational coefficient statistical index, in order to determine the significance of Scenario 2 as against Scenario 1, described as follows:

$$AI(\%) = \left(\frac{CC_{Scen2} - CC_{Scen1}}{CC_{Scen2}}\right) * 100 \tag{12}$$

Wherein CCScen2 denotes the coefficient of correlation for Scenario 2, whereas CCScen1 denotes a similar statistical index for Scenario 1. From Table 8, it is clear that Scenario 2 is more satisfactory than Scenario 1, with meaningful improvements observed in every station, which ranged from 0.5% to 5%. Predictive accuracy was meaningfully enhanced after introducing Scenario 2 for every station. As in the case for pH, Scenario 2 showed more satisfactory performance than Scenario 2, with meaningful improvements observed in AI, which ranged from 3% in Station 2 to 5% in Station 3.

Conversely, less improvement was gained with AN, wherein AI was equal to 0.5 in Stations 1 and 3. Even though it is clear that Scenario 2 was less efficient with AN, accuracy does increase by 2% once it is applied to Station 3. Furthermore, the findings indicate that Scenario 2 not only showed improved accuracy for certain parameters, but this particular model had the ability to capture temporal patterns in water-quality parameters. This enabled the scheme to apply meaningful improvements to station scenarios.

Table 8.

7.9 MODEL VALIDATION

Models must be verified whenever resulting outputs and observed values are near enough to satisfy all validation criteria (Palani et al., 2008). To investigate the effectiveness of this proposed scheme, validation of the enhanced wavelet de-noising method using the Neuro-Fuzzy Inference System (WDT-ANFIS), in accordance with field measurements collected from 2009 to 2010, is therefore applied. The scatter plots among the forecast and observed values for all 5 selected parameters for water quality are depicted in Figure 12. Clearly, the majority of forecast water-quality parameters had closely approximated actual observations. As well, R^2 must be as near 1 as possible, with values that exceed 0.9 implying very satisfactory model execution, values from 0.6 to 0.9 implying fairly good execution, and values below 0.5 indicating unsatisfactory execution. Based on these criteria, the WDT-ANFIS model's ability to predict both pH and SS concentrations is very satisfactory (in that R^2 values are at least 0.9) for every station but for AN, wherein models showed merely decent performances (in that R^2 values were below 0.9) for Station 3. Based on these findings, WDT-ANFIS can be said to demonstrate good predictive performance. For predictions of water-quality parameters using AI, other researchers have advanced network modelling strategies that apply differing types of AI as well as input datasets. Moatar et al. (Moatar et al., 1999) applied solar radiation and discharge levels in predictions of pH, with an R^2 value equal to 0.86. For predictions of AN, WDT-ANFIS predictive performance in this research managed better in comparison (R^2 ranging from 0.88 to 0.96) with ANN predictive performance. Cigizoglu (Cigizoglu, 2004) utilised ANN models that were trained and then tested with daily flows, for predicting SS concentrations a day ahead, with R^2 values ranging from 0.75 to 0.81 (with upstream flows as inputs). A comparable prediction for SS was similarly claimed by Zhu et al. (Zhao et al., 2007). For predictions of SS, the WDT-ANFIS predictive performance in this research managed better in comparison (R^2 ranging from 0.91 to 0.95) to previous studies. The proposed scheme demonstrated efficiency in its predictions of the concentrations of water-quality parameters for the Johor River, which corresponds to the findings of other research. The findings also show that the proposed scheme is a useful alternative that offers a comparatively fast algorithm, featuring decent theoretical properties for predicting water-quality parameters, which could be extended to predictions of other water-quality parameters.

Figure 12.

8. CONCLUSION

The study proposes the use of enhanced Wavelet De-noising Techniques using Neuro-Fuzzy Inference Systems (WDT-ANFIS) according to historical water-quality parametric data. The effectiveness of each model was examined in order to predict key parameters that could be affected as a result of urbanisation surrounding rivers. This area of research accords with the available secondary data for each water-quality parameter of Johor River. The parameters comprise ammoniacal nitrogen (AN), suspended solid (SS), and pH. Dual scenarios were presented: the first (Scenario 1) was designed to confirm prediction models for water-quality parameters at each stations according to 12 input parameters, whereas the second (Scenario 2) is designed to confirm prediction models for water-quality parameters according to 12 input parameters, as well as the parametric values from prior upstream stations. In evaluating the impact of input parameters on this scheme, validation of enhanced Wavelet De-noising Techniques using Neuro-Fuzzy Inference Systems (WDT-ANFIS), in accordance with measurements taken from 2009 to 2010, was thereby employed. The findings showed the challenge of determining reliable schemes based on MLP-ANN models, from the high variances as well as inherent non-linear associations among the water-quality parameters that emerge as a result of the stochastic quality and chemical-based process. Furthermore, MLP-ANN was subject to slow convergence during training, as a result of the requirement for comparatively large numbers of hidden neurons. In the example of RBF-ANN, its predictive capability for water-quality parameters in training phases was decent, but showed less precision during validation and test phases. The findings indicated that ANFIS determined solutions faster than alternative MLP-ANN and RBF-ANN methods and is the most precise and reliable method for processing large volumes of non-linear as well as non-parametric data. Of note is the performance of the WDT-ANFIS scheme, which exceeded that of ANFIS and improved predictive accuracy for every quality parameter, in that this model achieves higher prediction accuracy overall. Generally, WDT-ANFIS can therefore be seen as having the best network architecture, since it outperformed ANFIS. The findings indicate that WDT-ANFIS not only offered a means to improve accuracy but it also features the ability to capture temporal patterns in water quality. This enables it to provide meaningful improvements in the generation of forecasts. Consequently, the ANFIS model appears more capable at capturing the more complex and dynamic processes that are hidden within the data for WQP, following enhancement with WDT. In comparisons between Scenarios 1 and 2, Scenario 2 achieved higher accuracy in terms of simulating the patterns and magnitudes for every water-quality parameter, at every station. The suggested WDT-ANFIS model in Scenario 2 gave predictions for water-quality parameters that ably mimicked patterns (dynamics) in recorded values, aside from extreme outliers observed within this period. Furthermore, validation of WDT-ANFIS, according to measurements collected from 2009 to 2010, demonstrated that WDT-ANFIS performed well in predicting both pH and SS concentrations (with R^2 values of at least 0.9) for every station but for AN, wherein models still showed decent performances (with R^2 values lower than 0.9) for Station 3. Since forecasts of water quality are readily influenced by external environments, the acquired model would at times generate findings that deviated much from the observed values. In general, the methodology of the proposed models development for water quality has proved its effectiveness. However, it should be highlighted that there are no structured methods today to identify which network structure that can best in predicting water quality parameters. Moreover, the optimal selection of the hyper parameters still requires to be achieved by augmenting the AI model with other advanced meta-heuristic optimization algorithms. Overall, this study integrates several analytical and modelling techniques that could become useful to institutions that are committed to river basin management within Malaysia. Furthermore, the approach utilised in this research could lay ground for better decision-making that assists policy makers in maintaining and improving river basin management.

Acknowledgments: The authors would like to appreciate the technical and financial support received from research grant coded J510050822 by Innovation & Research Management Center (iRMC), Universiti Tenaga Nasional (UNITEN) and from research grant coded UMRG RP025A-18SUS funded by the University of Malaya

Conflicts of Interest: The authors declare no conflict of interest.

References

Abrahart, R.J., White, S.M., 2001. Modelling sediment transfer in Malawi: comparing backpropagation neural network solutions against a multiple linear regression benchmark using small data sets. Phys. Chem. Earth, Part B Hydrol. Ocean. Atmos. 26, 19–24. https://doi.org/10.1016/s1464-1909(01)85008-5

Avci, E., 2007. An expert system based on Wavelet Neural Network-Adaptive Norm Entropy for scale invariant texture classification. Expert Syst. Appl. 32, 919–926. https://doi.org/10.1016/j.eswa.2006.01.025

Bell, W.R., Martin, D.E.K., 2004. Computation of asymmetric signal extraction filters and mean squared error for ARIMA component models. J. Time Ser. Anal. 25, 603–623. https://doi.org/10.1111/j.1467-9892.2004.01920.x

Bowden, G.J., Dandy, G.C., Maier, H.R., 2005. Input determination for neural network models in water resources applications. Part 1—background and methodology. J. Hydrol. 301, 75–92. https://doi.org/10.1016/j.jhydrol.2004.06.021

Campolo, M., Andreussi, P., Soldati, A., 1999. River flood forecasting with a neural network model. Water Resour. Res. 35, 1191–1197. https://doi.org/10.1029/1998wr900086

Chang, F.-J., Chang, Y.-T., 2006. Adaptive neuro-fuzzy inference system for prediction of water level in reservoir. Adv. Water Resour. 29, 1–10. https://doi.org/10.1016/j.advwatres.2005.04.015

Chang, F.-J., Chen, Y.-C., 2001. A counterpropagation fuzzy-neural network modeling approach to real time streamflow prediction. J. Hydrol. 245, 153–164. https://doi.org/10.1016/S0022-1694(01)00350-X

Chang, Y.-T., Chang, L.-C., Chang, F.-J., 2005. Intelligent control for modeling of real-time reservoir operation, part II: artificial neural network with operating rule curves. Hydrol. Process. 19, 1431–1444. https://doi.org/10.1002/hyp.5582

Cigizoglu, H.K., 2004. Estimation and forecasting of daily suspended sediment data by multi-layer perceptrons. Adv. Water Resour. 27, 185–195. https://doi.org/10.1016/j.advwatres.2003.10.003

DAWSON, C.W., WILBY, R., 1998. An artificial neural network approach to rainfall-runoff modelling. Hydrol. Sci. J. 43, 47–66. https://doi.org/10.1080/02626669809492102

DID, 2000. Urban Stormwater Management Manual for Malaysia.

DOE, 2007. Malaysia Environmental Quality Report 2007. Malaysia Environ. Qual. Rep. 1–86. https://doi.org/10.1007/s13398-014-0173-7.2

Dogan, E., Sengorur, B., Koklu, R., 2009. Modeling biological oxygen demand of the Melen River in Turkey using an artificial neural network technique. J. Environ. Manage. 90, 1229–1235. https://doi.org/10.1016/j.jenvman.2008.06.004

Dohan, K., Whitfield, P.H., 1997. Identification and characterization of water quality transients using wavelet analysis. I. Wavelet analysis methodology. Water Sci. Technol. 36, 325–335. https://doi.org/10.2166/wst.1997.0229

Firat, M., Güngör, M., 2007. River flow estimation using adaptive neuro fuzzy inference system. Math. Comput. Simul. 75, 87–96.

Hsu, K., Gupta, H.V., Sorooshian, S., 1995. Artificial Neural Network Modeling of the Rainfall-Runoff Process. Water Resour. Res. 31, 2517–2530.

Hull, V., Parrella, L., Falcucci, M., 2008. Modelling dissolved oxygen dynamics in coastal lagoons. Ecol. Modell. 211, 468–480. https://doi.org/10.1016/j.ecolmodel.2007.09.023

Ibrahim, R., 2001. River Water quality Status In Malaysia, in: National Conference On Sustainable River Basin Management In Malaysia.

Jang, J.-S.R., 1993. ANFIS: adaptive-network-based fuzzy inference system. Syst. Man Cybern. IEEE Trans. 23, 665–685.

Karunanithi, N., Grenney, W.J., Whitley, D., Bovee, K., 1994. Neural Networks for River Flow Prediction. J. Comput. Civ. Eng. 8, 201–220. https://doi.org/10.1061/(asce)0887-3801(1994)8:2(201)

Khadse, G.K., Patni, P.M., Kelkar, P.S., Devotta, S., 2007. Qualitative evaluation of Kanhan river and its tributaries flowing over central Indian plateau. Environ. Monit. Assess. 147, 83–92. https://doi.org/10.1007/s10661-007-0100-x

Kim, B., Park, J.H., Kim, B.-S., 2002. Fuzzy logic model of Langmuir probe discharge data. Comput. Chem. 26, 573–581. https://doi.org/10.1016/s0097-8485(02)00021-9

Kişi, Ö., 2006. Daily pan evaporation modelling using a neuro-fuzzy computing technique. J. Hydrol. 329, 636–646. https://doi.org/10.1016/j.jhydrol.2006.03.015

Koklu, R., 2006. Dissolved oxygen estimation using artificial neural network for water quality control, Fresenius Environmental Bulletin.

Kuo, J.-T., Hsieh, M.-H., Lung, W.-S., She, N., 2007. Using artificial neural network for reservoir eutrophication prediction. Ecol. Modell. 200, 171–177.

https://doi.org/10.1016/j.ecolmodel.2006.06.018

Lek, S., Delacoste, M., Baran, P., Dimopoulos, I., Lauga, J., Aulagnier, S., 1996. Application of neural networks to modelling nonlinear relationships in ecology. Ecol. Modell. 90, 39–52. https://doi.org/10.1016/0304-3800(95)00142-5

Li, S.-T., Shue, L.-Y., 2004. Data mining to aid policy making in air pollution management. Expert Syst. Appl. 27, 331–340. https://doi.org/10.1016/j.eswa.2004.05.015

MINNS, A.W., HALL, M.J., 1996. Artificial neural networks as rainfall-runoff models. Hydrol. Sci. J. 41, 399–417. https://doi.org/10.1080/02626669609491511

Moatar, F., Fessant, F., Poirel, A., 1999. pH modelling by neural networks. Application of control and validation data series in the Middle Loire river. Ecol. Modell. 120, 141–156. https://doi.org/10.1016/s0304-3800(99)00098-8

Muttil, N., Chau, K.W., 2006. Neural network and genetic programming for modelling coastal algal blooms. Int. J. Environ. Pollut. 28, 223. https://doi.org/10.1504/ijep.2006.011208

Palani, S., Liong, S.-Y., Tkalich, P., 2008. An ANN application for water quality forecasting. Mar. Pollut. Bull. 56, 1586–1597. https://doi.org/10.1016/j.marpolbul.2008.05.021

Park, J., Sandberg, I.W., 1993. Approximation and Radial-Basis-Function Networks. Neural Comput. 5, 305–316. https://doi.org/10.1162/neco.1993.5.2.305

Rumelhart, D.E., Hinton, G.E., Williams, R.J., 1986. Learning representations by back-propagating errors. Nature 323, 533–536. https://doi.org/10.1038/323533a0

Sheta, A.F., De Jong, K., 2001. Time-series forecasting using GA-tuned radial basis functions. Inf. Sci. (Ny). 133, 221–228. https://doi.org/10.1016/s0020-0255(01)00086-x

Singh, K.P., Basant, A., Malik, A., Jain, G., 2009. Artificial neural network modeling of the river water quality—A case study. Ecol. Modell. 220, 888–895. https://doi.org/10.1016/j.ecolmodel.2009.01.004

Soyupak, S., Karaer, F., Gürbüz, H., Kivrak, E., Sentürk, E., Yazici, A., 2003. A neural network-based approach for calculating dissolved oxygen profiles in reservoirs. Neural Comput. Appl. 12, 166–172. https://doi.org/10.1007/s00521-003-0378-8

Stern, C., Garson, G.D., 1999. Neural Networks. An Introductory Guide for Social Scientists. Contemp. Sociol. 28, 753. https://doi.org/10.2307/2655607

Sugeno, M., Kang, G.T., 1988. Structure identification of fuzzy model. Fuzzy Sets Syst. 28, 15–33. https://doi.org/10.1016/0165-0114(88)90113-3

Tirtom, H., Engin, M., Engin, E.Z., 2008. Enhancement of time-frequency properties of ECG for detecting micropotentials by wavelet transform based method. Expert Syst. Appl. 34, 746–753. https://doi.org/10.1016/j.eswa.2006.10.009

Wavelet Toolbox - MATLAB [WWW Document], n.d.

Yu, L., Lai, K.K., Wang, S., 2008. Multistage RBF neural network ensemble learning for exchange rates forecasting. Neurocomputing 71, 3295–3302. https://doi.org/10.1016/j.neucom.2008.04.029

Zaqoot, H.A., Ansari, A.K., Unar, M.A., Khan, S.H., 2009. Prediction of dissolved oxygen in the Mediterranean Sea along Gaza, Palestine – an artificial neural network approach. Water Sci. Technol. 60, 3051–3059. https://doi.org/10.2166/wst.2009.730

Zhao, Y., Nan, J., Cui, F., Guo, L., 2007. Water quality forecast through application of BP neural network at Yuqiao reservoir. J. Zhejiang Univ. A 8, 1482–1487. https://doi.org/10.1631/jzus.2007.a1482

Figures

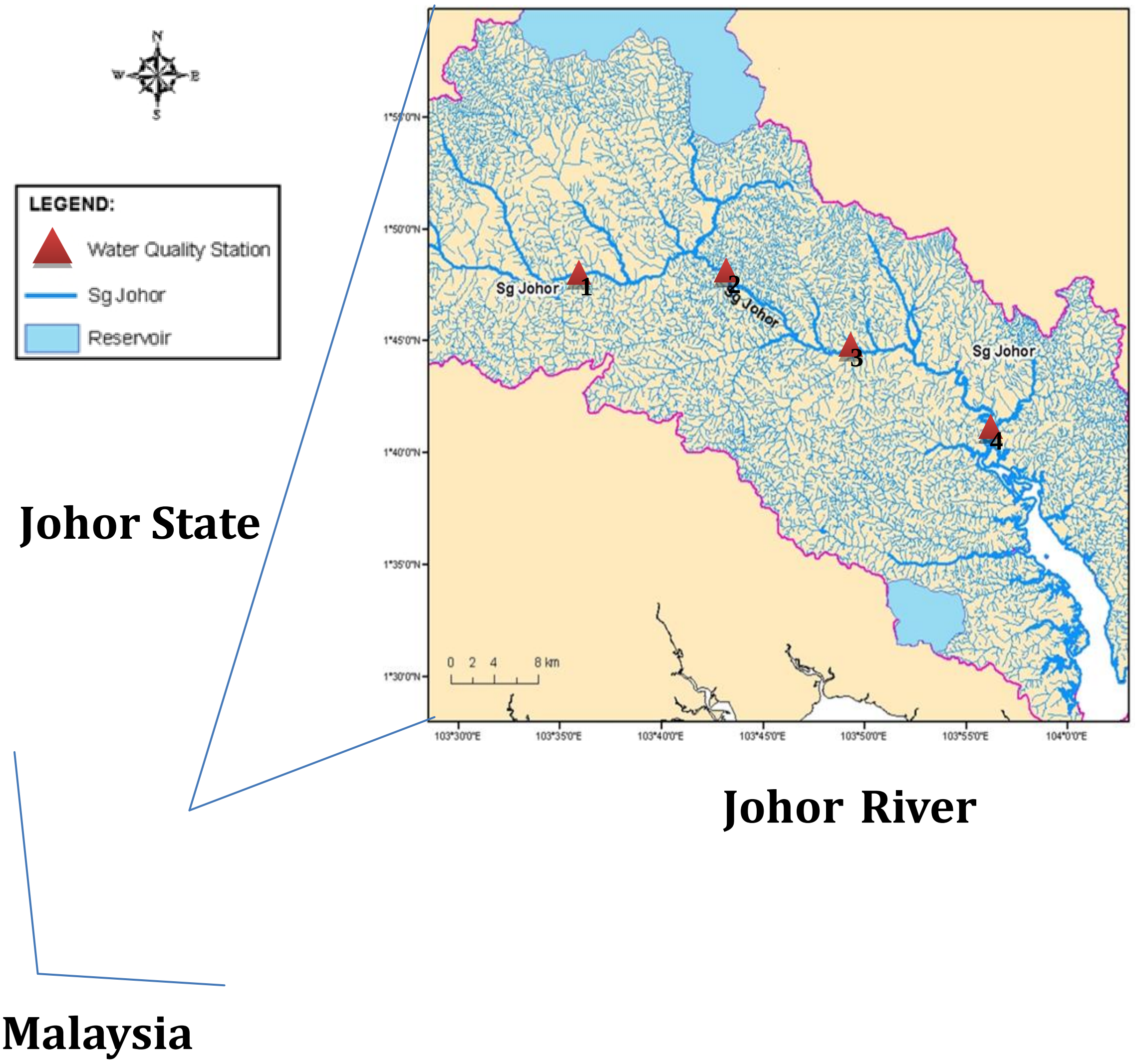


Figure 1. A map showing the geographical setting of the survey area with four field monitoring stations on the main stream

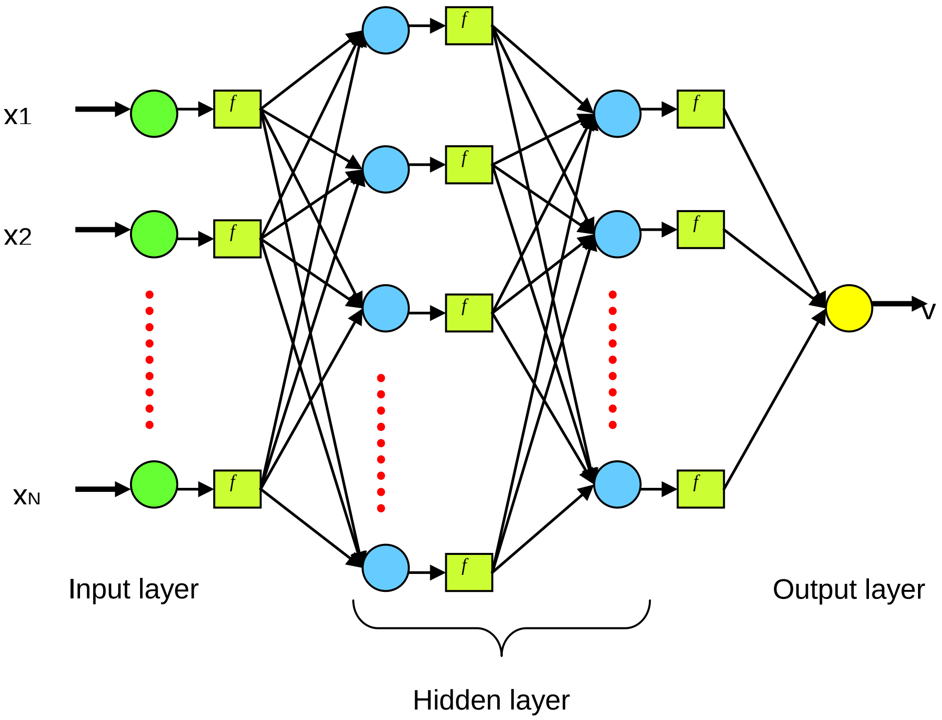


Figure 2. A multi-layer perceptron neural network architecture.

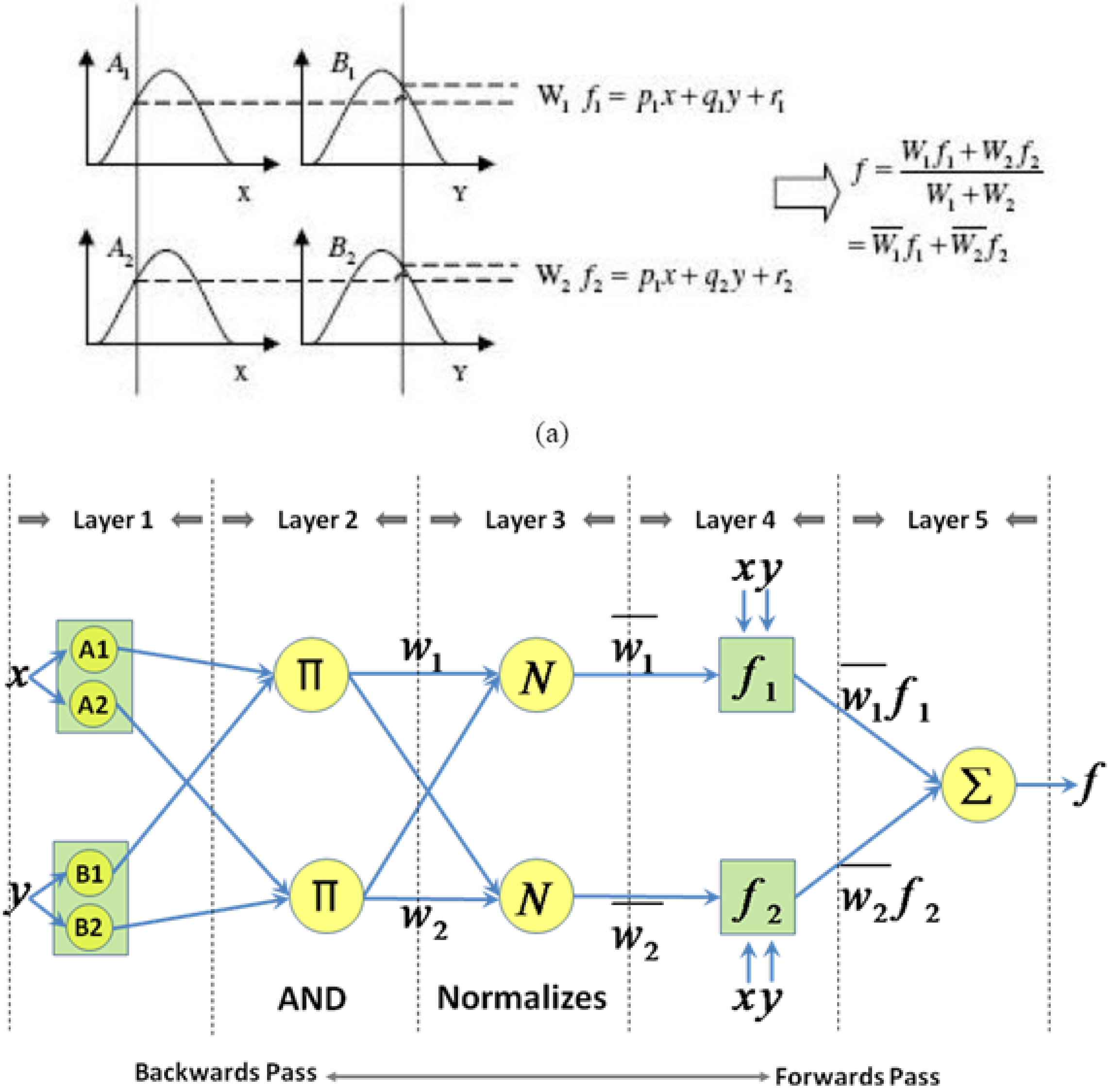


Figure 3. (a) A two-input first-order Sugeno fuzzy model with two rules; (b) An equivalent ANFIS structure.

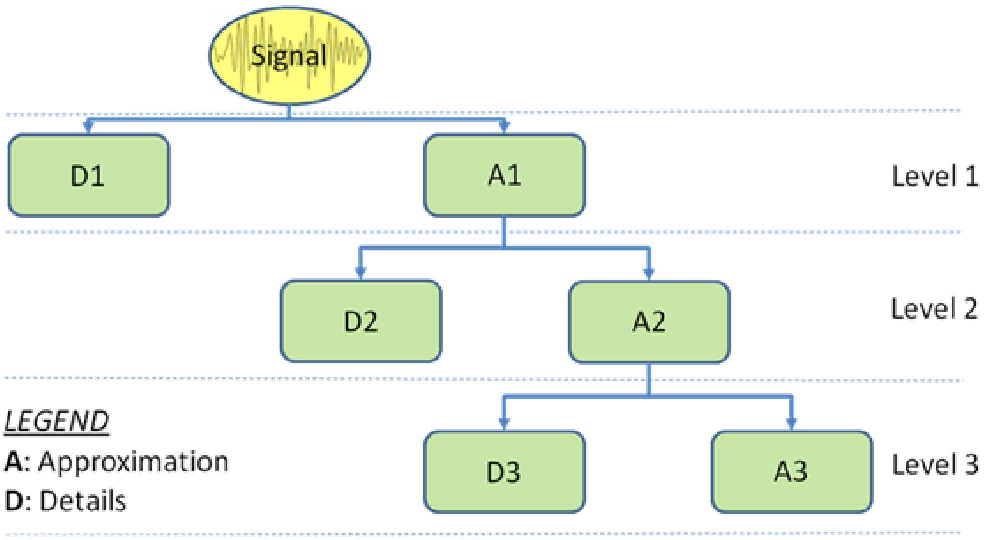


Figure 4. A schematic representation of the pyramid structure representing the WMRA.

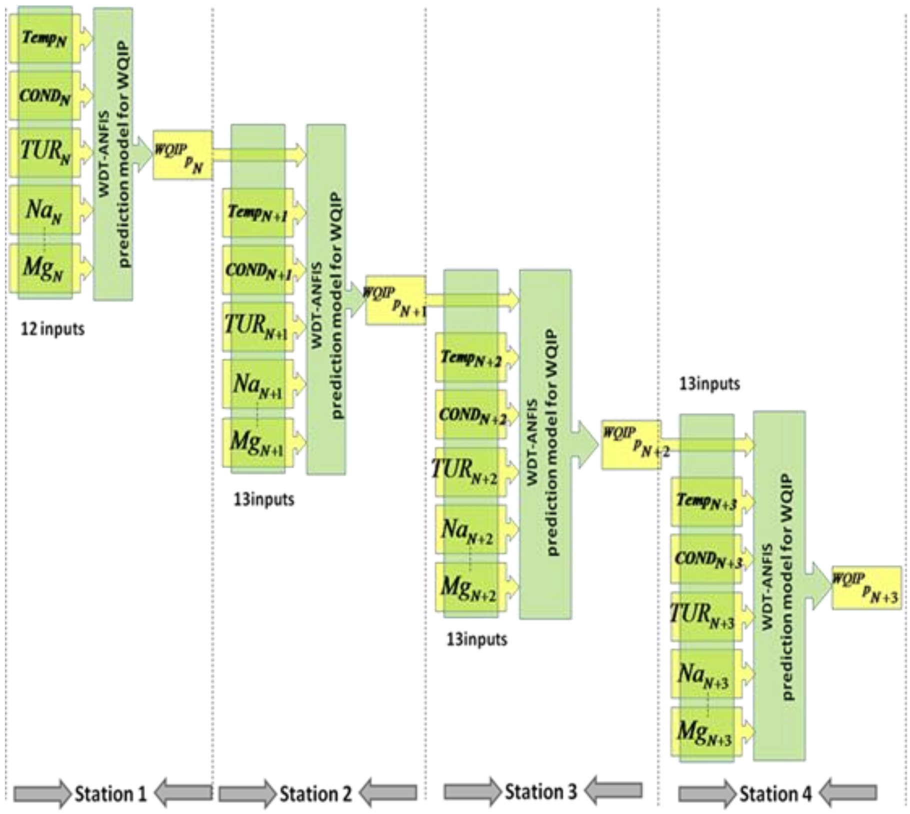


Figure 5. Schematic representation of the proposed networks for Scenario 2.

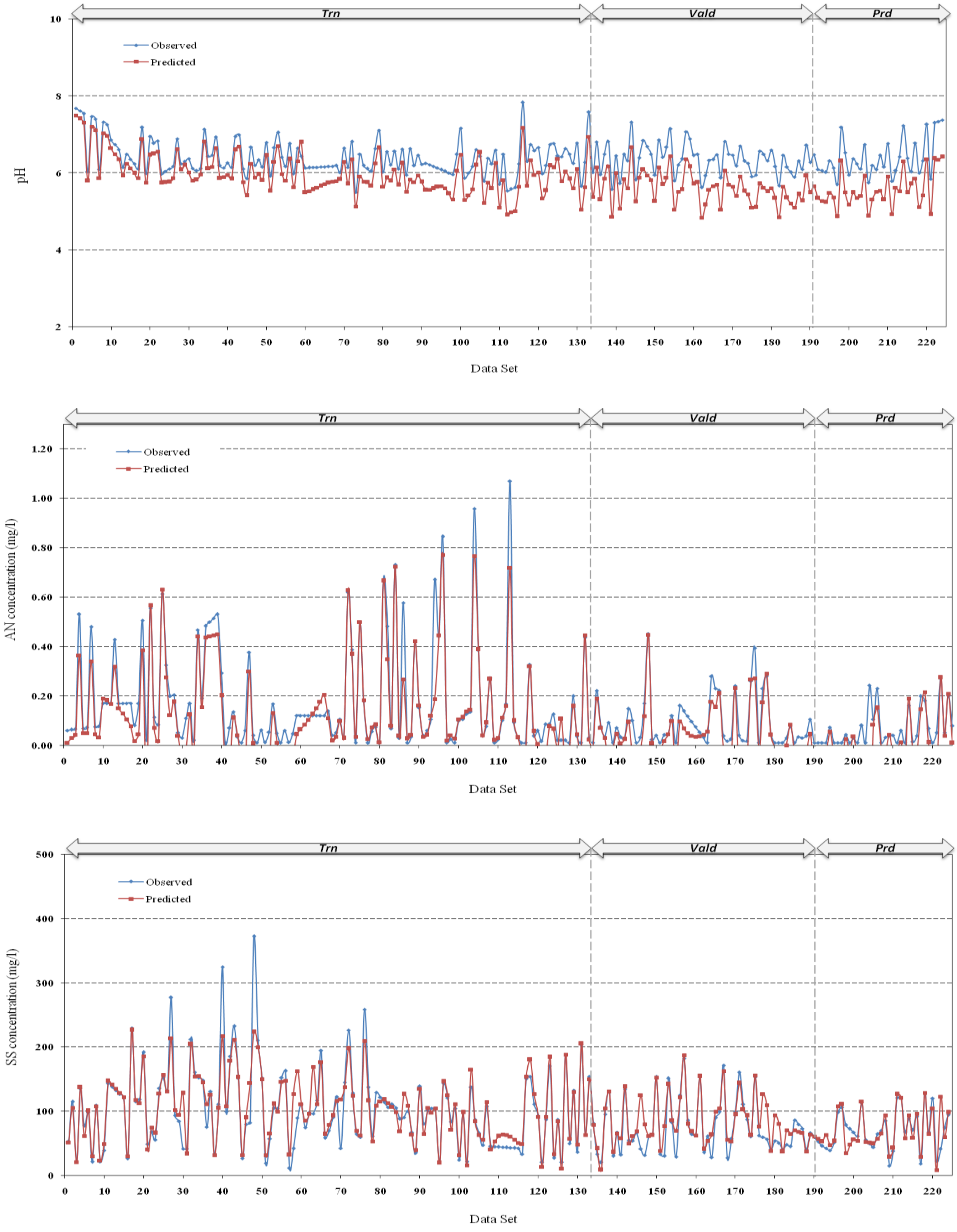
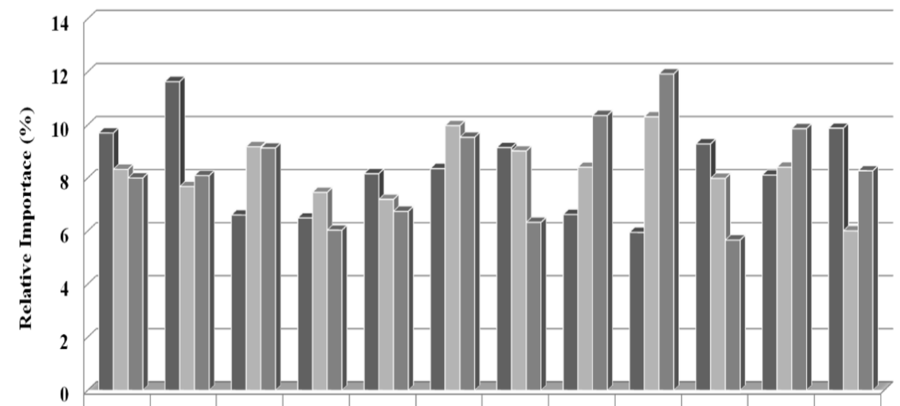


Figure 6. Performance of the MLP-ANN model: A comparison between the predicted and observed values.



	COND	TEMP	NO3	SAL	TURB	Cl	PO4	K	MG	NA	FE	E-coli
pH	9.6992	11.638	6.6094	6.4990	8.1682	8.3608	9.1518	6.6285	5.9566	9.2899	8.1155	9.8817
SS	8.3322	7.6819	9.1824	7.4634	7.1992	9.9799	9.0220	8.4105	10.311	7.9916	8.4116	6.0133
AN	8.0057	8.0942	9.1360	6.0392	6.7516	9.5393	6.3340	10.361	11.928	5.667	9.8666	8.2763

Figure 7. Relative importance of each input parameter.

Figure 8. Performance of the ANFIS model: A comparison between the predicted and observed values.

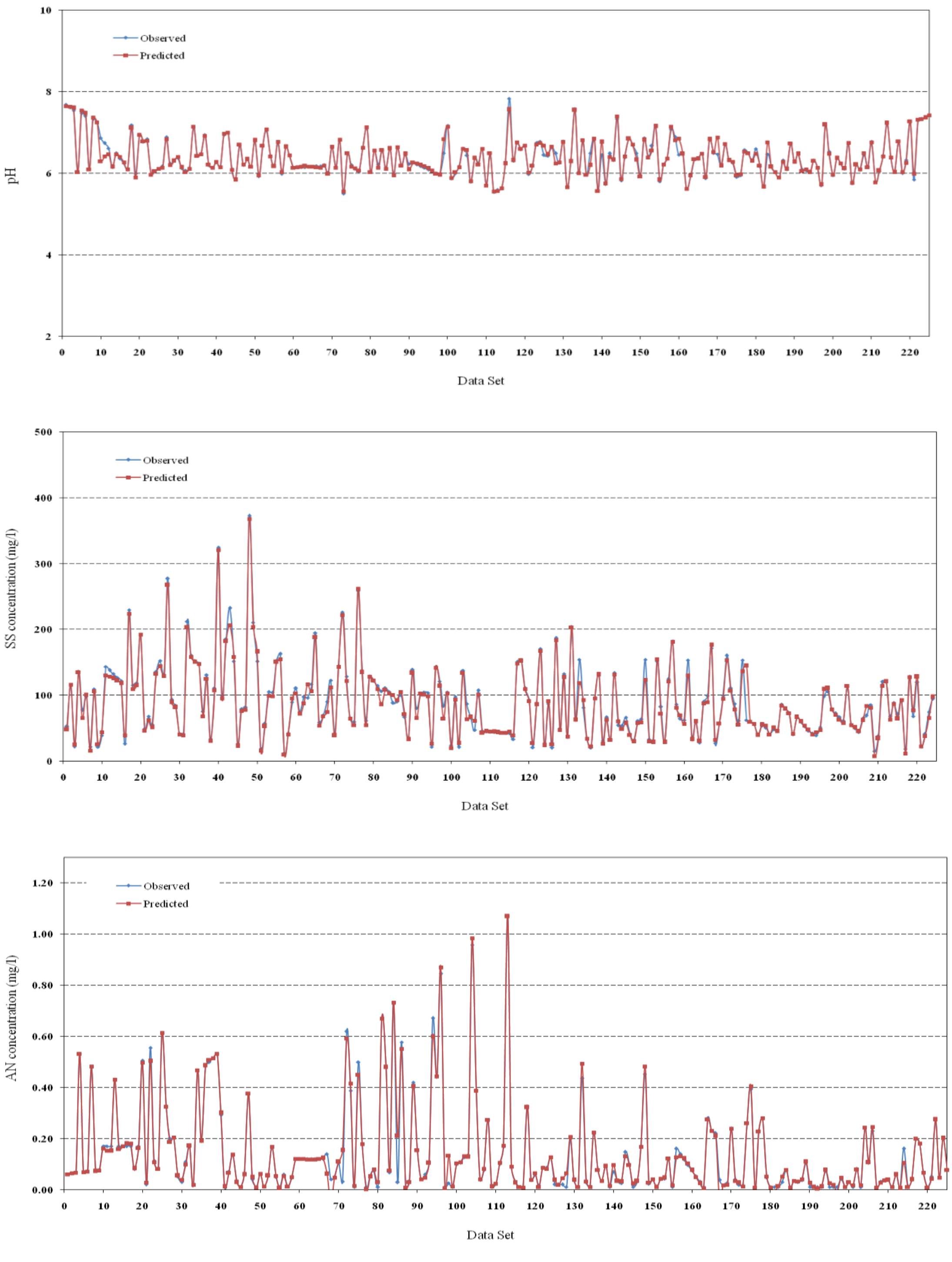


Figure 9. Performance of the WDT-ANFIS model: A comparison between the predicted and observed values.

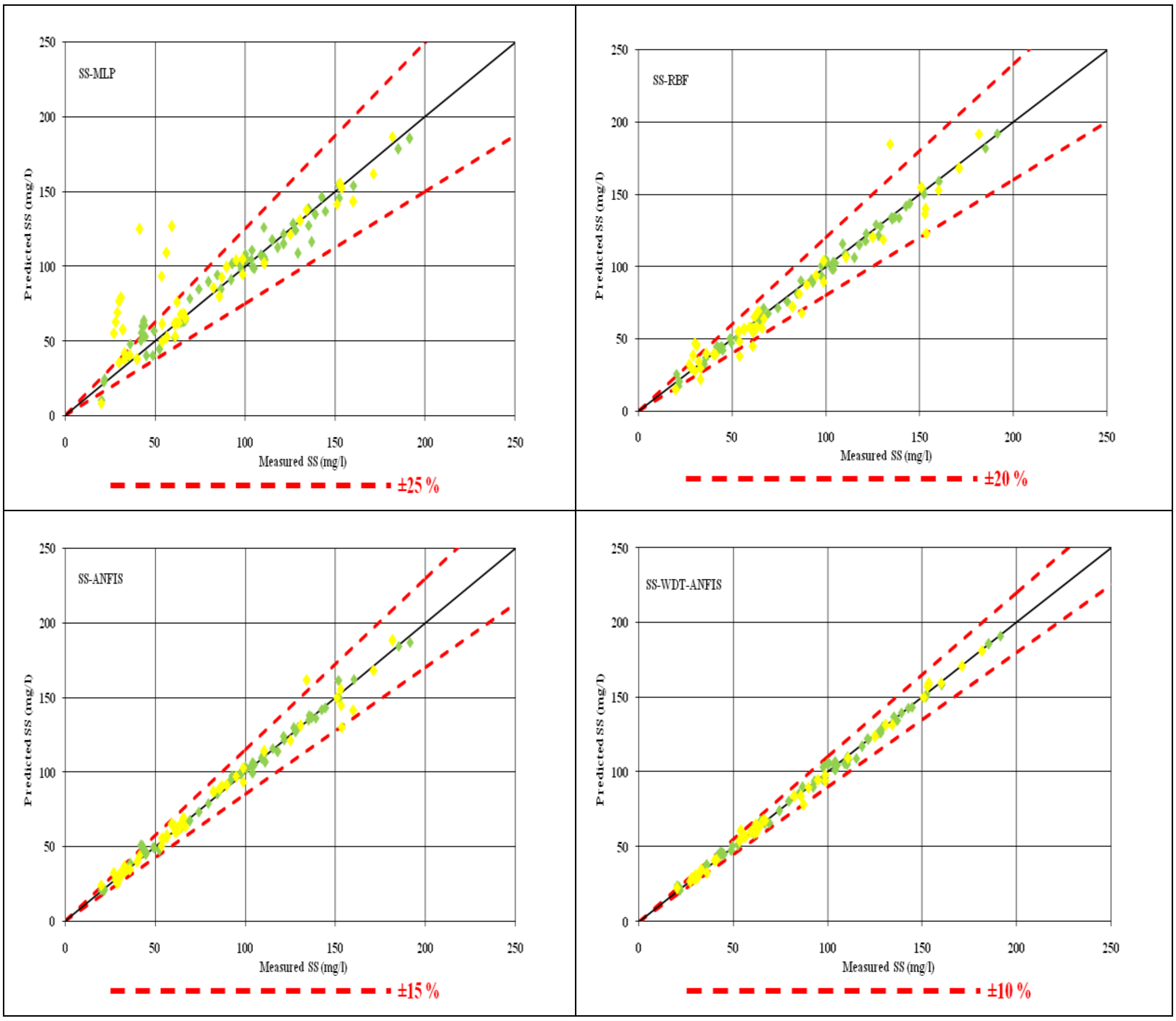


Figure 10. Comparison between the predicted SS versus the observed SS utilizing different techniques.

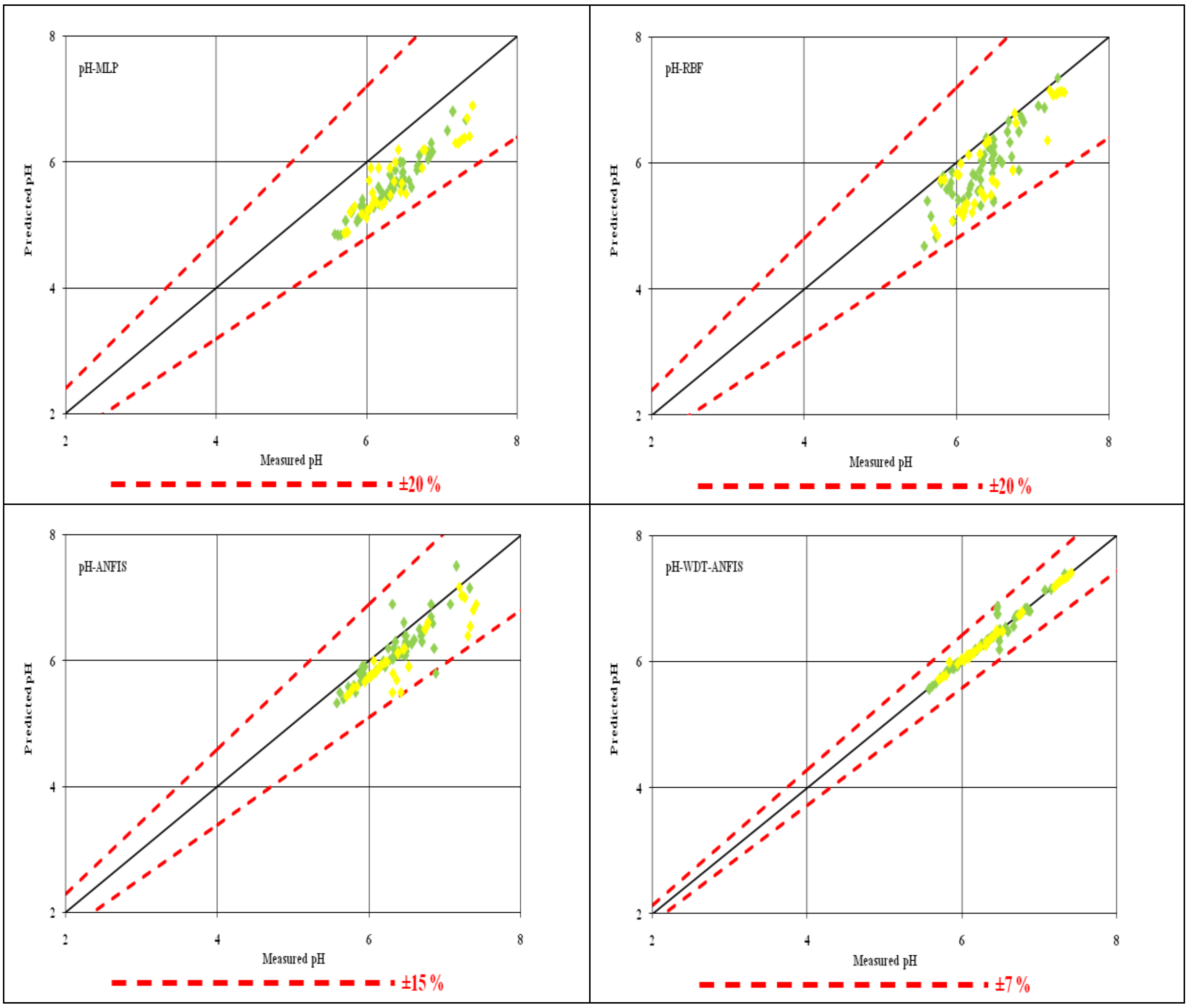


Figure 11. Comparison between the predicted pH versus the observed pH utilising different techniques.

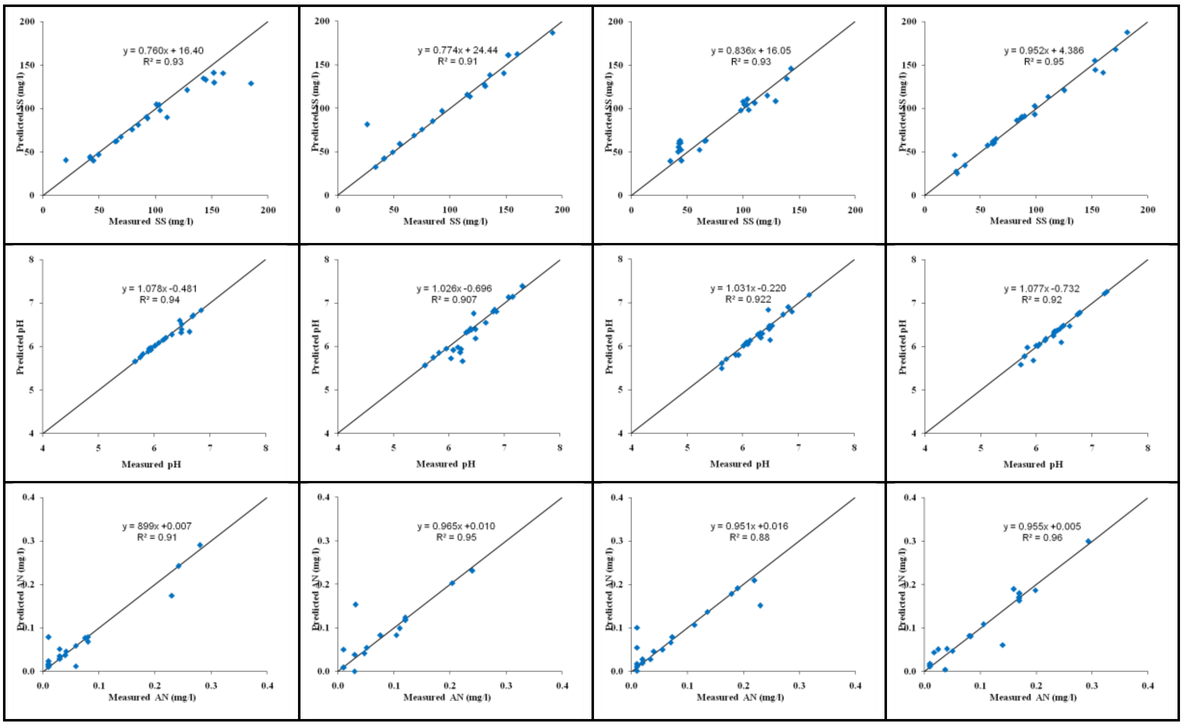


Figure 12. WDT-ANFIS model verification for each water quality parameter at each station.

Tables

Table 1. Input parameters used in previous studies for the ANN model.

Author(s) and year	Input variable	Location(s)
Rabia (Koklu, 2006)	BOD, Temp, Water discharge, NO2-N, NO3-N	N/A
Kuo *et al*. (Kuo et al., 2007)	pH, Chl-a, NH4N, No3N, temp, month	Te-Chi Reservoir, Taiwan
Ying *et al*. (Zhao et al., 2007)	Turbidity, Temp, pH, Hardness, Alkalinity, Chloride, NH_4-N, NO_2-N	Yuqiao reservoir, China
Palani *et al*. (Palani et al., 2008)	DO, Chl-a, temp	Singapore coastal, Singapore
Zaqoot *et al*. (Zaqoot et al., 2009)	Conductivity, Turbidity, Temp, PH, Wind speed	Mediterranean Sea along Gaza, Palestine
Singh *et al*. (Singh et al., 2009)	pH, TS, T-AlK, T-Hard, CL, PO4, K, Na, NH4N, No_3N, COD	Gomti, India

Table 2. Basic statistical analysis for input parameters.

	Unit	**Mean**	**Minimum**	**Maximum**	**SD**	**CV**
			SN01			
TEMP	o C	27.03	24.08	30.33	0.83	3.08
COND	µS	55.42	32.00	92.00	13.82	24.93
SAL	ppt	0.64	0.01	2.93	0.36	56.00
TUR	NTU	0.03	0.01	0.20	0.05	152.38
NO3	mg/l	163.50	15.50	775.00	130.61	79.88
CL	mg/l	5.27	1.00	18.00	2.49	47.16
PO4	mg/l	0.04	0.01	1.08	0.12	283.32
FE	mg/l	4.61	1.00	10.30	1.74	37.63
K	mg/l	0.87	0.10	2.40	0.44	50.59
MG	mg/l	3.13	1.22	11.54	1.42	45.18
NA	mg/l	0.87	0.08	2.32	0.44	51.20
E-COLI	cfu/100ml	3844.98	40.00	48000.00	6377.64	165.87
			SN02			
TEMP	o C	27.16	24.08	29.82	1.11	4.10
COND	µS	62.64	28.00	300.00	38.78	61.91
SAL	ppt	0.02	0.01	0.07	0.01	54.16

TUR	NTU	127.79	30.70	370.00	77.64	60.76
NO3	mg/l	0.73	0.12	5.55	0.69	93.53
CL	mg/l	5.66	1.00	24.00	3.28	57.89
PO4	mg/l	0.07	0.01	0.66	0.12	159.91
FE	mg/l	0.82	0.09	2.02	0.48	58.85
K	mg/l	4.63	0.90	7.80	1.56	33.76
MG	mg/l	0.80	0.10	1.40	0.33	40.69
NA	mg/l	3.27	1.40	26.70	3.33	101.77
E-COLI	cfu/100ml	2564.82	20.00	22000.00	3802.25	148.25
SN03						
TEMP	o C	26.14	23	31.93	1.38	5.07
COND	µS	54.16	26.07	373.00	45.62	84.24
SAL	ppt	9.56	0.01	61.00	20.43	213.64
TUR	NTU	113.33	0.01	820.00	139.73	123.29
NO3	mg/l	11.55	0.00	133.00	27.26	236.03
CL	mg/l	5.43	0.06	20.00	2.78	51.13
PO4	mg/l	0.09	0.00	1.02	0.22	233.34
FE	mg/l	1.21	0.15	5.60	1.35	111.53
K	mg/l	3.87	0.40	7.00	1.66	42.84
MG	mg/l	1.03	0.20	5.20	0.82	79.40
NA	mg/l	3.23	1.00	20.80	2.69	83.17
E-COLI	cfu/100ml	3498.07	0.00	86000.00	11402.45	325.96
SN04						
TEMP	o C	27.43	24.58	29.78	1.10	4.02
COND	µS	64.54	37.80	186.00	28.93	44.82
SAL	ppt	0.02	0.01	0.07	0.01	64.09
TUR	NTU	104.31	2.00	343.00	77.09	73.90
NO3	mg/l	0.66	0.06	3.22	0.40	61.13
CL	mg/l	7.32	2.00	28.00	5.60	76.50
PO4	mg/l	0.08	0.01	0.99	0.21	249.18
FE	mg/l	0.68	0.03	2.02	0.48	71.03
K	mg/l	4.03	0.40	6.40	1.22	30.30
MG	mg/l	0.94	0.20	2.90	0.54	57.05
NA	mg/l	4.15	1.60	24.00	3.79	91.28
E-COLI	cfu/100ml	4950.04	0.00	41000.00	7419.36	149.88

Table 3. Basic statistical analysis for three water quality parameters.

	Unit	Mean	Minimum	Maximum	SD	CV
			SN01			
PH	-	6.39	5.49	7.83	0.45	7.07
SS	mg/l	91.01	11.00	372.00	56.26	61.81
NH3-NL	mg/l	0.14	0.01	1.07	0.18	129.30
SN02						
PH	-	6.22	5.43	7.28	0.36	5.77
SS	mg/l	73.44	7.00	274.00	50.16	68.30
NH3-NL	mg/l	0.10	0.01	0.45	0.11	103.81
			SN03			
PH	-	6.36	5.67	8.41	0.48	7.59
SS	mg/l	72.61	1.00	574.00	83.44	114.91
NH3-NL	mg/l	0.15	0.01	2.46	0.38	254.94
			SN04			
PH	-	6.29	5.59	8.09	0.41	6.56
SS	mg/l	47.98	1.00	146.00	32.05	66.80
NH3-NL	mg/l	0.15	0.01	0.83	0.20	131.79

Table 4. Correlation coefficient between WQP and the input parameters.

	PH	**SS**	**NH3-NL**	**PH**	**SS**	**NH3-NL**	**PH**	**SS**	**NH3-NL**	**PH**	**SS**	**NH3-NL**
		SN01			SN02			SN03			SN04	
TEMP	0.316	-0.171	-0.137	-0.425	0.361	0.014	-0.022	0.090	0.083	-0.295	0.154	-0.076
COND	-0.029	0.301	0.208	-0.113	0.061	0.144	0.216	0.002	-0.069	-0.290	0.083	0.094
NO3	0.228	0.131	0.383	-0.364	-0.101	0.067	-0.183	-0.279	0.201	-0.264	-0.196	0.054
SAL	0.202	-0.043	0.393	0.835	-0.118	-0.115	0.844	-0.071	-0.028	0.757	-0.147	-0.073
TURB	-0.167	0.766	0.137	0.071	0.061	0.000	-0.079	-0.200	0.191	-0.008	0.131	0.221
Cl	-0.114	0.354	0.411	-0.063	0.287	0.084	0.146	-0.076	-0.316	-0.302	0.067	0.245
PO4	0.181	-0.148	0.065	0.025	0.121	-0.083	0.077	-0.114	0.454	0.088	0.052	0.569
K	-0.306	0.184	0.253	-0.005	0.014	-0.108	-0.012	0.039	0.018	0.325	0.013	-0.248
MG	0.038	0.191	0.376	0.247	-0.023	0.152	0.115	-0.104	-0.192	0.020	-0.074	0.142
NA	0.127	0.088	0.400	0.106	0.283	0.077	-0.027	0.104	0.269	-0.268	0.176	0.025
FE	0.023	-0.080	-0.038	-0.165	0.143	-0.001	0.152	-0.045	0.017	-0.345	-0.024	0.106
E-coli	-0.085	0.315	0.007	0.142	0.024	0.014	0.223	-0.095	0.036	-0.042	0.143	0.367

Table 5. ANN architecture for each parameter.

Parameter	No. of neuron	RMSE	Maximum error (%)	TFHL	TFOL	TA
pH	18	0.15	3.22	TS	PL	LMA
SS	17	0.30	3.46	LS	PL	LMA
AN	17	0.26	3.12	TS	PL	LMA

TFHL: Transfer function between input layer and hidden layer; TFOL: Transfer function between hidden layer and output layer; TA: Training algorithm; LS: Log sigmoid; TS: Tan sigmoid; PL: Pure-line; LMA: Levenberg–Marquardt algorithm.

Table 6. The number and types of MFs for each module.

Parameter	AFNIS Module	
	MFs (Type)	MFs (Number)
PH	gbellmf	3 4
SS	gbellmf	4
NH3-NL	gbellmf	3 4 4

Table 7. The running time (seconds) of training process for each model

Model	MLP	RBF	ANFIS	WDT-ANFIS
pH	51	44	67	78
SS	53	46	71	81
AN	49	43	64	75

Table 8. A summary of correlation coefficients for Scenario 1, Scenario 2 and the AI %.

Model	SNO2		SNO3		SNO4		AI (%)		
	Scen1	Scen2	Scen1	Scen2	Scen1	Scen2	SNO2	SNO3	SNO4
pH	0.95	0.98	0.94	0.98	0.93	0.98	3.1	4.1	5.1
SS	0.96	0.97	0.97	0.98	0.97	0.98	1.1	1	1
AN	0.96	0.97	0.96	0.97	0.95	0.97	0.5	0.5	2

***Declaration of Interest Statement**

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

The authors declare no conflict of interest.

Cover Letter

Date: 16/08/2019

Editor-in-Chief
Journal of Hydrology

Dear Editor,

We would like to rsubmit the enclosed revised manuscript "**Machine Learning Methods for Better Water Quality Prediction**" for your consideration for publication in the ***Journal of hydrology***. The authors would like to sincerely thank the Editor in chief, associated editor and the reviewers for the time spent on reviewing our manuscript for possible publication in this admired journal. The authors valued the comprehensive comments and the valuable suggestions given by the reviewers.

We believe that this manuscript is appropriate for publication in ***Journal of Hydrology*** because it is providing AI based model for water quality prediction. In addition, our article lays a foundation for the development of intelligent which should be of broad interest to your readership.

This manuscript has not been published and is not under consideration for publication elsewhere. We have no conflicts of interest to disclose. All authors have read and approved the final version of the manuscript.

Thank you for your consideration, and we look forward to hearing from you at your earliest convenience.

Yours Sincerely,

Rusul Khaleel Ibrahim
Department of Civil Engineering
Faculty of Engineering
University of Malaya
50603 Kuala Lumpur
+6011-11915-730

Graphical Abstract (for review)

Graphical Abstract