

International Journal of Sustainable Agricultural Management and Informatics

ISSN online: 2054-5827 - ISSN print: 2054-5819

<https://www.inderscience.com/ij sami>

Water quality index prediction using artificial neural network: a case study of Selangor River, Malaysia

Jia Jun Tan, Senthil Kumar Arumugasamy, Fang Yenn Teo

DOI: [10.1504/IJSAMI.2025.10067962](https://doi.org/10.1504/IJSAMI.2025.10067962)

Article History:

Received:	21 June 2023
Last revised:	18 August 2023
Accepted:	19 September 2023
Published online:	03 December 2024

Water quality index prediction using artificial neural network: a case study of Selangor River, Malaysia

Jia Jun Tan

Department of Civil Engineering,
University of Nottingham Malaysia,
Semenyih, 43500 Selangor, Malaysia
Email: evyjt2@nottingham.edu.my

Senthil Kumar Arumugasamy*

Department of Chemical and Environmental Engineering,
University of Nottingham Malaysia,
Semenyih, 43500 Selangor, Malaysia
Email: Senthil.Arumugasamy@nottingham.edu.my

*Corresponding author

Fang Yenn Teo

Department of Civil Engineering,
University of Nottingham Malaysia,
Semenyih, 43500 Selangor, Malaysia
Email: FangYenn.Teo@nottingham.edu.my

Abstract: Rapid urban development often leads to deterioration of river water quality, and water quality index (WQI) is a number that represents the water quality of a water body. According to Department of Environment, Malaysia parameters used to calculate WQI are dissolved oxygen (DO), biochemical oxygen demand (BOD), chemical oxygen demand (COD), potential of hydrogen (pH), suspended solids (SS) and ammoniacal nitrogen (AN). Data was collected from the ASEAN Working Group on Water Resources Management website in Sungai-Selangor section. Two ANN models were developed; a prediction model to predict the current WQI, and a forecasting model to predict the future WQI. The prediction model gave good results with very low overall root mean squared error (1.15), an excellent overall regression value (0.97874), and a high correlation with the actual WQI (99.94%). The forecasting model did not provide good result with the overall RMSE of 4.80 and overall regression value of 0.752.

Keywords: artificial neural networks; ANNs; water quality index; WQI; dissolved oxygen; DO; biochemical oxygen demand; BOD; chemical oxygen demand; COD; potential of hydrogen; pH; suspended solids; SS; ammoniacal nitrogen; AN; Malaysia.

Reference to this paper should be made as follows: Tan, J.J., Arumugasamy, S.K. and Teo, F.Y. (2025) 'Water quality index prediction using artificial neural network: a case study of Selangor River, Malaysia', *Int. J. Sustainable Agricultural Management and Informatics*, Vol. 11, No. 1, pp.48–71.

Biographical notes: Jia Jun Tan completed his MEng in Civil Engineering from University of Nottingham Malaysia in 2023. He is currently working as a Graduate Civil Structure Engineer at Arup Jururunding Sdn. Bhd.

Senthil Kumar Arumugasamy is an Assistant Professor of Process Modelling and Control at University of Nottingham Malaysia. His major research interest includes model predictive control, and artificial neural networks. He is also focusing on enzymatic polymerisation for synthesising polymers using enzymes as catalyst. He has published more than 50 peer reviewed journal articles (h-index-14) and has presented in more than 30 international conferences.

Fang Yenn Teo is a Professor of Water Engineering at the University of Nottingham Malaysia. He is the Chair of International Association for Hydro-Environment Engineering and Research, Malaysia Chapter. He is the Professional Engineer with Practising Certificate; Professional Technologist; International Professional Engineer Register; ASEAN Chartered Professional Engineer; Fellow of the Institution of Engineers, Malaysia; Fellow of Higher Education Academy; and Fellow of ASEAN Academy of Engineering and Technology.

1 Introduction

Water is the most essential element on the earth whether it is for human activities or socio-economic development. Nowadays, 90% of sewage in developing countries is discharged untreated directly into water bodies, with 2 million tons of agriculture, sewage, and industrial waste being discharged into the world's water every day (UNESCO World Water Assessment Programme, 2003). This is a very concerning problem that has caused contamination by microorganisms (bacteria or viruses) on drinking water supplies in developing countries like Indonesia due to urbanisation and lack of sewerage (Luo et al., 2019). Furthermore, Flörke et al. (2013) reported that the global withdrawal of industrial and domestic water has increased by 77.7% with 300–1,345 km³ withdrawn over the last six decades (1950–2010), 88% of the extracted water was returned to the system as wastewater with limited treatment. This phenomenon is especially concerning in the Asia-Pacific region which is now experiencing rapid growth in economic as most of the countries are developing countries. Therefore, it can be concluded that the global water resources are under immense pressure as the water demand increases due to expanded industrial and economic activities, and population growth (Chapagain et al., 2022). The worsening of the water quality can only contribute more harm into this problem (Boretti and Rosa, 2019).

1.1 Water quality index

The health of a river is represented by a number called the water quality index (WQI) of the river. WQI summarises a large amount of information of water quality into just one number (Štambuk-Giljanović, 1999). According to the Department of Environment (DOE, Malaysia), WQI is dependent on the six parameters that are regularly recorded which are potential of hydrogen (pH), biochemical oxygen demand (BOD), chemical

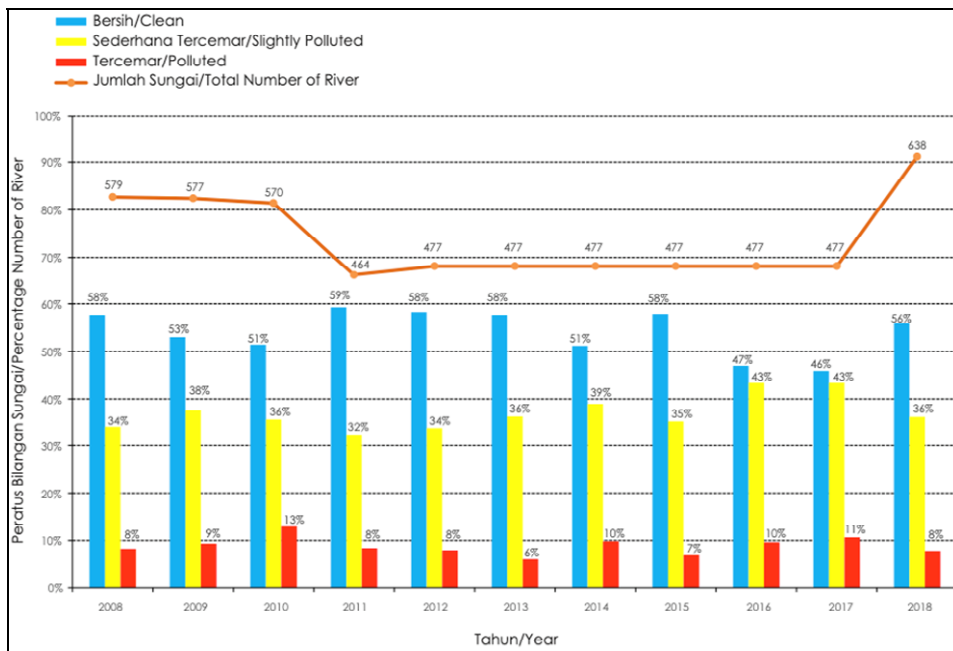
oxygen demand (COD), concentration of ammoniacal nitrogen (AN), suspended solid (SS), and dissolved oxygen (DO). These parameters will then be converted into their relative sub-indexes, which will then be used to calculate the WQI of the river to determine the water quality of the river. These parameters are influenced by the nature as well as the human activities around the water source (Muryani, 2021), therefore by identifying which parameter is the main factor of affecting the WQI, the source of the activity can be tracked down and relative actions can be taken. The river will then be classified into Class I to V according to the value of its WQI.

1.2 River water quality in Malaysia

The status of the river water quality has been a concern because the Malaysia rivers are generally regarded as polluted (Najah et al., 2021). As mentioned in the Environmental Quality Report 2018 (Sekitar, 2018) from DOE, Malaysia, out of the 638 rivers that were monitored that year, 357 (56%) were determined as clean, 231 (36%) were determined as slightly polluted, and 50 (8%) were determined as polluted. Figure 1 shows the river water quality trend from 2008 to 2018 extracted from Environmental Quality Report 2018. It is also mentioned that suspended solids (SS), BOD, and AN were the significant factor that contributed to the river pollution. The sources of SS were mainly due to improper earthworks and land clearing activities, the BOD on the other hand is because of the effluent from agro-based and manufacturing industries or inadequate treatment of sewage, while the source of NH₃-N is determined to be from domestic waste and animal farming.

Therefore, by using artificial neural network (ANN) to predict the current and future WQI, action can be taken earlier to prevent the worsening of the river pollution. It can also save a lot of cost of water quality monitoring because the ANN model are able to quickly compute the WQI with fewer input parameters and input conditions. Water quality modelling techniques have evolved as an acceptable tool to support the surface water management. Modelling techniques are utilised to carry out a systematic and methodical analysis in order to have a better understanding of the relationships between the cause and effect and accessing the impact of changes in ambient water quality (Othman et al., 2020). Water quality is the bridge between the sources of pollution and the water quality processes of the respective water body. Therefore, it can be concluded that the ANN model is a representation of the chemical, biological, and physical water quality processes and mechanisms that occur in a water body.

Applications of ANN in the area of ecological sciences, water engineering, and environmental sciences have been increasing since 1990s and in recent years. ANN have been used extensively for prediction purposes in some water related areas. The AI techniques have been applied and approved recently as an appropriate tool to model composite nonlinear phenomena in water bodies system and hydrology (Othman et al., 2020). There have been researches made to predict and forecast the WQI using ANN models (Palani et al., 2008; Al-Badaai and Shuhaimi-Othman, 2014; Najah et al., 2021). In multi-layer ANN, the neurons are placed in a similar pattern as that of the human brain cells, with each neuron connecting to other neurons with certain coefficients. ANN is a representation of complex, nonlinear functions with many parameters that can be adjusted, trained, or calibrated to match the ANN's own output to the measured output on a known dataset (Palani et al., 2008).

Figure 1 River water quality trend from 2008 to 2018 (see online version for colours)

Source: Sekitar (2018)

Much research has been conducted in the past regarding this topic. Othman et al. (2020) approached the ANN prediction to compute the WQI of the Klang river basin, Malaysia from the WQI parameters when one of the parameters is missing to determine which one of the six WQI parameters (pH, BOD, COD, TSS, AN and DO) is the most influential. The ANN models that is developed in this research was constructed using the Statistica neural networks (SNN) and several input combinations were tested. Dawood et al. (2021) built an ANN model based on the historical data of the City of RI Pedregal, Peru to train the network and to generate a water quality failure in water distribution networks (WDNs) forecast. The network utilised in this research was the extreme learning machine (ELM) model and this research utilised sensitivity analysis to determine which factors have the highest influence to the model result. The factors included were the pipe diameter, pipe thickness, pipe age, pipe installation, water pressure, water quality, and practices of operation and maintenance. It was found that the water pressure, water quality, and practices of operation and maintenance are the most influential ones amongst all the factors. Kulisz et al. (2021) utilised the neural network library in MATLAB and Simulink computing environments to build an ANN model for groundwater WQI prediction. The water samples were collected from 19 wells located near a shale gas extract site in the Lublin, an area in eastern Poland. The researches approached this research by first using seven WQI parameters which are electrical conductivity, calcium, magnesium, sodium ions, potassium, phosphates, and pH the as the inputs and the actual WQI as the output of the ANN model, then slowly decrease the number of inputs in the subsequent tests. It was found that the best result was obtained when the five input neurons which are electrical conductivity, calcium, magnesium, sodium ions, and pH

were chosen as the inputs of the ANN model. The results showed that the ANN model was able to predict the WQI with acceptable results.

Wu et al. (2010) used statistical analysis methods (CA, PCA and fuzzy logic approach) to evaluate the trophic status of water quality in Daya Bay, South China Sea. Cluster analysis was used to identify similar groups of temporal and spatial variations of water quality, the temporal and spatial patterns of trophic status were also determined by principal component analysis and the fuzzy logic approach identified the temporal and spatial patterns of trophic status and the corresponding driving factors.

Burgan et al. (2013) analysed the water quality trends from 2006–2011 of Akarcay River. Measured physical and inorganic chemical parameters in Araplı River, Afyon sewage output, sugar factory output and Bolvadin Bridge of river were used. Statistical properties of data and to understand the variation in time, parameters – time correlation and regression analysis and also parameter-discharge correlations were calculated. According to these analyses results, current pollution status for last five years, variation of quality parameters and interaction among them were determined.

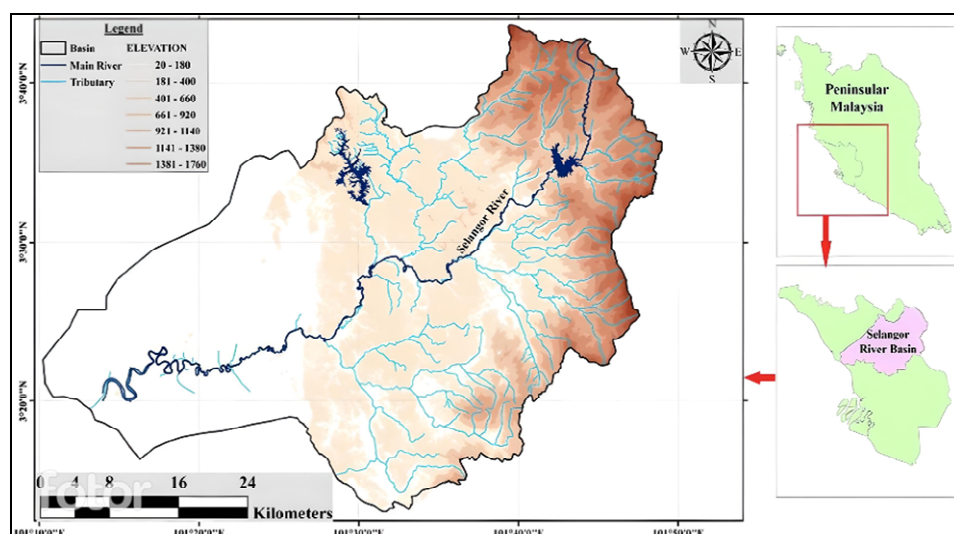
ANN models in general have the limitation of not being successful in providing good prediction outside the range of the dataset used for training. Also, it requires a lot of computational power. In addition, as they are rightly refereed as black box models, it is difficult to explain how the data is processed in ANN to provide the predicted results. Another important limitation is that they need a lot of data for developing good quality models. In this study the major source of error in predicting and forecasting the WQI would be the quality of the data, if the quality is not good then the possibility of better prediction is not that good, i.e., predicted data closely matching with the real data not possible.

In this study, the pH, BOD, COD, concentration of AN, total suspended solid (TSS), and dissolved oxygen (DO) were used as inputs and WQI was used as outputs for ANN modelling. Therefore, by predicting the current and future WQI using ANN, actions can be taken beforehand to prevent the worsening of the water quality in the given water body. ANN have been extensively utilised in extensively for prediction and forecasting purposes in some water related areas. In this study, the main objective is to use ANN model to predict the current and future WQI of Sungai Selangor, by using the various numbers of water quality parameters such as pH, BOD, COD, AN, TSS, and DO. From the literature studies it can be seen the input parameter has been different in each of these studies and also the region chosen for the studies are also different. Hence the novelty of this work is that even though the parameter for the current study is similar to that of Othman et al. (2020), the river chosen for the study is different (Sungai Selangor).

2 Materials and methodology

2.1 Description of study area

Malaysia is a tropical country located at latitude 4.2150° and longitude 101.9758° between Thailand and Singapore. Malaysia has two parts, West Malaysia (Peninsular Malaysia), and East Malaysia (Sabah and Sarawak). Sungai Selangor basin is located within the state of Selangor, situated in the centre and on the west coast of Peninsular Malaysia as shown in Figure 2. Sungai Selangor Basin has a catchment area of about 2,200 km², nearly a quarter of the total area of Selangor State.

Figure 2 Map of Selangor River Basin (see online version for colours)

Source: Seyam and Othman (2015)

Table 1 Statistical properties

	<i>pH</i>	<i>BOD</i>	<i>COD</i>	<i>Amm-N</i>	<i>TSS</i>	<i>DO</i>
Mean	0.99	2.82	2.69	819.75	226.65	48.07
Standard error	0.12	0.21	0.38	156.07	7.79	1.78
Median	0.00	2.00	0.00	40.00	200.00	50.00
Mode	0.00	2.00	0.00	500.00	151.00	0.00
Standard deviation	2.24	3.74	6.75	2,804.98	139.93	31.94
Sample variance	5.03	14.00	45.58	7,867,915.22	19,580.46	1,020.05
Kurtosis	1.50	68.35	5.11	78.81	9.72	2.86
Skewness	1.85	7.49	2.45	7.90	2.53	0.82
Minimum	0.00	0.00	0.00	0.00	0.00	0.00
Maximum	8.00	43.00	32.00	33,300.00	1,120.00	200.00
Sum	319.10	911.90	870.20	264,780.35	73,209.00	15,528.10
Count	323.00	323.00	323.00	323.00	323.00	323.00
Confidence level (95.0%)	0.25	0.41	0.74	307.05	15.32	3.50
Upper confidence interval	1.23	3.23	3.43	1,126.81	241.97	51.57
Lower confidence interval	0.74	2.41	1.96	512.70	211.34	44.58
Coefficient of variation	227.09	132.52	250.59	342.17	61.74	66.43

Selangor is the most developed state in Malaysia with a population of approximately 6.56 million. Due to its strategic location, the Selangor State has been experiencing rapid development which can lead to severe pollution to its river basin (Bassi and Kumar, 2017). Out of the four major river basins in Selangor namely Langat, Klang, Bernam, and Selangor, Sungai Selangor basin is the most important basin to the country as it is

recognised as the largest source of water, providing up to 60% of the total water required in both Selangor and Kuala Lumpur (Subramaniam, 2004). Besides that, it is also a well-known firefly ecotourism. Table 1 provides the details of the statistical properties.

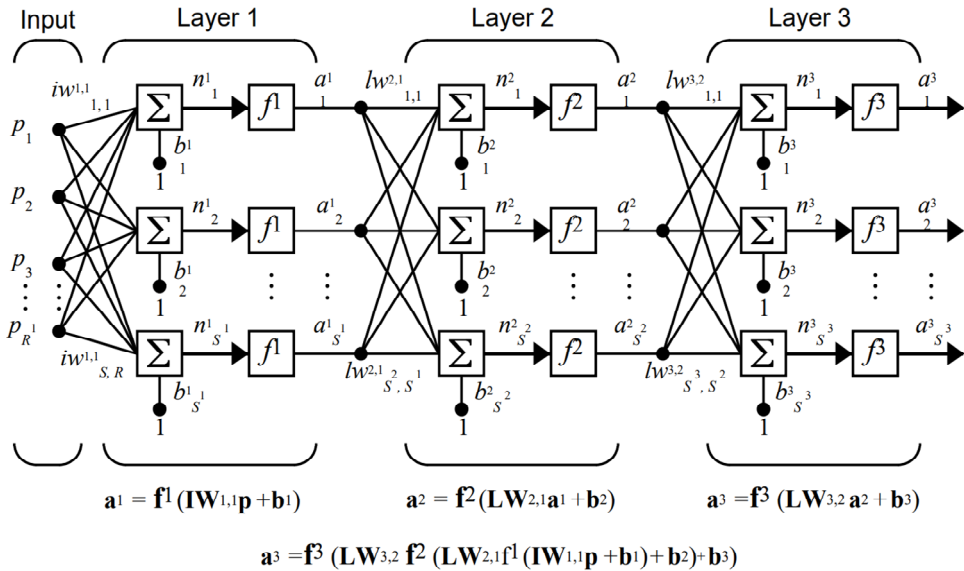
2.2 River water quality monitoring data

The data used in this study were obtained from the ASEAN Working Group on Water Resources Management website, the 26-year data ranges from February 1993 to November 2019 with significant gaps in between the dataset. The data are provided by the Department of Irrigation and Drainage Malaysia (DID) and only the six parameters which are pH, BOD, COD, AN, TSS, and DO as mentioned above. These data are specifically chosen to be because the standard of calculating the WQI in Malaysia is based on these six parameters.

2.3 Artificial neural network

ANN is a computational model that can utilise learning algorithms to simulate a complex system behaviour and approximate nonlinear correlations between the given input and output data without any preceding information about the internal relations among their components (Makarynska and Makarynsky, 2008). Typically, ANN need a certain amount of past data to be able to train until satisfactory to ensure that the predicted results are not too far off from the original output data.

Figure 3 A three-layer MLP network with different numbers of neurons in different layers



Source: Beale et al. (2010)

The MLP neural network is a well-built model designed to perform well in modelling complex and nonlinear phenomena. A feedforward MLP network usually have one input layer, followed by one or more hidden layers of sigmoid neurons and finally an output

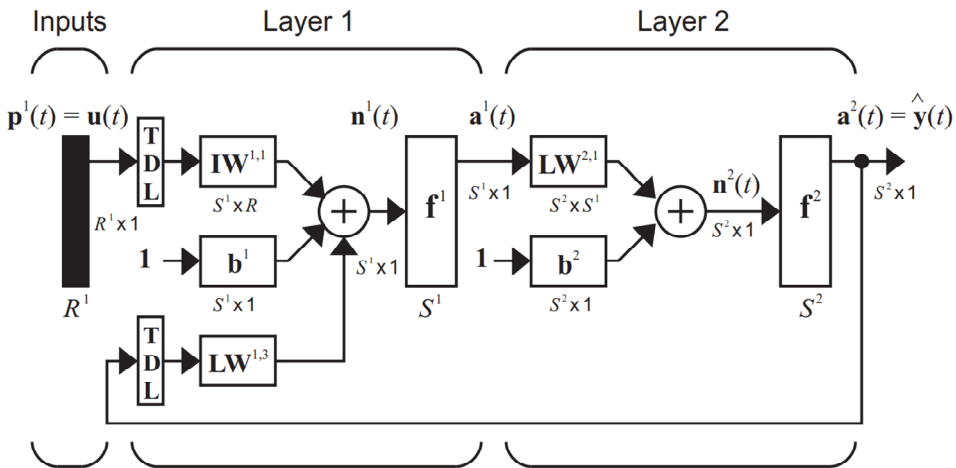
layer of linear neurons (Beale et al., 2010). This model is typically called ‘prediction model’. Figure 3 shows the same three-layer network with different notation. Layers 1 and 2 are hidden layers while layer 3 is known as the output layer. The network shown below has R^1 inputs, S^1 neurons in the first layer, S^2 neurons in the second layer, and so on. Each layer has a weight matrix (W), a bias vector (b), and an output vector (a). The outputs of each layer are the inputs of the following layer. An artificial neuron in a typical MLP neural network receives a set of inputs with weight (IW) and bias vector, then uses the summation function to calculate the input vector (n), and finally uses the activation function (f) to produce an output (a).

On the other hand, the ANN can also be used in predicting future outputs using the past inputs and outputs. This model is typically called ‘forecasting model’. One of the most popular forecasting models that is utilised in this field is the nonlinear autoregressive models with exogenous inputs (NARX) recurrent neural models. As compared to the other recurrent neural architectures, NARX neural networks have limited feedback architectures that only come from the output neuron instead of the hidden neurons. It is proven that the NARX neural networks are computationally powerful in theory (Siegelmann et al., 1997). It has also been reported that gradient-descent learning can be more effective in NARX networks than in other recurrent architectures with ‘hidden states’ (Horne and Giles, 1995). Figure 4 shows a general two-layer NARX network.

2.4 Data pre-processing

Since there is a huge data gap in the provided data, interpolation is needed to fill in the gaps from September 1999 to January 2005, January 2008 to January 2009, and December 2009 to December 2011. The gap from May 2013 to April 2019 is unable to be interpolated because there is only two sets of data in 2019 which is extremely insufficient for MATLAB to interpolate, therefore in this study, the two sets of data in 2019 will be excluded to increase the quality of the data.

Figure 4 A general two-layer NARX network



Source: Beale et al. (2010)

Therefore the data is now a 20-year data that ranges from February 1993 to May 2013. The interpolated data have a regular time interval which is at the 1st and 15th of January, March, May, July, September, and November. The reasoning behind this time interval is to mimic the behaviour of the original data. After interpolating the data, the six parameters were converted into their respective sub-indexes before calculating the WQI. The pH, BOD, COD, AN, TSS, and DO was converted to SIpH, SIBOD, SICOD, SIAN, SISS, and SIDO respectively. This conversion is needed because it minimises the impact of some unrealistic raw data to the WQI value. The best fit equations used for the estimation of the six sub-index values are shown in Table 2.

Table 2 Best fit equations of the sub-index

<i>Sub-index parameter</i>	<i>Value</i>	<i>Condition</i>
SIpH	$17.2 - 17.2 \text{ pH} + 5.02 \text{ pH}^2$	$\text{pH} < 5.5$
	$-242 + 95.5 \text{ pH} - 6.68 \text{ pH}^2$	$5.5 \leq \text{pH} < 7$
	$-181 + 82.4 \text{ pH} - 6.05 \text{ pH}^2$	$7 \leq \text{pH} < 8.75$
	$536 - 77.0 \text{ pH} + 2.76 \text{ pH}^2$	$\text{pH} \geq 8.75$
SIBOD	$100.4 - 4.23 \text{ BOD}$	$\text{BOD} \leq 5$
	$108e^{-0.055} \text{ BOD} - 0.1 \text{ BOD}$	$\text{BOD} > 5$
SICOD	$-1.33 \text{ COD} + 99.1$	$\text{COD} \leq 20$
	$103e^{-0.0157} \text{ COD} - 0.04 \text{ COD}$	$\text{COD} > 20$
SIAN	$100.5 - 105 \text{ AN}$	$\text{AN} \leq 0.3$
	$94e^{-0.573\text{AN}} - 5 \text{AN} - 2 $	$0.3 < \text{AN} < 4$
	0	$\text{AN} \geq 4$
SISS	$97.5e^{-0.00676\text{SS}} + 0.05\text{SS}$	$\text{SS} \leq 100$
	$71e^{-0.0061\text{SS}} - 0.015\text{SS}$	$100 < \text{SS} < 1,000$
	0	$\text{SS} \geq 1,000$
SIDO	0	$\text{DO} \leq 8$
	100	$\text{DO} \geq 92$
	$-0.395 + 0.030 \text{ DO}^2 - 0.00020 \text{ DO}^3$	$8 < \text{DO} < 92$

Source: Othman et al. (2020)

After calculating the sub-indexes, WQI can be calculated using equation (1) which complies to the DOE, Malaysia standard.

$$\begin{aligned} \text{WQI} = & 0.22 \text{ SIDO} + 0.19 \text{ SIBOD} + 0.16 \text{ SICOD} \\ & + 0.16 \text{ SISS} + 0.15 \text{ SIAN} + 0.12 \text{ SIpH} \end{aligned} \quad (1)$$

where $0 \leq \text{WQI} \leq 100$.

Figure 5 and 6 show the incomplete WQI dataset and the WQI dataset after interpolation respectively. With the six parameters as the inputs and the calculated WQI as the output for the ANN model, the data still need to be normalised. Normalisation is a data preparation technique that is used extensively in machine learning like ANNs modelling, the process of normalisation is transforming the columns in a dataset to the same scale. There are two types of normalisation method that is widely used in machine learning, namely min-max scaling and standardisation scaling. In this study, min-max

scaling is utilised since it is the default normalisation method in MATLAB ANN modelling. The equation for min-max scaling is shown in equation (2).

$$X = (X - X_{\min}) / (X_{\max} - X_{\min}) \quad (2)$$

Figure 5 Incomplete WQI dataset (see online version for colours)

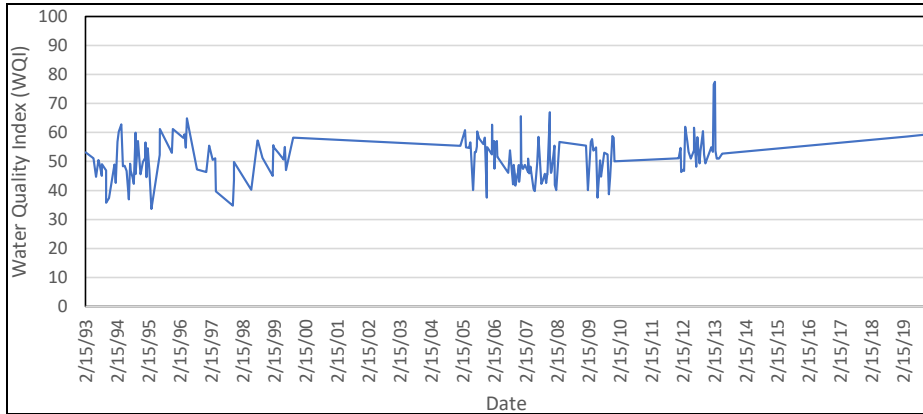
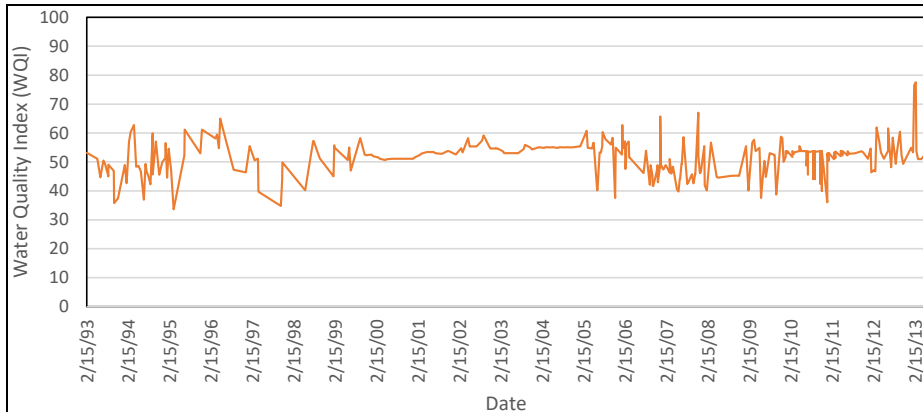


Figure 6 WQI dataset after interpolation (see online version for colours)



2.5 Data partition

The data were categorised into training, testing, and validation data. The training dataset is used during the learning process and to determine the adjusted weight and biases of a network. The test dataset is independent to the training dataset that is used for calibration to prevent overtraining the network. The validation dataset is to avoid overfitting by tuning the number of hidden neurons in each layer of the ANN model (Palani et al., 2008). Throughout this study, 70% of the data is categorised as training data, 15% of the data is categorised as testing and validation data respectively and the method of dividing the data into their respective roles is set to randomised.

2.6 Architecture, training algorithms, and parameter optimisation techniques

After the data is divided into three parts data is subjected to processing. Processing includes trial and error method being used with hidden neurons to optimise the best performing hidden neurons that can given good results in terms of regression (R) and lowest errors. In this study 1 to 15 hidden neurons were tested to optimise the best hidden neurons. In ANNs, there are 12 training algorithms (TAs) falling under six different classes namely six different classes of TAs are used:

- 1 additive momentum [gradient descent with momentum (GDM)]
- 2 self-adaptive learning rate (variable learning rate gradient descent and GDM and adaptive learning rate)
- 3 resilient back propagation
- 4 conjugate gradient back propagation [scaled conjugate gradient (SCG), conjugate gradient with Powell/Beale restarts (CGP), Fletcher-Powell conjugate gradient and Polak-Ribière conjugate gradient]
- 5 Quasi-Newton [Levenberg-Marquardt (LM), BFGS Quasi-Newton and one step secant]
- 6 Bayesian regulation propagation [Bayesian regularisation (BR)].

In this study Quasi-Newton (LM) was used. With TA and number of hidden neurons chosen the model is developed by varying the hidden neurons and the training is carried out till a high regression value is obtained.

2.7 Performance parameters of the ANN models

Evaluating the performance of an ANN model is essential. For the feedforward ANN model, it is evaluated by comparing the predicted results to the actual outcome of the data to calculate the mean squared error (MSE), root mean squared error (RMSE), mean absolute error (MAE), mean absolute percentage error (MAPE), and accuracy of the data by using the equations below. While for the NARX recurrent neural model, the only performance parameter is MSE.

$$MSE = \frac{1}{N} \sum (X_{\text{observed}} - X_{\text{predicted}})^2 \quad (3)$$

$$RMSE = MSE^{1/2} \quad (4)$$

$$MAE = \frac{1}{N} \sum |X_{\text{observed}} - X_{\text{predicted}}| \quad (5)$$

$$MAPE = \frac{1}{N} \sum |(X_{\text{observed}} - X_{\text{predicted}})/X_{\text{observed}}| \quad (6)$$

$$Accuracy = 1 - ((X_{\text{observed}} - X_{\text{predicted}})/X_{\text{observed}}) * 100 \quad (7)$$

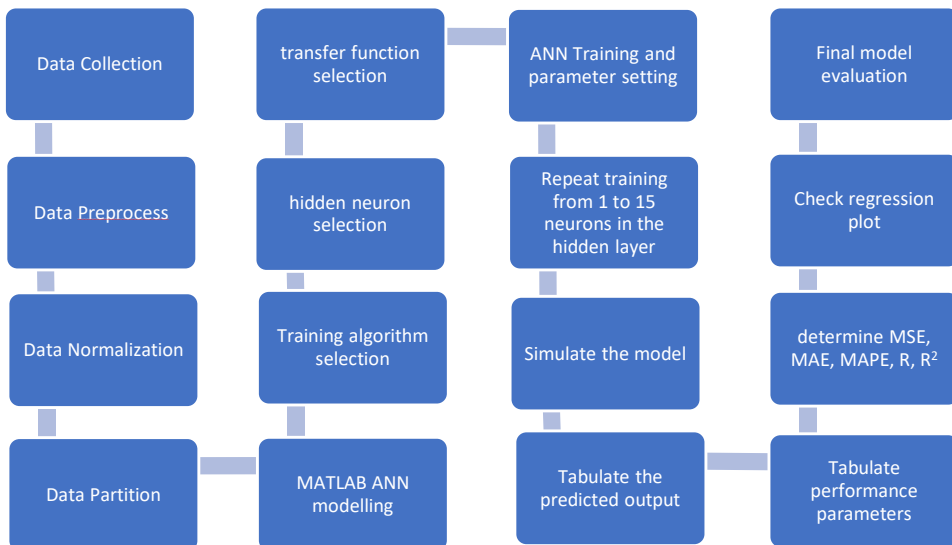
where N is the total number of observations data.

2.8 ANN model setup

2.8.1 Feedforward network model

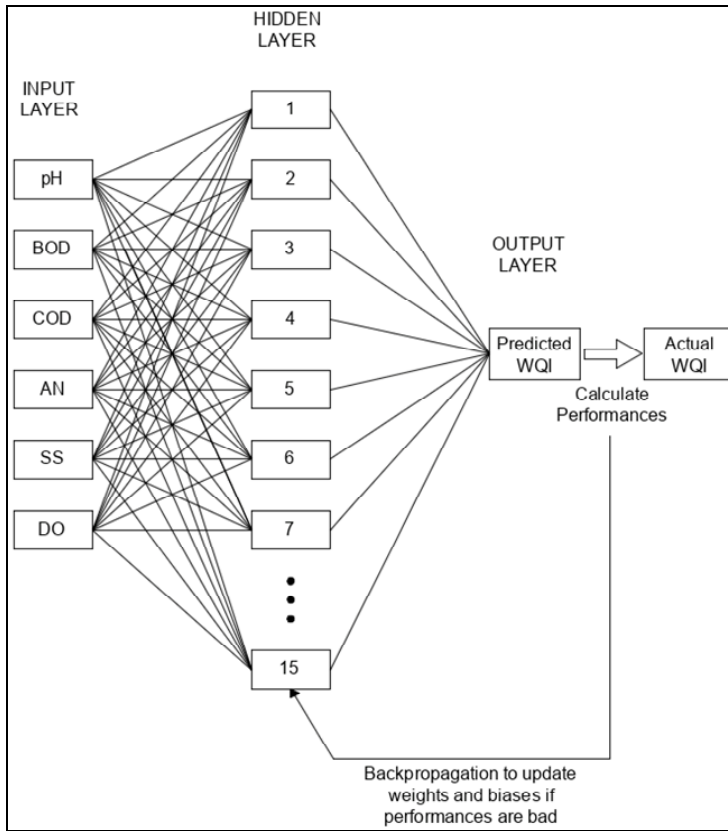
Generally, the steps of building an ANN model follow the flowchart shown in Figure 7. In this study, the ANNs model was constructed using MATLAB. In MATLAB, the first step is to prepare the input and output for the ANNs model. Then, 70% of the data is chosen for training, 15% of the data is chosen for both validation and testing. To make sure the data is categorised randomly, 'dividerand' is chosen as the divide function. With all of the steps done, the model can now be constructed. The network function used is 'feedforwardnet' and the training function is set as 'trainlm' (LM backpropagation TA). The network is then trained to provide an output. The hidden neurons of the ANNs model is set as a range from 1 until 15 and the respective MSE, MAE, MAPE and accuracy for each ANNs models with the respective hidden neurons are tabulated. This is done so that the hidden neuron with the best outcome can be chosen. Figure 8 shows the ANN structure of the feedforward network model.

Figure 7 The process of building an ANNs model (see online version for colours)



2.8.2 NARX recurrent neural model

For predicting the future data, a different network named 'narxnet' is chosen to carry out time series prediction as it is designed as a feedforward time delay neural network to predict the future value of a dataset. The data is imported and divided in the same manner as used for FANN. Then, the 'tonndata' command is used to convert the data to standard neural network cell array form. After that, an open loop network can be created by setting the training function as 'trainbr' (BR backpropagation TA) because this training function is capable of analysing more complex problems.

Figure 8 ANN structure of the feedforward network model

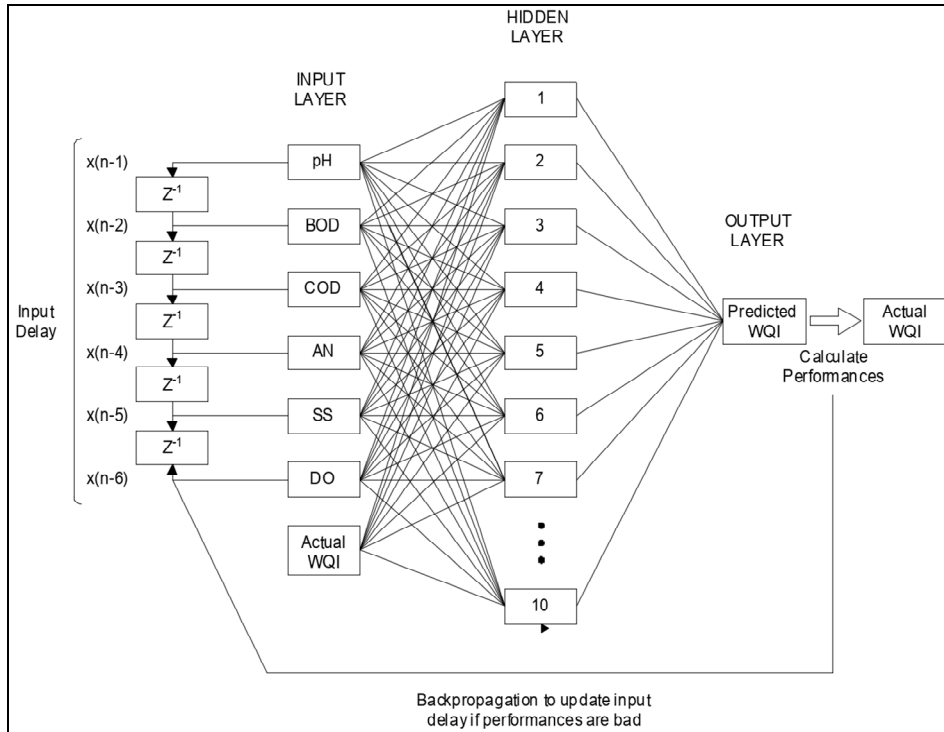
Next, the input delays, feedback delays, and the hidden neurons are set as from 1 to 6, from 1 to 80, and 10 respectively. There is no definitive set of rule for determining the ranges, it is purely based on trial and error to get the best performance of the model. After that, the network can be trained by using the 'prepaerts' function to prepare the input and target time series data for network training. Then, the network is closed to proceed to run the multi-step prediction. This is the step where the future values of WQI that is out of the range of the given data can be predicted. Finally, the plot of the original dataset and the predicted dataset can be plotted. Figure 9 shows the ANN structure of the NARX recurrent neural model.

2.9 Model validation and verification

After getting the output from the ANN models, specific rules to evaluate the performance of the model are applied to ensure the quality of the output. The prediction model and forecasting model have different evaluation process due to the nature of the modelling techniques and different types of network used. For the feedforward network model, the parameters used to evaluate the performance are MSE, MAE, MAPE, accuracy, and regression. The MSE represents the squared difference between the actual output and the model output, therefore by rooting is, RMSE is introduced and the direct difference

between the datasets can be observed. For an example, if the MSE between 1 and 3 is 4, while the RMSE between 1 and 3 is 2. Dikmen et al. (2005) stated that the RMSE and MAE should be close to zero in order justify the model as reliable, but the way to justify whether the value of RMSE and MAE are close to zero depends on the dataset itself.

Figure 9 ANN structure of the NARX recurrent neural model



Therefore, to be able to accurately justify it, MAPE is introduced because MAPE expresses the accuracy of the model output as a percentage of error. For example, if the MAPE of the model is 3, it means that the model output is off by 3%. This gives a clearer idea on the performance of the model. After knowing the how much is the model output is off when compared to the actual output, it is still unclear that if the offset is an overestimation or an underestimation. To solve this problem, the accuracy of the model is calculated. For an instance, if the accuracy of the model is 101.8910, on average, the model is overestimating the output by 1.8910%. Finally, the regression plot of the model shows the relationship between the outputs of the network and the actual results. In a perfect training scenario, the network outputs and the actual results would be exactly equal, therefore the regression value would be 1, but this is practically impossible. Table 3 shows what different ranges of regression values tell about the performance of the model. High value of regression does not necessarily means that the model is performing well, sometimes the regression value of a good model can be low and vice versa. This is why various performance parameters are needed to convey the reliability of a model.

Table 3 Description of different ranges of regression value

<i>Range of regression value</i>	<i>Description</i>
$r < 0.3$	None or very weak effect size
$0.3 < r < 0.5$	Weak or low effect size
$0.5 < r < 0.7$	Moderate effect size
$r > 0.7$	Strong effect size

Source: Zikmund et al. (2013)

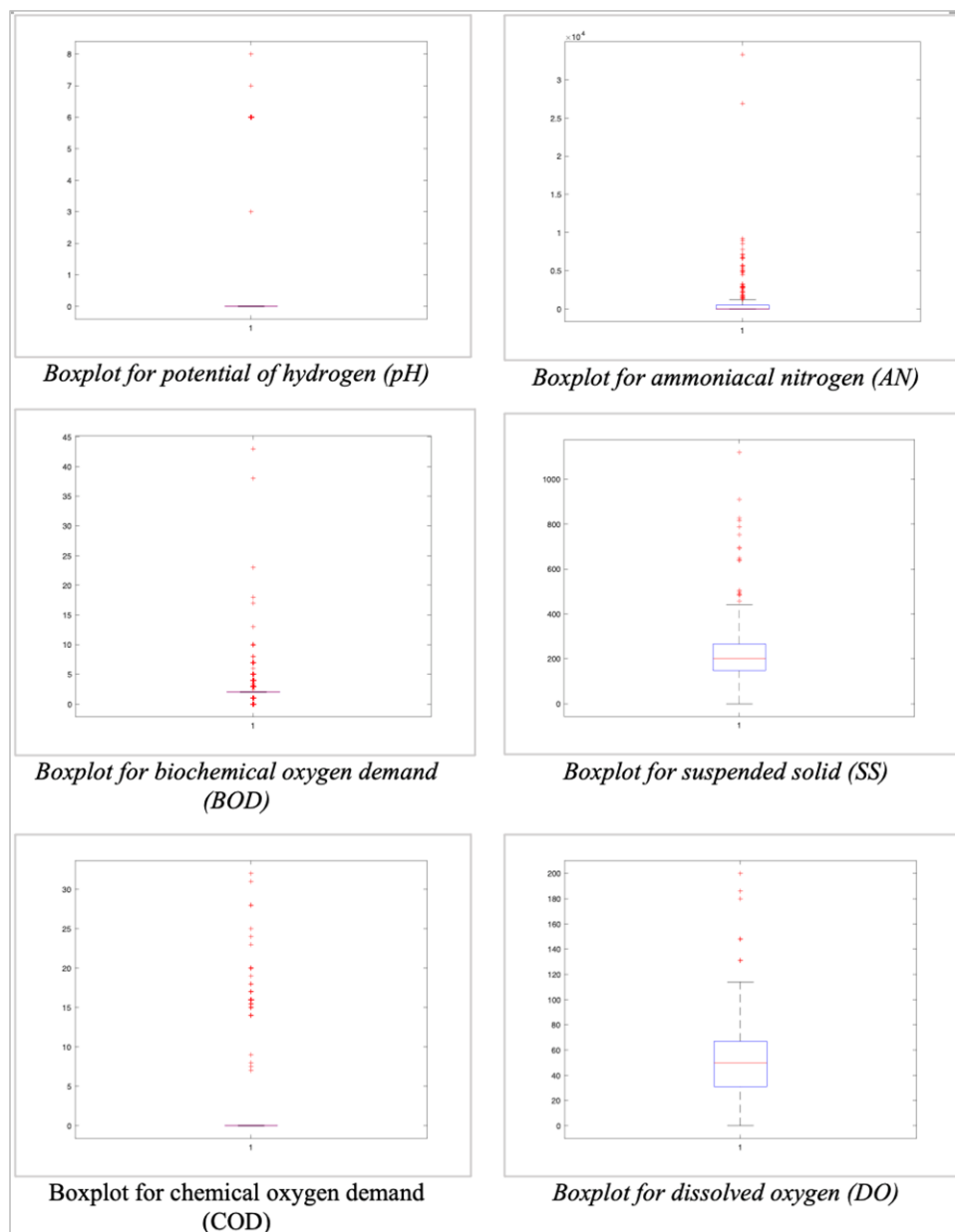
On the other hand, for the NARX recurrent neural model, the performance can only be evaluated by RMSE and regression value. The other parameters cannot be calculated due to the limitations where the future forecasted values do not have the existing data to compare with. In a nutshell, only by analysing the values of all the parameters mentioned above, the performance of the models can be justified.

3 Results and discussion

3.1 WQI parameters evaluation and classification

Since majority of the data is missing, interpolation for each WQI parameters is needed to fill in the gaps between the respective datasets. But the main challenge in this step is that the dataset provided have an abnormally huge range that is not logical. Figure 10, shows the boxplots for pH, BOD, COD, AN, TSS, DO respectively. Normally, a boxplot contains a horizontal red line in the middle that stands for the median of the data, with a horizontal blue line on top and below the red line representing the 75th percentile and 25th percentile respectively. The horizontal grey line at the very top and bottom represents the upper adjacent and lower adjacent respectively. As clearly shown in figure pH, BOD, COD, and AN datasets are not able to form a normal boxplot because most of the data are clustered in the lower end of the respective values with some data that is unreasonably high or low. In Figure 10 the shape of the boxplot seems to be reasonable. Figure 10 shows that the median of the TSS dataset is 200, the 75th percentile and 25th percentile are 266.25 and 147.75 respectively, and the upper adjacent and lower adjacent are 440 and 0 respectively. For Figure 10, it shows that the median of the DO dataset is 50, the 75th percentile and 25th percentile are 67 and 31 respectively, and the upper adjacent and lower adjacent are 114 and 0 respectively. However, the readings for DO is out of the normal range based on Table 2, typically, the readings of DO range from 1 to 7, but the dataset provided gives an abnormally high value that exceeds the range by a lot.

Therefore, in conclusion, only the SS readings in the whole dataset is normal. Figure 6 shows that most of the time, the Selangor River has a WQI value of a range from 40 to 60. This generally means that the Selangor River has been in a polluted state since 1993 to date.

Figure 10 Boxplots of WQI parameter (see online version for colours)

3.2 Prediction of current WQI using feedforward network model

The feedforward network model is developed to predict the WQI by only using the six WQI parameters (pH, BOD, COD, AN, TSS, DO) as inputs for the model, and actual WQI as output for the model to learn. After the model output is produced, the comparison between the predicted WQI and the actual WQI, and the regression plots are shown in

Figure 11 and 12. The comparison clearly shows that there is a high correlation between the predicted WQI and actual WQI. The regression plots suggest that the model is able to produce an output that behaves extremely similar to the actual output with the regression values of training, test, validation, and overall dataset being very close to 1, which are 0.97318, 0.99469, 0.98619 and 0.97874 respectively. For the values of the performance parameters which are the RMSE, MAE, MAPE, and accuracy for every hidden neurons from 1 to 15 are tabulated in Tables 4 and 5. Since the validation dataset is used to evaluate the ANN model, the ANN model with the lowest validation RMSE is considered as the best performing model out of the 15 hidden neurons, which in this case is the model with 11 hidden neurons having a RMSE value of 0.94.

Other than the RMSE value, the other performance parameters for every dataset on the model with 11 hidden neurons tend to have excellent results if not the best results. The overall RMSE, validation MAE, overall MAE, overall MAPE, and overall accuracy have the best results when compared to the other models, with the values being 1.15, 0.60, 0.61, 0.06 and 99.94 respectively. The training RMSE, training MAE, training MAPE, and test accuracy have the second best results with the values of 1.20, 0.61, 0.03 and 99.76 respectively. The test RMSE, test MAE, test MAPE and validation accuracy have the third best results with the values of 1.16, 0.61, 0.38 and 99.83. Lastly, the validation MAPE and training accuracy are the fifth and seventh best performing results with the values of 0.17 and 100.23 respectively. Table 6 summarises the results ranking for each performance parameters of the feedforward network model with 11 hidden neurons. It can be seen that all the performance parameters of the overall dataset have the best results which is further proving the fact that the model with the lowest validation MSE is the best performing model.

The MAPE values and the accuracy of the model says that the model output is only off by 0.06% when compared to the actual data, and the accuracy is 99.94% similar to the actual data. Together with the regression value of the model, it can be concluded that this prediction model is very satisfactory.

Figure 11 Comparison of the predicted WQI and the actual WQI (see online version for colours)

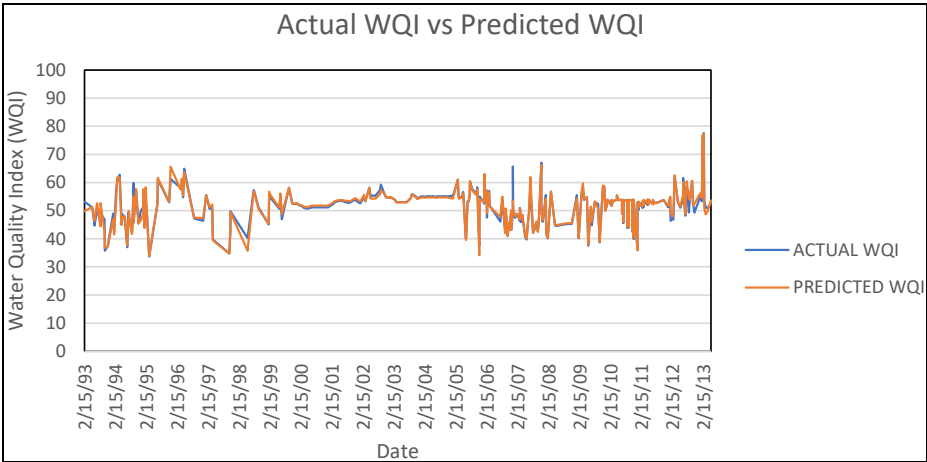
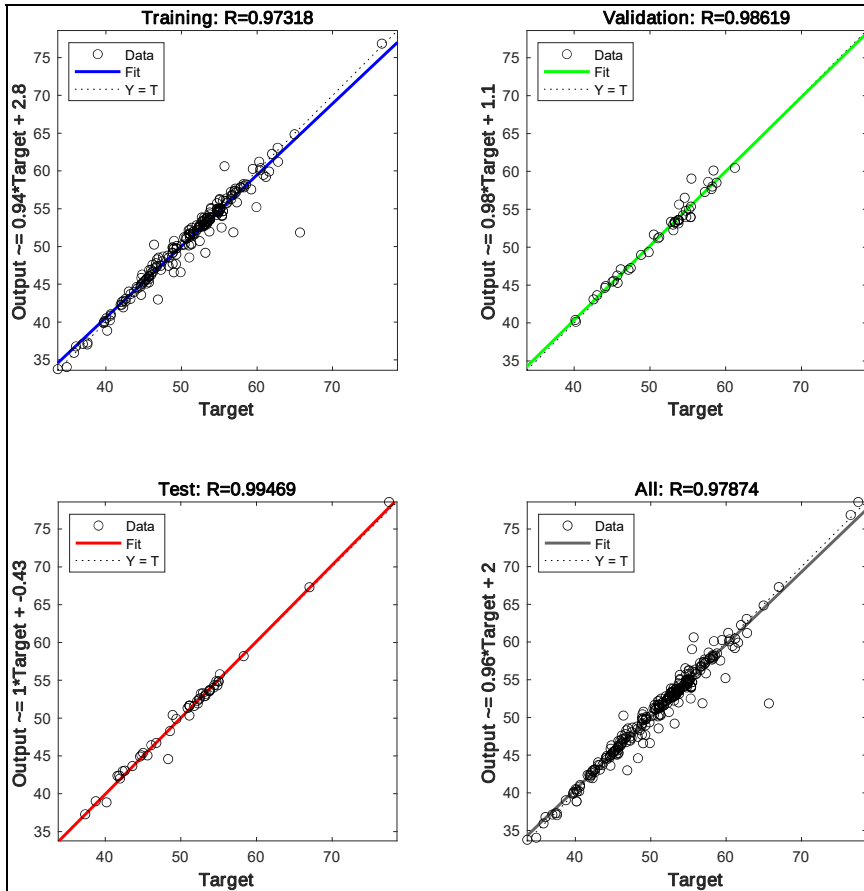


Figure 12 Training, test, validation and overall regression plots of prediction model (see online version for colours)



3.3 Multi-step prediction of future WQI using NARX recurrent neural model

The NARX recurrent neural model is developed to predict the future WQI values by using the currently available data such as the six WQI parameters (pH, BOD, COD, AN, SS and DO) as inputs and the currently available WQI as output. Figure 13 shows the graph of actual and predicted WQI. It can be seen that the NARX recurrent neural model could not predict the WQI after the 400th dataset as the pattern of the WQI become less and less clearer as it approaches the end. Figure 14 shows the regression plots for the training, test, and overall datasets with the values of 0.76295, 0.70941 and 0.75200 respectively, which is considered a strong effect size.

The only performance parameter for this model is RMSE. The training RMSE, test RMSE, and the closed loop performance, which is the overall RMSE of the forecasting model, with the values of 3.74, 4.86, and 4.80 as shown in Table 7 are stating out the fact that the forecasted WQI have an acceptable amount of difference with the actual WQI. This means that in general, each forecasted WQI only have a difference of 4.80 from the actual WQI which is a considerably small range considering the WQI can only range

from 0 to 100. This proved that the ANNs modelling techniques are able to produce a satisfactory result despite the sub-par quality and the incomplete dataset. In a nutshell, the performance of the forecasting model is acceptable because of the low value of RMSE.

Table 4 MSE and MAE for training, test, validation, and overall dataset

<i>No. of hidden neuron</i>	<i>Training dataset RMSE</i>	<i>Testing dataset RMSE</i>	<i>Validation dataset RMSE</i>	<i>Overall dataset RMSE</i>	<i>Training dataset MAE</i>	<i>Testing dataset MAE</i>	<i>Validation dataset MAE</i>	<i>Overall dataset MAE</i>
1	2.51	2.46	1.88	2.42	1.37	1.46	1.21	1.36
2	2.48	2.95	2.43	2.55	1.55	1.46	1.41	1.51
3	2.52	2.04	1.46	2.32	1.46	1.31	0.98	1.36
4	1.37	1.77	2.28	1.60	0.96	1.15	1.11	1.01
5	2.12	2.85	1.44	2.16	1.20	1.43	1.10	1.22
6	1.42	2.01	1.75	1.58	0.84	1.21	1.13	0.94
7	1.53	0.71	1.13	1.38	0.70	0.43	0.76	0.67
8	1.88	1.17	1.61	1.75	1.14	0.90	1.18	1.11
9	1.43	2.24	1.68	1.62	0.83	1.40	0.93	0.93
10	1.10	2.39	2.26	1.57	0.65	0.98	1.19	0.78
11	1.20	1.16	0.94	1.15	0.61	0.61	0.60	0.61
12	1.26	2.37	1.00	1.45	0.58	0.97	0.60	0.64
13	1.31	0.81	1.57	1.29	0.72	0.58	1.01	0.74
14	1.20	2.21	1.70	1.48	0.64	0.83	0.91	0.71
15	1.27	1.57	4.03	1.98	1.01	1.15	2.04	1.19

Table 5 MAPE and accuracy for training, test, validation, and overall dataset

<i>No. of hidden neurons</i>	<i>Training dataset MAPE</i>	<i>Testing dataset MAPE</i>	<i>Validation dataset MAPE</i>	<i>Overall dataset MAPE</i>	<i>Training dataset ACC</i>	<i>Testing dataset ACC</i>	<i>Validation dataset ACC</i>	<i>Overall dataset ACC</i>
1	0.43	0.88	0.11	0.41	100.05	98.86	100.23	99.59
2	0.26	0.02	0.3	0.13	100.84	101.39	100.86	99.87
3	0.6	0.52	0.52	0.26	99.86	100.99	100.91	99.74
4	0.18	0.81	0.72	0.1	100.51	99.2	99.48	99.9
5	0.13	1.46	0.07	0.33	100.33	98.99	100.41	99.67
6	0.1	1.08	0.63	0.33	100.46	101.63	101.2	100.33
7	0.34	0.14	0.27	0.25	99.85	100.22	99.35	99.75
8	0	1.36	1.14	0.37	100.39	102.27	102.1	100.37
9	0.23	0.62	0.04	0.24	100.72	101.9	100.17	100.24
10	0.48	1.47	0.32	0.61	99.43	98.56	99.93	99.39
11	0.03	0.38	0.17	0.06	100.23	99.76	99.83	99.94
12	0.06	0.92	0.01	0.18	100.08	98.4	99.98	99.82
13	0.03	0.38	0.24	0.11	100.16	100.63	100.59	100.11
14	0.05	0.55	0.46	0.12	100.13	100.93	100.73	100.12
15	1.27	0.66	0.6	1.08	102.18	101.39	102.91	101.08

Table 6 Results ranking for each performance parameters of the feedforward network model with 11 hidden neurons

<i>Performance parameter</i>	<i>Value</i>	<i>Ranking amongst other models</i>
RMSE_TRAIN	1.20	2nd
RMSE_TEST	1.16	3rd
RMSE_VAL	0.94	1st
RMSE_ALL	1.15	1st
MAE_TRAIN	0.61	2nd
MAE_TEST	0.61	3rd
MAE_VAL	0.60	1st
MAE_ALL	0.61	1st
MAPE_TRAIN	0.03	2nd
MAPE_TEST	0.38	3rd
MAPE_VAL	0.17	5th
MAPE_ALL	0.06	1st
ACC_TRAIN	100.23	7th
ACC_TEST	99.76	2nd
ACC_VAL	99.83	3rd
ACC_ALL	99.94	1st

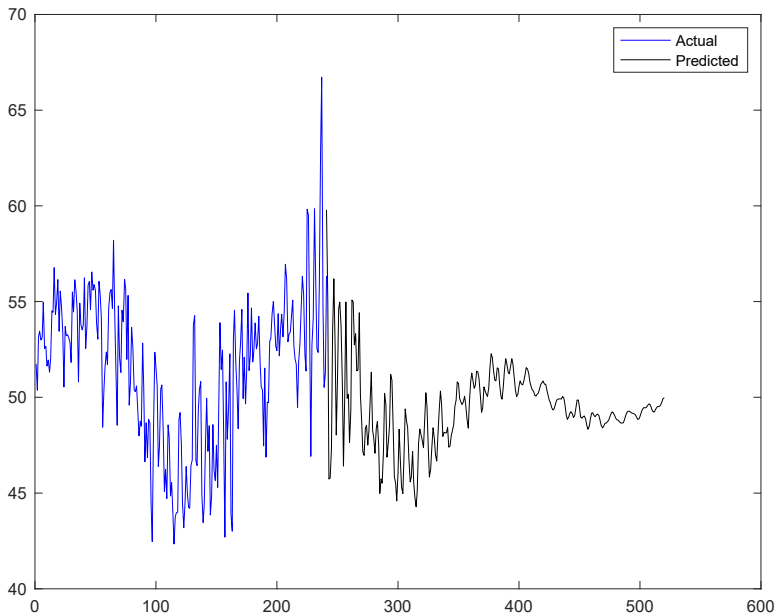
Figure 13 Graph of actual and predicted WQI (see online version for colours)

Figure 14 Training, test, and overall regression plots of NARX recurrent neural model (see online version for colours)

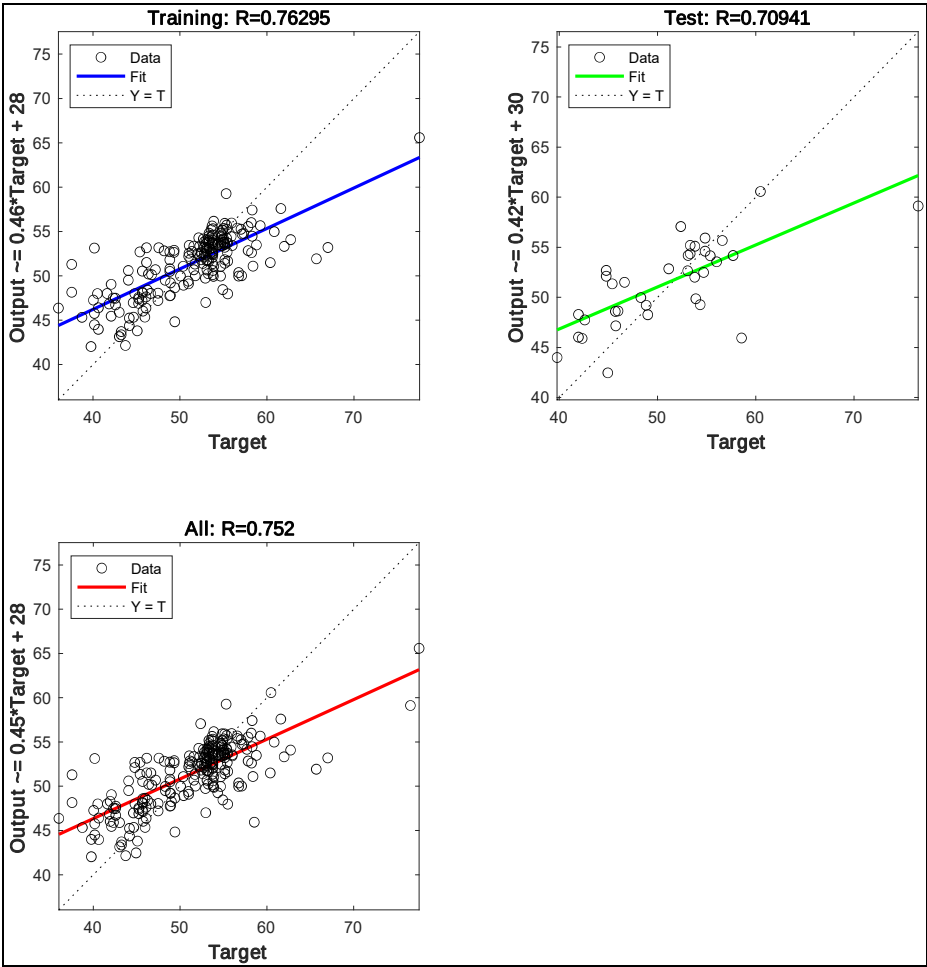


Table 7 MSE for training, test, and overall dataset

Performance parameter	Value
Training RMSE	3.74
Test RMSE	4.86
Overall RMSE	4.80

4 Conclusions and recommendations

Two ANN models were built to predict the current and future WQI of Sungai Selangor, Malaysia. For the feedforward network model, it predicts the WQI by only using the six WQI parameters (pH, BOD, COD, AN, TSS and DO) as inputs and the actual WQI as outputs to learn the nonlinear correlations between the inputs and output and produce an

output that matches the actual WQI. Despite the largely unknown factors that is controlling the river water quality, the limited data size, and the incomplete dataset, an excellent result was able to be produced and the correlation between the predicted WQI and the actual WQI was found to be satisfactory. Regression of 0.97318, 0.99469, 0.98619 and 0.97874 respectively for training, test, validation, and overall dataset. RMSE value of 0.94 was obtained for validation dataset.

The NARX recurrent neural model on the other hand is to predict the future WQI with the current six WQI parameters as inputs and actual WQI as output for the model. The results show that even with a corrupted dataset, satisfactory results were still able to be generated by the network, which means the predicted WQI of the NARX recurrent neural model is a great match with the actual WQI and high correlation can be found between the actual WQI and predicted WQI. The results could be better if more quality datasets were provided. Regression value of 0.76295, 0.70941 and 0.75200 respectively are obtained for the training, test, and overall datasets.

In a conclusion, ANN modelling is a very useful and promising tool that utilises artificial intelligence to predict the current and future WQI with acceptable accuracies. The model can act as proof to show whether it is able to predict the WQI successfully. When successful the model can be used for data (operating parameters) from different rivers where in the predicted WQI can be obtained without determining the WQI experimentally. This in turn will reduce the time and cost for determining WQI of a new river and most importantly guide the stake holders to overcome the issue of water contamination. Also, stakeholders or utility managers can take preventive measures to avoid any disease outbreak due to water contamination. As future study, sensitivity analysis can be carried out to identify the input parameter which has maximum impact on the WQI. In further studies, new types of algorithms that are more appropriate for time series forecasting should be used in order to achieve a better result. Additionally, data that is more complete and up to date should be utilised instead of using incomplete data. This can eliminate the probability of the dataset affecting the model performance and narrow down the sources of possible errors to be able to pinpoint the errors with ease.

References

- Al-Badaii, F. and Shuhaimi-Othman, M. (2014) 'Heavy metals and water quality assessment using multivariate statistical techniques and water quality index of the Semenyih River, Peninsular Malaysia', *Iranica Journal of Energy & Environment*, Vol. 5, No. 2, pp.132–145.
- Bassi, N. and Kumar, M.D. (2017) 'Water quality index as a tool for wetland restoration', *Water Policy*, June, Vol. 19, No. 3, pp.390–403.
- Beale, M.H., Hagan, M.T. and Demuth, H.B. (2010) 'Neural network toolbox', *User's Guide, MathWorks*, Vol. 2, pp.77–81.
- Boretti, A. and Rosa, L. (2019) 'Reassessing the projections of the World Water Development Report', *NPJ Clean Water*, Vol. 2, No. 1, p.15.
- Burgan, H.İ., İçağa, Y., Bostanoğlu, Y. and Kilit, M. (2013) 'Water quality tendency of Akarçay River between 2006–2011.
- Chapagain, S.K., Mohan, G., Rimba, A.B., Payus, C., Sudarma, I.M. and Fukushima, K. (2022) 'Analyzing the relationship between water pollution and economic activity for a more effective pollution control policy in Bali Province, Indonesia', *Sustainable Environment Research*, Vol. 32, No. 5, pp.1–14.

- Dawood, T., Elwakil, E., Novoa, H.M. and Delgado, J.F.G. (2021) 'Toward urban sustainability and clean potable water: prediction of water quality via artificial neural networks', *Journal of Cleaner Production*, April, Vol. 291, p.125266.
- Dikmen, I., Birgonul, M.T. and Kiziltas, S. (2005) 'Prediction of organizational effectiveness in construction companies', *Journal of Construction Engineering and Management*, February, Vol. 131, No. 2, pp.52–261.
- Flörke, M., Kynast, E., Bärlund, I., Eisner, S., Wimmer, F. and Alcamo, J. (2013) 'Domestic and industrial water uses of the past 60 years as a mirror of socio-economic development: a global simulation study (2013)', *Global Environmental Change*, February, Vol. 23, No. 1, pp.144–156.
- Horne, B.G. and Giles, C.L. (1995) *An Experimental Comparison of Recurrent Neural Networks*, pp.697–704, MIT Press, Cambridge, MA.
- Kulisz, M., Kujawska, J., Przysucha, B. and Cel, W. (2021) 'Forecasting water quality index in groundwater using artificial neural network', *Energies*, Vol. 14, No. 18, p.5875.
- Luo, P., Kang, S., Apip, Zhou, M., Lyu, J., Aisyah, S., Binaya, M., Regmi, R.K. and Nover, D. (2019) 'Water quality trend assessment in Jakarta: a rapidly growing Asian megacity', *PLoS One*, July, Vol. 14, No. 7.
- Makarynska, D. and Makarynsky, O. (2008) 'Predicting sea-level variations at the Cocos (Keeling) Islands with artificial neural networks', *Computers & Geosciences*, Vol. 34, No. 12, pp.1910–1917.
- Muryani, E. (2021) 'Literature review: water quality and public health problems in developing countries', in *AIP Conference Proceedings*, November, Vol. 2363, No. 1, AIP Publishing.
- Najah, A., Teo, F.Y., Chow, M.F., Huang, Y.F., Latif, S.D., Abdullah, S., Ismail, M. and El-Shafie, A. (2021) 'Surface water quality status and prediction during movement control operation order under COVID-19 pandemic: case studies in Malaysia', *International Journal of Environmental Science and Technology*, Vol. 18, pp.1009–1018.
- Othman, F., Alaaeldin, M.E., Seyam, M., Ahmed, A.N., Teo, F.Y., Fai, C.M., Afan, H.A., Sherif, M., Sefelnasr, A. and El-Shafie, A. (2020) 'Efficient river water quality index prediction considering minimal number of inputs variables', *Engineering Applications of Computational Fluid Mechanics*, Vol. 14, No. 1, pp.751–763.
- Palani, S., Liong, S.-Y. and Tkalich, P. (2008) 'An ANN application for water quality forecasting', *Marine Pollution Bulletin*, Vol. 56, No. 9, pp.1586–1597.
- Sekitar, J.A. (2018) *Laporan Kualiti Alam Sekeliling 2018: Environmental Quality Report 2018*, Jabatan Alam Sekitar.
- Seyam, M. and Othman, F. (2015) 'Long-term variation analysis of a tropical river's annual streamflow regime over a 50-year period', *Theoretical and Applied Climatology*, July, Vol. 121, Nos. 1–2, pp.71–85.
- Siegelmann, H.T., Horne, B.G. and Giles, C.L. (1997) 'Computational capabilities of recurrent NARX neural networks', *IEEE Trans Syst. Man Cybern. B Cybern.*, April, Vol. 27, No. 2, p.208.
- Štambuk-Giljanović, N. (1999) 'Water quality evaluation by index in Dalmatia', *Water Research*, Vol. 33, No. 16, pp.3424–3440.
- Subramaniam, V. (2004) *Managing Water Supply in Selangor and Kuala Lumpur*, The Board of Engineers Malaysia.
- UNESCO World Water Assessment Programme (2003) *UN WWAP (2003)*, United Nations Educational, Scientific and Cultural Organization (UNESCO), Berghahn Books.
- Wu, M.L., Wang, Y.S., Sun, C.C., Wang, H., Dong, J.D., Yin, J.P. and Han, S.H. (2010) 'Identification of coastal water quality by statistical analysis methods in Daya Bay, South China Sea', *Marine Pollution Bulletin*, Vol. 60, No. 6, pp.852–860.
- Zikmund, W.G., Babin, B.J., Carr, J.C. and Griffin, M. (2013) *Business Research Methods*, Cengage Learning, South-Western, Australia.

Abbreviations

WQI	Water quality index
DO	Dissolved oxygen
BOD	Biochemical oxygen demand
COD	Chemical oxygen demand
pH	Potential of hydrogen
SS	Suspended solids
AN	Ammoniacal nitrogen
ANN	Artificial neural network
RMSE	Root mean squared error
GRNN	General regression neural network
MAE	Mean absolute error
WDN	Water distribution network
ELM	Extreme learning machine
AVP	Average validity percent
DID	Department of Irrigation and Drainage Malaysia
NARX	Nonlinear autoregressive models with exogenous inputs
MSE	Mean squared error
MAPE	Mean absolute percentage error
