❑   126

Estimation of water quality index using artificial intelligence approaches and multi-linear regression

Muhammad Sani Gaya[1], Sani Isah Abba[2], Aliyu Muhammad Abdu[3], Abubakar Ibrahim Tukur[4], Mubarak Auwal Saleh[5], Parvaneh Esmaili[6], Norhaliza Abdul Wahab[7]

[1,3,5]Department of Electrical Engineering, Kano University of Science &Technology, Wudil, Nigeria
[2]Department of PPD &M Department, Yusuf Maitama Sule University, Kano, Nigeria
[4]Department of Geography, Kano University of Science and Technology, Wudil, Nigeria
[6]Department of Electrical Engineering, Near East University, Lefkosa-North Cyprus, via Mersin 10 Turkey, Turkey
[7]Department of Control and Mechatronics, Universiti Teknologi Malaysia, Malaysia

Article Info

ABSTRACT

Water quality index is a measure of water quality at a certain location and over a period of time. High value indicates that the water is unsafe for drinking and inadequate in quality to meet the designated uses. Most of the classical models are unreliable producing unpromising forecasting results. This study presents Artificial Intelligence (AI) techniques and a Multi Linear Regression (MLR) as the classical linear model for estimating the Water Quality Index (WQI) of Palla station of Yamuna river, India. Full-scale data of the river were used in validating the models. Performance measures such as Mean Square Error (MSE), Root Mean Squared Error (RMSE) and Determination Coefficient (DC) were utilized in evaluating the accuracy and performance of the models. The obtained result depicted the superiority of AI models over the MLR model. The results also indicated that, the best model of both ANN and ANFIS proved high improvement in performance accuracy over MLR up to 10% in the verification phase. The difference between ANN and ANFIS accuracy is negligible due to a slight increment in performance accuracy indicating that both ANN and ANFIS could serve as reliable models for the estimation of WQI.

Corresponding Author:

Muhammad Sani Gaya,
Department of Electrical Engineering,
Kano University of Science and Technology,
Wudil Kano-Nigeria.
Email: muhdgayasani@gmail.com

1.   INTRODUCTION

The concern for water quality (WQ) is quite essential for health, water resources and environmental purposes [1]. The demand by billions of individuals for clean, safe and adequate freshwater on the planet enticed the practioners and research communities to be much engaged in modeling and monitoring of water quality and to address this universal concern [2]. WQ can be described as a physical, chemical and biological characteristics of water which can be used to predict the water quality that aid in determining the extent of water purity.

Water quality index (WQI) is applied worldwide to resolve the data management issues and assess success and failures in management strategies for improving WQ [3]. In order to determine the overall status of WQ the number of sensitive parameters need to critically be identify. Since no single variables can sufficiently assess the WQ, therefore, the WQ is generally assessed by computing the broad range of

parameters. As a result, large amount of data set is generated which requires to be presented in a meaningful way to decision makers, local planners and general public. In a view of this, WQI have been developed to convert the large data set in to a single index [4].

The reduction of water quality as a result of inadequate proper sanitation and pollutants coming from industries and the unreliability of most of the available mechanistic models in yielding promising forecasting results necessitated the vehement need for adopting others techniques and approaches [5]. Different methods have been used to measure and predict the quality of water in order to reduce the time consuming by collecting the data from the large data set and classify the quality using machine learning [3], but the main issues with machine learning method are high level of error susceptibility and acquisition relevant data set. Recently, a keen interest in studying the broad concept of artificial intelligence was developed, that communicate with the traditional model [6]. Despite several researchers such as [4-8] have used different neural network approaches in handling WQI. Nevertheless, most of the available models focus more on monitoring and analysis of water quality index. Therefore, this paper centres on estimating the water quality index through comparing the artificial intelligent approaches with conventional method applied to the Palla station along Yamuna River, India.

Artificial Neural Network (ANN) is an AI-based approach that not only proved to be effective in handling large amount of dataset, complex nonlinear input and output relationship but also flexible and powerful computational tool [5, 9-10]. Adaptive Neuro-Fuzzy Inference System (ANFIS) as another AI-based model has found to be successful tool which incorporate the approach of fuzzy Sugeno model that derived the benefit of both ANN and fuzzy logic in a single system [11].

The performances of the models were evaluated using commonly used measures. The paper is organized as follows: section 2 describes the research method and section 3 presents results and discussion while section 4 gives the conclusion.

2.   RESEARCH METHOD
2.1. Study area
The biggest tributary of River Ganga is Yamuna River, this river is as sacred and prominent as the immense River Ganga itself. As the holy river, Yamuna covered 1,376 km, almost 57 million residents of North part of India rely upon it. A total catchment area of Yamuna is 366,223 km^2 which comprises of 42 percent of the river ganga basin located in the territory of India. Delhi as capital territory received almost 70 percent of its drinking water from Yamuna River while discharges almost 10,000 m^3/s yearly. But due to urbanization and inadequate water treatment plant, the River leaves Delhi as polluted water [12-13]. Figure 1 shows the location of Palla station along Yamuna River basin in India. The daily WQ data were obtained from the CPCB for years 1999 to 2012.
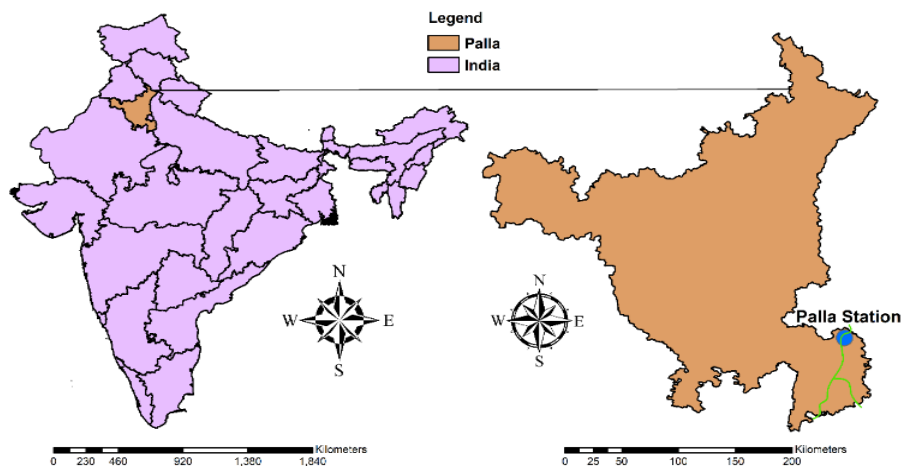

Figure 1. Geographic location of the study Palla Station, Yamuna River India

2.2. Modelling
In this study, ANN, ANFIS and MLR models were proposed for the estimation of WQI of the river, data set were partitioned into two parts, 70% of the data were employed for calibration phase and the 30% of

the data for verification purposes. Selection of dominant inputs parameters is one of the important parts in any AI based modeling. The functional expressions for the WQI are presented in (1-5). MATLAB 9.3 (R2017b) was used for the analysis of ANN and ANFIS while MLR model was developed using regression tool of EViews software 9.5 version.

$$WQI_1 = \psi(DO) \tag{1}$$

$$WQI_2 = \psi(DO, pH) \tag{2}$$

$$WQI_3 = \psi(DO, pH, BOD) \tag{3}$$

$$WQI_4 = \psi(DO, pH, BOD, NH_4) \tag{4}$$

$$WQI_5 = \psi(DO, pH, BOD, NH_5, WT) \tag{5}$$

where WQI_n depicts the water quality index, ψ is the function of Dissolved oxygen (DO), pH, Biological Oxygen Demand (BOD), Ammonium nitrate (NH4), and Water temperature (WT).

2.3. Multilinear regression analysis (MLR)

Multi-linear regression analysis is the model applied based linear relationship between the dependent variable and independent variable. MLR is based on the concept of least squares, which is the value of the estimated parameter is expressed as a linear function [14]. As it is stated in (6).

$$y = b_0 + b_1 x_1 + b_2 x_2 + ...b_i x_i \tag{6}$$

where x_1, is the value of the ith predictor, b_0 is the regression constant, and b_i is the coefficient of the ith predictor.

2.4. Artificial neural network (ANN)

ANNs are mathematical model aims to handle non-linear relationship of input – output dataset. Historically, are information processing tools derived from analogy with biological nervous system of brain. ANN has proved to be an effective tool in predicting nonlinear systems and quite capable of handling complex noisy data set [15-16], the prediction accuracy of ANN is high [17]. Back propagation (BP) algorithm is the most common used technique among the classification of ANN. In BP, each input training data flows via the system and passes to the output layer, the error of the training is generated and propagates backward until the desired target of the network is achieved [18]. The primary aim of BPNN is to reduce the error in order for the network to learn the training data. Sigmoid and the Lavenberg-Marquardt (LM) were used as activation function and algorithm, respectively. LM used in training MLP model because of its outstanding performance [19]. Before model training at the initial stage, the data for both input and output were normalized within a scale of 0 and 1 using the as: Figure 2 shows the structure of ANN.

$$X_i = \frac{x_i - x_{min}}{x_{max} - x_{min}} \tag{7}$$

Where X_i is the normalized quantity, x_i is un-normalized quantity, x_{min} is the minimum and x_{max} is the maximum quantity of the data set
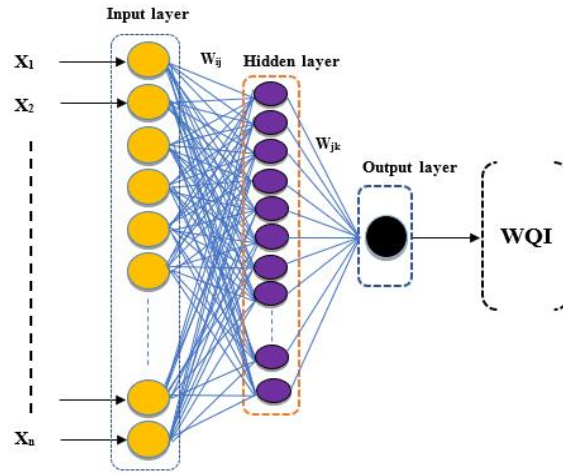
Figure 2. Structure of ANN that map a given inputs and output

2.5. Adaptive neuro-fuzzy inference system (ANFIS)

The combination of artificial neural network with the fuzzy system creates a robust hybrid system that is able to solve a complex nature of relationship. ANFIS as one of the AI models has the ability to overcome the limitations of fuzzy inference and ANN. ANFIS model combine the ability of ANN and Fuzzy logic to create a process that has the ability of handling complex non-linear interactions between a set of input and output [20]. The general structure of ANFIS is shown in Figure 3. For a typical ANFIS, assuming the FIS that contains two inputs 'x' and 'y' and one output 'f', a first order Sugeno fuzzy has following rule:

$$Rule(1) : if \; \mu(x) \; is \; A_1 \; and \; \mu(y) \; is \; B_1 \; then \; f_1 = p_1 x + q_1 y + r_1$$

$$Rule(2): if \; \mu(x) \; is \; A_2 \; and \; \mu(y) \; is \; B_2 \; then \; f_2 = p_2 x + q_2 y + r_2$$

Membership functions parameters for x and y inputs are A_1, B_1, A_2, B_2, outlet functions' parameters are $p_1, q_1, r_1, p_2, q_2, r_2$, a five-layer neural network arrangement followed the formulation and structure of ANFIS. For more explanation of ANFIS, refer to the study in [6].
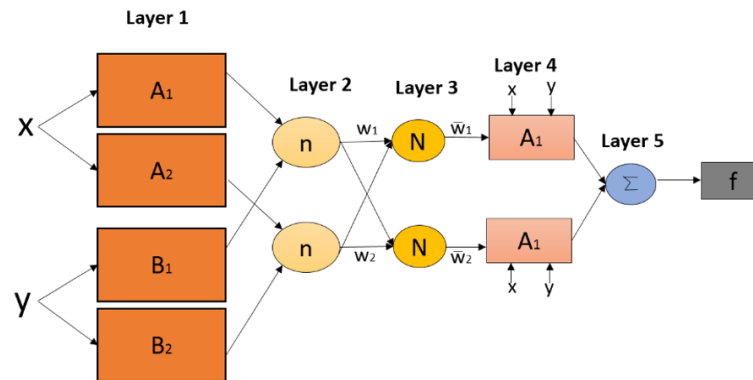

Figure 3. Geographic location of the study Palla Station, Yamuna River India

2.6. Performance evaluation criteria

The performance efficiency of the model can be assessed through different statistical measures, including Determination Coefficient (DC), Root Mean Square Error (RMSE), Mean Square Error (MSE) etc. Therefore, in order to evaluate the performance of ANN, ANFIS and MLR models, DC, RMSE and MSE were employed in this study [21]. The equation of DC and RMSE are given as:

$$DC = 1 - \frac{\sum_{i=1}^{N}\left(WQI_{oi} - WQI_{pi}\right)^2}{\sum_{i=1}^{N}\left(WQI_{oi} - \overline{WQI_{oi}}\right)^2} \tag{8}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}\left(WQI_{oi} - WQI_{pi}\right)^2}{N}} \tag{9}$$

$$MSE = \frac{1}{N}\sum_{i=1}^{N}\left(WQI_{oi} - WQI_{pi}\right) \tag{10}$$

where N, WQI_{oi}, $\overline{WQI_{oi}}$, WQI_{pi} are data number, observed data, average value of the observed data and computed values, respectively. DC ranges between and $-\infty$ and 1 with a perfect score of 1.

3.   RESULTS AND ANALYSIS

In this paper, MLR, ANN and ANFIS were used to estimate the WQ at Pala station in Yamuna River, and their individual performance accuracy were compared. For all these models, MATLAB 9.3 (R2017b) software was used for the analysis of ANN and ANFIS while MLR model was developed using regression tool of EViews software 9.5 version. For the estimation of river parameters, different input parameters have been employed, as appropriate input selection is essential [22]. Pearson and Spearman correlation analysis methods were performed to choose the inputs parameters. Five different models and input combinations were trained based on the number and types of input, for all the methods the model types were defined as MLR-I up to MLR-V, ANN-I up to ANN-V and ANFIS-I up to ANFIS-V indicating the type of models from one to five for MLR, FFNN, and ANFIS, respectively.

3.1. Result of MLR model

MLR model was applied as the classical conventional method for modeling the linear interactions of the system. It is often used as the reference comparison model with non-linear models. The equation (11) was obtained for the best model to estimate the performance of WQI, From Table 1, it indicates that the best performing mode was MLR-V which has a total of 5 input variables, the results indicate that the MLR model is best with the highest number of input variables.

$$WQI = 0.8670 + 0.031DO + 0.095pH + 0.3381BOD - 0.3961NH_4 - 0.2471WT \tag{11}$$

The negative values in the estimation serve no purpose in the modeling of WQI. As shown in the Table 1, the MLR performance was satisfactory for the prediction of WQI at Palla. This is proved by the value of MSE=0.00131, DC=0.8919 and RMSE=0.03625 in the verification phase. Figure 4 present the scatter and times series plots for measured and estimated WQI values for MLR model in a verification phase. The measure and estimated values were well superposed and the discrepancies between the measured and estimated values were small which indicate high prediction accuracy.

Table 1. MLR estimation results

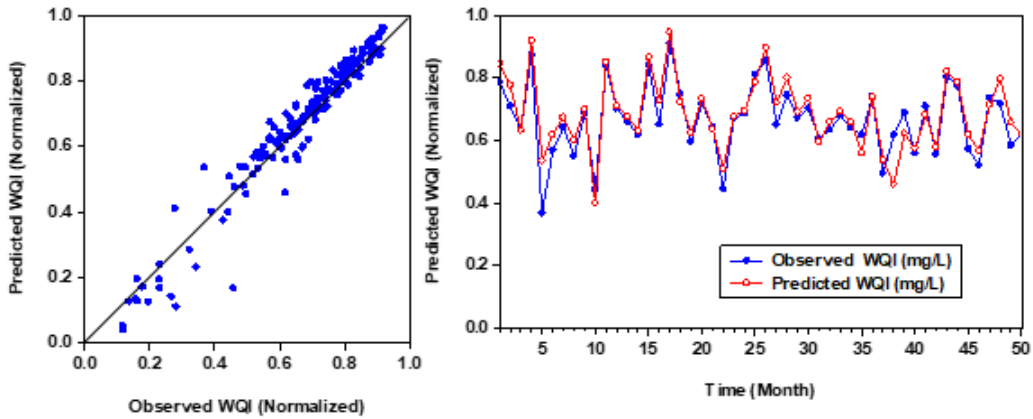| Stations	Model	Input Variables	Calibration			Validation		
			MSE	DC	RMSE	MSE	DC	RMSE
Palla	MLR-I	DO	0.00026	0.8970	0.0161	0.00143	0.8819	0.0378
	MLR-II	DO, pH	0.00028	0.9071	0.0167	0.00135	0.8886	0.0367
	MLR-III	DO, pH, BOD	0.0003	0.8009	0.0172	0.0004	0.8651	0.02
	MLR-IV	DO, pH, BOD, NH4,	0.00031	0.8912	0.0177	0.00133	0.8912	0.0365
	MLR-V	DO, pH, BOD, NH4, WT	0.00029	0.8960	0.017	0.00131	0.8919	0.03625

Figure 4. Evaluation of observed versus predicted scatter and time series plot of MLR model

3.2. Result of ANN model

In ANN-feed forward was trained by the algorithm called Lavenberg-Marquartd. ANN was trained with a sigmoid activation function which is non-linear exponential function. It's paramount importance to make an appropriate selection of a hidden neurons and architecture of the network in order to prevent over-learning in the calibration stage. The result of ANN model is presented in Table 2. The prediction accuracy of ANN was superior than MLR model, the best model to estimate WQI was obtained to be ANN-II with the values of MSE, DC and RMSE are 9.0E-8, 0.9974 and 0.0003, respectively as shown in Table 2. Figure 5, shows the scatter and time series for measured and estimated WQI values for ANN model in a verification phase. From the comparison of Figure 4-5 it is clear that ANN are more fitted and the accuracy proved high merit over MLR model. This can also be justified by MSE between ANN and MLR models. The robustness of ANN could be attributed to the great advantage of ANN to handle complex and nonlinear system, unlike the MLR models which is base on the assumptions of linear input - output relationship.

Table 2. ANN estimation results

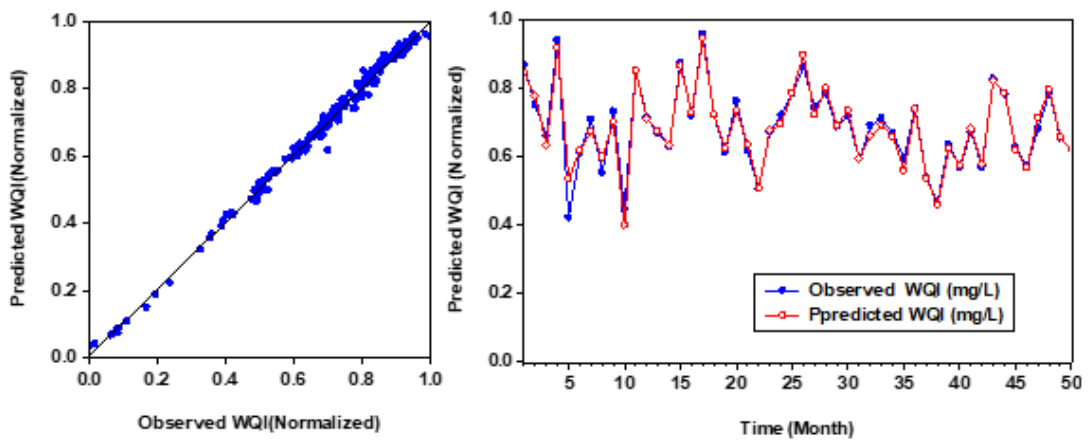| Station	Model	Model Structure	Calibration			Validation		
			MSE	DC	RMSE	MSE	DC	RMSE
	ANN-I	(1 - 1 - 1)	0.0000036	0.9957	0.0006	0.00000009	0.9946	0.00033
	ANN-II	(2 - 2- 1)	0.00000121	0.9976	0.0011	0.00000009	0.9974	0.0003
Palla	ANN-II	(3 - 3 - 1)	0.00000676	0.9951	0.0026	0.00000009	0.9975	0.0003
	ANN-IV	(4 - 4- 1)	0.00001089	0.994	0.0033	0.00000036	0.9954	0.0006
	ANN-V	(5 - 6 - 1)	0.00001024	0.9573	0.0032	0.00001156	0.9921	0.0034

Figure 5. Evaluation of observed versus predicted scatter and time series plot of ANN model

3.2. Result of ANFIS model

ANFIS as a hybrid algorithm was employed with a suitable inference system called Takgi-Sugeno-Kang which worked based on several rules and membership function. In this study the ANFIS model consist of five input variables and one output variable in order to estimate the WQI at Palla station. ANFIS were trained base on five different models in each station, the triangular and gaussian membership function were tried to find the best model. For the purpose of this research, 5, (2, trimf, constant) indicates that, a model with 5 input variables, 2 triangular membership function input and constant output. Table 3 indicates that, the value of MSE, RMSE and DC are 8.41E-6, 0.0029 and 0.9909, respectively. ANFIS II was obtained to be the best model with two input combinations as shown in Table 3. Despite the superiority of ANN model over ANFIS model the performance accuracy of ANFIS model proved to be reliable in estimation of WQI of Palla station. Figure 6 shows scatter and time series plots for the measured and estimated WQI values for ANN model in a verification phase. From the Figure 6 it is clear that the ANFIS estimates were closer to the observed WQI value than MLR value.

However, by comparing Figures 4-6 for MLR, ANN and ANFIS modes it is clear from the Figures, that ANN and ANFIS best model proved high improvement in performance accuracy over MLR up to 10% in the verification phase. The difference between ANN and ANFIS accuracy is negligible indicating that both the models outperformed MLR model interm of estimation accuracy. The results can also be justified by presenting the box plot of the best three models, in order to demonstrate how closely the estimations models are with the observed values, as illustrated in Figure 7. According to the Figure it is appear that ANN and ANFIS prediction values resembled the observed values. In comparing with MLR and ANFIS, ANN had the best fitting because the closer the data point to be best line of fit the better the predictions (see the scatter plots). Hence, ANN can be used as reliable and superior to ANFIS and MLR for the estimation of WQI.

Table 3. ANFIS estimation results

| Station	Model	Model Structure	Calibration			Validation		
			MSE	DC	RMSE	MSE	DC	RMSE
Palla	ANFSI-I	5, (2, trimf, Constant)	0.00000961	0.9078	0.0031	0.00001764	0.9179	0.0042
	ANFIS-II	4, (2, trimf, Constant)	0.00000676	0.9929	0.0026	0.00000841	0.9900	0.0029
	ANFIS-II	3, (2, trimf, Constant)	0.00001521	0.9906	0.0039	0.00000961	0.9807	0.0031
	ANFIS-IV	2, (2, trimf, Constant)	0.00046225	0.9287	0.0215	0.00081225	0.9469	0.0285
	ANFIS-V	2, (2, trimf, Constant)	0.00001225	0.9557	0.0035	0.00001369	0.9384	0.0037
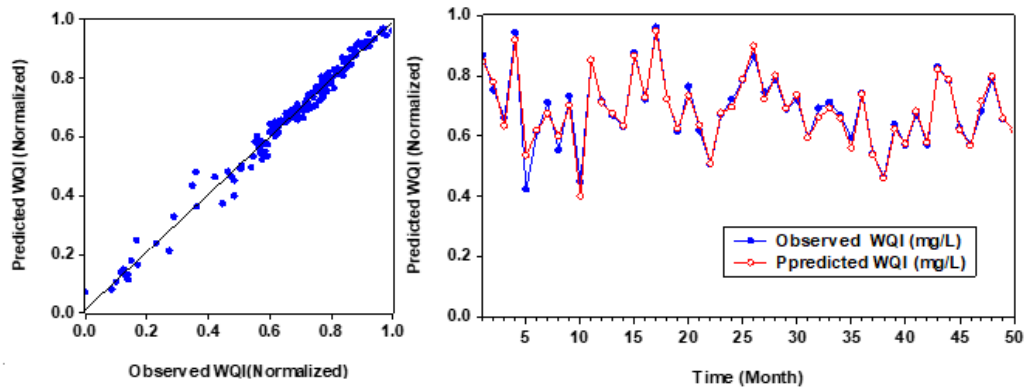
Figure 6. Evaluation of observed versus predicted scatter and time series plot of ANFIS model

It can be seen from the Figure 7 that the ANN and ANFIS model outperformed the MLR model, rhe predictions of the intelligent models are extremely closer to the observed WQI. The intelligent models were highly accurate in estimating the WQI by achieving the DC close to unity.
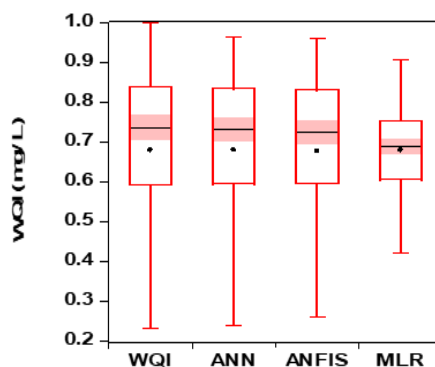
Figure 7. Comparison of Box plots of the observed WQI and predicted of best models for ANN,
ANF and MLR

4. CONCLUSION

The paper has presented MLR, ANN, and ANFIS models for estimation of Water Quality Index (WQI), with the water quality variables as inputs. The obtained results indicated that the artificial intelligence-based models (ANN and ANFIS) outperformed conventional model (MLR) up to 10% in the verification phase. The AI models were able to accurately follow the trajectories of the observed water quality index. Although the performance of ANN is slightly better than the ANFIS, but the ANN and ANFIS models outperformed MLR model in estimating the WQI. ANN and ANFIS models are more reliable in the estimation of WQI at Palla station of Yamuna River India. In order to increase the accuracy and uncertainties problems of the models and to explore the contribution of each input combinations, further research should be carried out by employing more AI based models in estimation of WQI. The intelligent models could serve as reliable and useful tools in estimating the water quality index of the river.

REFERENCES

[1]   Sharma, D., and Kansal, A. (2011). Water quality analysis of River Yamuna using water quality index in the national capital territory , India (2000 – 2009), *0*, 147–157. https://doi.org/10.1007/s13201-011-0011-4
[2]   Elkiran, G. (2004). The Water Budget Analysis of Girne Region , North Cyprus, (October), 6–8.
[3]   Muhammad, S. Y. et al., (2015). Classification Model for Water Quality using Machine Learning Techniques, *9*(6), 45–52.
[4]   Gazzaz, N. M. et al., (2012). Artificial neural network modeling of the water quality index for Kinta River (Malaysia) using water quality variables as predictors. *Marine Pollution Bulletin*, *64*(11), 2409–2420. https://doi.org/10.1016/j.marpolbul.2012.08.005
[5]   Singh, K. P. et al., (2009). Artificial neural network modeling of the river water quality — A case study, *220*, 888–895. https://doi.org/10.1016/j.ecolmodel.2009.01.004.
[6]   Nourani, V. et al., (2013). Application of the Artificial Neural Network to monitor the quality of treated water, *3*(1), 38–45.
[7]   Wang, W.C. et al., (2015). Improving forecasting accuracy of medium and long-term runoff using artificial neural network based on EEMD decomposition. *Environmental Research*, *139*(October 2017), 46–54. https://doi.org/10.1016/j.envres.2015.02.002
[8]   Khan, Y. and Chai, S. S. (2016). Ensemble of ANN and ANFIS for Water Quality Prediction and Analysis - A Data Driven Approach, *9*(2), 117–122.
[9]   Nourani, V. et al., (2011). Two hybrid Artificial Intelligence approaches for modeling rainfall – runoff process. *Journal of Hydrology*, *402*(1–2), 41–59. https://doi.org/10.1016/j.jhydrol.2011.03.002
[10] Abba, S. I. and Elkiran, G. (2017). Effluent prediction of chemical oxygen demand from the astewater treatment plant using artificial neural network application. In *Procedia Computer Science* (Vol. 120). https://doi.org/10.1016/j.procs.2017.11.223
[11] Yetilmezsoy, K. et al., (2011). Artificial Intelligence -Based Prediction Models, 193–218.
[12] Abba, S. I. et al., (2017). River water modelling prediction using multi-linear regression, artificial neural network,

and adaptive neuro-fuzzy inference system techniques. *Procedia Computer Science*, *120*, 75–82. https://doi.org/10.1016/j.procs.2017.11.212

[13] Parmar, K. S. and Bhardwaj, R. (2015). River Water Prediction Modeling Using Neural Networks, Fuzzy and Wavelet Coupled Model. *Water Resources Management*, *29*(1), 17–33. https://doi.org/10.1007/s11269-014-0824-7

[14] Abdullahi, J. and Elkiran, G. (2017). Prediction of the future impact of climate change on reference evapotranspiration in Cyprus using artificial neural network. *Procedia Computer Science*, *120*, 276–283. https://doi.org/10.1016/j.procs.2017.11.239

[15] Gaya, M. S. et al., (2017). Estimation of turbidity in water treatment plant using hammerstein-wiener and neural network technique. *Indonesian Journal of Electrical Engineering and Computer Science*, *5*(3), 666–672. https://doi.org/10.11591/ijeecs.v5.i3.pp666-672

[16] Yasmin, N.S.A. et al., (2017). Estimation of pH and MLSS using neural network. *TELKOMNIKA*, Vol.15, No.2, June 2017, pp. 912~918. http://doi.org/ 10.12928/TELKOMNIKA.v15i2.6142

[17] Yusuf, Z. et al., (2018). Neural network model development with soft computing techniques for membrane filtration process. *International Journal of Electrical and Computer Engineering (IJECE), Vol. 8, No. 4, August 2018, pp. 2614~2623*. http://doi.org/ *10.11591/ijece.v8i4.pp2614-2623*

[18] Nourani, V. et al., (2015). Wavelet-entropy data pre-processing approach for ANN-based groundwater level modeling. *JOURNAL OF HYDROLOGY*, *524*, 255–269. https://doi.org/10.1016/j.jhydrol.2015.02.048

[19] Committee, A. T. (2000). Artificial neural networks in hydrology I: Preliminary concepts. *Journal of Hydrologic Engineering*, *5*(2), 115–123.

[20] Mirbagheri, S. A. et al., (2010). Neuro-fuzzy models employing wavelet analysis for suspended sediment concentration prediction in rivers. *Hydrological Sciences Journal*, *55*(7), 1175–1189. https://doi.org/10.1080/02626667.2010.508871

[21] Legates, D. R. and McCabe Jr., G. J. (2005). Evaluating the Use of "Goodness of Fit" Measures in Hydrologic and Hydroclimatic Model Validation. *Water Resources Research*, *35*(1), 233–241. https://doi.org/10.1029/1998WR900018

[22] Emamgholizadeh, S. et al., (2014). Prediction of water quality parameters of Karoon River (Iran) by artificial intelligence-based models. *International Journal of Environmental Science and Technology*, *11*(3), 645–656. https://doi.org/10.1007/s13762-013-0378-x

BIOGRAPHIES OF AUTHORS

Muhammad Sani Gaya received B.Eng in Electrical Engineering in 2002, M.Eng (Electrical Engineering.) in 2008 both from Bayero University Kano, Nigeria and PhD degree from the Department of Control and Mechatronics Engineering, Universiti Teknologi Malaysia in 2014. His current research interest centres around system identification, feedback control, IMC and inverse control of uncertain/ nonlinear systems based on neural network, fuzzy and ANFIS. Currently, working with Department of Electrical Engineering, Kano University of Science and Technology, Wudil.

Sani Isah Abba received his bachelor degree from Bayero University, Kano (BUK), master's degree from Sharda University, India, and Ph.D. degree from Near East University, TRNC. His current research interest includes water resources, environmental modeling and simulation, climatology, water quality and water and wastewater treatment plant. He also has additional skills and expertise in data-driven algorithms, Artificial intelligence and data processing software.

Aliyu Muhammad Abdu received his B.Eng. in Electrical Engineering from Kano University of Science and Technology (KUST) in 2009 and his M. Eng. degree in Electrical Engineering from Universiti Technologi Malaysia (UTM) in 2014. His research interest includes Digital Image Processing, Machine Learning, Computer Vision and Pattern Recognition. His current research is in the application of these models to plant disease detection. Aliyu is currently a Ph. D student, member of the Computer Vision, Video, and Image Processing (CVVIP) research group of the department of Microcomputer and Electronics department at School of Electrical Engineering UTM.