International Journal of Recent Engineering Science (IJRES) Volume 6 Issue 2, 14-19, Mar-Apr 2019
ISSN: 2349 – 7157 / doi: 10.1445/23497157/IJRES-V6I2P103

Application of data-driven algorithms for the forecasting of non-linear parameter

S.I. Abba*, A.S. Maihula[1], M.B. Jibril[2], A.M. Sunusi[3], M.A. Ahmad[4], M. A. Saleh[2]

*[,3] *Civil Engineering Unit, Department of P.P.D. and M, Yusuf Maitama Sule University, Kano. Nigeria*
[1]*Department of Civil Engineering, Kano University of Science and Technology, Wudil, Kano. Nigeria*
[2]*Department of Electrical Engineering, Kano University of Science and Technology, Wudil, Kano. Nigeria*
[4]*Department of Civil Engineering, Ministry of Works, Housing & Transport, Kano State, Nigeria*

Abstract

Water quality index (W.Q.I.) is a widely used tool in different parts of the world to solve the problems of data management and to evaluate success and failures in management strategies for improving water quality. This study aimed to develop two Non-linear models (i.e., Adaptive neuro-fuzzy inference system (ANFIS) and Artificial neural network (ANN)) and a conventional linear model (viz: Autoregressive integrated moving average (ARIMA) models, in modelling the non-linear W.Q.I. at Kinta River, Malaysia and Yamuna River, India (Agra station). The performance of the models was assessed through Mean Square Error (M.S.E.), Root Mean Square Error (RMSE), Determination Coefficient (R^2), and Mean Absolute Percentage Error (MAPE). The obtained result depicted the non-linear models (ANF5S and ANN) can averagely increase the performance accuracy of linear models (ARIMA) up to 25% and 18% at Kinta and Yamuna River, respectively, in the verification phase. The overall results also demonstrated that the ANFIS model outperformed the other models, with the average increased up to 23% in the verification phase. Hence, serve as a suitable and reliable tool in forecasting the W.Q.I. in both of the regions.

Keywords - *ANFIS, ANN, ARIMA, Kinta River, W.Q.I., Yamuna River*

I.INTRODUCTION

The concept of water quality is fundamental to the study of environmental engineering and water resources because they explore the relationship between water requirements and the form and extent of permissible departure from purity [1]. Because of the water scarcity and long-term water imbalance, more than one billion individuals required clean, safe, and adequate fresh water on the planet; as a result, modelling and monitoring water quality and quantity will help to overcome this universal concern [2]. Water quality (W.Q.) can be described as the physical, chemical, and biological characteristics of water, which can predict the water quality that helps determine the extent of water purity. Water quality index (W.Q.I.) is a widely used tool in different parts of the world to solve the problems of data management and to evaluate success and failures in management strategies for improving water quality [3]. Lack of proper sanitation and activities from different industries have polluted the water due to the availability of natural water resources, results in the reduction of water quality and necessitates the desire need for adopting other techniques and approaches [4]. However, different traditional methods have been used to measure and predict the quality of water to reduce the time consuming by collecting the data from the large data set and classify the quality using machine learning. Recently, a keen interest in studying the broad concept of artificial intelligence was developed that communicate with the traditional model [1,5]. Although several researchers have used a different approach in predicting the W.Q. by using assigned weight, there is a need to employ robustness models' techniques to improves model diversity.

Non-linear artificial intelligence (A.I.) such as Artificial Neural Network (ANN) and Adaptive Neuro-fuzzy Inference System (ANFIS) are approaches that not only proved to be effective in handling a large amount of dataset, complex non-linear input and output relationship but also flexible and powerful computational tool [6, 7, 8]. Several researchers, such as [9, 10, 11], have conducted some in different linear and non-linear models, and the obtained results indicate the superior performance of non-linear over a linear model. Hence, the present work proposed applying linear (ARIMA) and non-linear models (ANFIS and ANN) in forecasting the non-linear W.Q.I. at two different regions.

II. MATERIALS AND METHODS

A. Case Study, Proposed Methodology and Data

Kinta River is located at Kinta district between the latitude and longitude of N 040 07' 102' and N 040 40' 115' and S1010 01' 284' and S 1010 09' 400' respectively. It has been used for farming, industries, and domestic affairs with a total length of about 100Km and 2500 Km^2 area. The river was divided into several subdivisions both along the upstream, undulating, and downstream Stations

2PK22 and 2PK24 are placed at upstream, while stations 2PK25 and 2PK34 are at the centre of the river, and stations 2PK19 and 2PK33 are at the downstream of the river [2]. On the other hand, the major side stream and the influent branch of Ganga River are attributed to the Yamuna River, with the catchment covering a length of about 1,376 km and covers approximately about 57 million demand of north Indian inhabitants. The river has been generally believed as the holy place where tradition, culture, and pilgrimage within the research area and across India are conducted. The total area of Yamuna catchment is 3,66,223 $km^{2,}$ and its annual discharge is nearly 10,000 m^3/s, which approximately provides 70% of water to Delhi for domestic and other activities. [8]. Fig. 1 shows the Kinta and Yamuna River (Agra station) for two different regions.

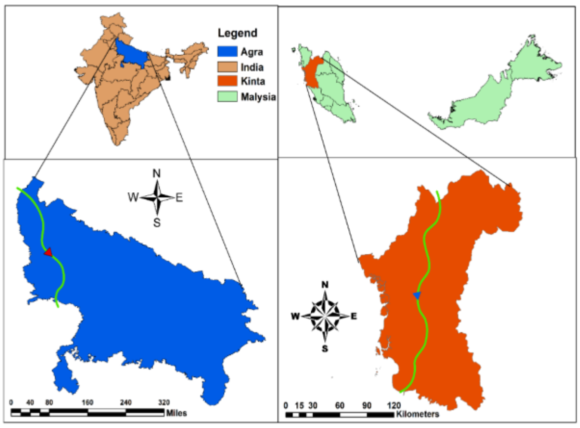

Fig. 1: Geographic location of the two regions (Kinta and Yamuna River)

In this study, non-linear AI-based models viz: ANN and ANFIS and a traditional linear model ARIMA were proposed for forecasting the W.Q.I. in a river. The science of data and descriptive statistical analysis is crucial in any AI-based model to recognize the relationship and the degree of strength between the variables. The normalization of the data was conducted before the model calibration, which is usually performed to increase the accuracy and speed of the model (see, Equation. 1) [12]. The available monthly W.Q. data set originated from the Department of Environment (D.O.E.) Malaysia and Central Pollution Control Board (CPCB) for Kinta and Yamuna River, respectively, includes measured D.O. (mg/L), B.O.D. (mg/L), pH, Ammonia (NH_3) (mg/L), W.T. (°C) and Water Quality Index (W.Q.I.). (Appendix A-Table 1) shows the regional statistical analysis of the data. 70% of the data were used for model calibration. At the same time, the remaining 30% were employed in the verification phase from 168 records. The suitable combinations of model input are selected using correlation analysis (see Appendix B-Table 2). Equations 2 are employed for the development of the models.

$$y = 0.5 + \left(0.5\left(\frac{x-\bar{x}}{x_{max}+x_{min}}\right)\right) \qquad (1)$$

Where y is the normalized data, x is the measured data, $\bar{x}$ is the mean of the measured data, x_{max} is the maximum value of the measured data, and xmin is the minimum value.

$$WQI = f(x)\begin{cases} BOD, NH_4 \\ BOD, NH_4, WT \\ BOD, NH_4, WT, DO, pH \end{cases} \qquad (2)$$

Where W.Q.I. is the water quality, *f(x)* is the function of model combinations.

B.Auto regressive integrated moving average (ARIMA)

ARIMA model is one of the most widely used classic models for time series forecasting and provides a consistent approach to the problems of forecasting. The major drawback of ARIMA is the pre-assumption of the linear model. The time-series data must be checked to be stationary in the model identification stage as it is essential in creating an ARIMA model [22].

$$y_t = \theta_0 + \emptyset_1 y_{t-1} + \emptyset_2 y_{t-2} + \cdots + \emptyset_p y_{t-p} + \alpha_t - \emptyset_1 \alpha_{t-1} - \emptyset_2 \alpha_{t-2} \ldots - \emptyset_q \alpha_{t-q} \qquad (3)$$

Where y_t is the observed value, α_t are error value, $\emptyset_i$ (i= 1, 2,…,p) and θ_j (j = 0,1,2, ... , q) are model parameters and p and q are called orders of the model. However, the three iterative stages used in ARIMA modelling are; estimation of the parameter, diagnostic checking, and identifying model.

C.Artificial Neural Network (ANN)

ANN is classified under the network structure functions and learning algorithm. ANN is a model designed to resembles the brain by learning and information processing based on a mathematical model [9]. Among different classes of ANN, FFNN with Back Propagation (B.P.) algorithm is the most common and widely used technique. BBNN, which is known as a multi-layer feed-forward neural network (FFNN). BPNN is most commonly accepted as a three-layer trained normally with Lavenberg - Marquardt algorithm, Fig. 2, which shows the architecture of the BPNN. The foremost concept of BPNN is that the weight is adjusted through the mean square error of the output until the error is minimized so that the network can learn the training data [12].

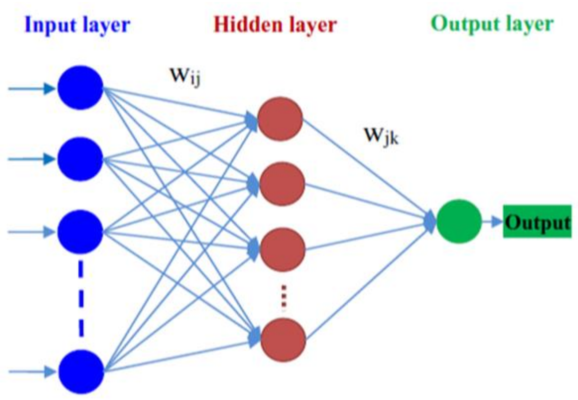

Fig. 2 Architecture of BPNN [21]

D. Adaptive Neuro-fuzzy Inference System (ANFIS)

ANFIS model combines the learning algorithms of both ANNs and Fuzzy logic to overcome the problems of complex non-linear between input-output variables [11]. The hybrid nature of ANFIS serves as an important tool in the simulation process, such as environmental prediction due to its hybrid nature. The combination of least square and gradient descent is used in ANFIS to optimize the network [14].

Assuming '*x*' '*y*' is the input and '*f*' is the output of a fuzzy inference system, the first order Sugeno type is the following rules (Eqs. 3 and 4).

Rule (1): if $\mu(x)$ is A_1 and $\mu(y)$ is B_1; then $f_1 = p_1x + q_1y + r_1$ (4)

Rule (2): if $\mu(x)$ is A_2 and $\mu(y)$ is B_2; then $f_2 = p_2x + q_2y + r_2$ (5)

where p_1, q_1, r_1 and p_2, q_2, r_2, are the parameters in the sub-sections of the model, and A_i and B_i are the linguistic labels obtained by a membership function. The full explanation of ANFIS can be found in [15].

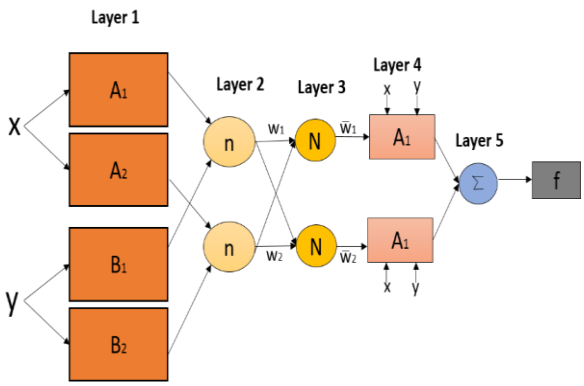

Fig. 3. The general structure of ANFIS [7]

E. Evaluation Criteria

The Evaluation Criteria of the model can be assessed through different statistical measures, including Determination Coefficient (R^2), Root Mean Square Error (RMSE), Mean Square Error (M.S.E.), Mean Absolute Percentage Error (MAPE), Mean Absolute Error (M.A.E.), Agreement Index (d), Standard Error of Prediction (S.E.P.), etc. But generally, the model efficiency performance should include at least one goodness-of-fit (e.g., R^2) and at least one absolute error measure (e.g., RMSE) [16]. Hence, R^2, RMSE, MAPE, and M.S.E. were employed in this study.

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(WQI_{obsi} - WQI_{comi})^2}{\sum_{i=1}^{N}(WQI_{obsi} - \overline{WQI}_{obsi})^2} \quad (6)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(WQI_{obsi} - WQI_{comi})^2}{N}} \quad (7)$$

$$MAPE = \frac{1}{N}\left[\sum_{i=1}^{N}\left|\frac{WQI_{obsi} - WQI_{comi}}{WQI_{obsi}}\right|\right] \quad (8)$$

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(WQI_{obsi} - WQI_{comi})^2 \quad (9)$$

Where N WQ_{obsi} $\overline{WQ}$ and WQ_{comi} *are* data number, observed data, the average value of the observed data, and computed values, respectively. R^2 ranges between and -∞ and 1, with a perfect score of 1.

III. Results and Discussion

For all these models, MATLAB 9.3 (R2017a) software was used for the analysis of ANFIS and ANN model. In contrast, the classical ARIMA model was developed using the regression tool of E-Views software 9.5 version. For the simulation of non-linear W.Q.I. parameters, different input parameters have been employed as appropriate input selection is essential in any modelling [17].

Results of Non-linear and linear models

Before the model calibration phase for the ANFIS and ANN non-linear models, three combinations, as shown in equation 2, are considered the core input of all the models. The performance efficiency for both regions is presented in Table 3. From the table, it can be seen that the model ANFIS-III (5, trimf, 2) for Kinta and Agra station with all the input parameters is found to be the best models in forecasting the W.Q.I. both for calibration and verification phase. This model (ANFIS-III (5, trimf, 2) indicates that a model with five input variables, trimf stands for triangular membership function and two as several membership functions. The best performance models for Kinta and Agra (ANFIS-III (5, trimf, 2) increased the average performance up to 30% and 38% concerning ANFIS-I and ANFIS-II, respectively. Likewise, for the ANN model, the best performing model was found to be ANN-III (5, 6, 1) and ANN-III (5, 8, 1) for Kinta and Agra, respectively (see Table 3). According to the table, it can be demonstrated that ANN-III (5, 6, 1) and ANN-III (5, 8, 1) models increased the average performance up to 24% and 11% for Kinta and Agra with regards to ANN-I and ANN-II in the verification phase. Table 3 also justified that the addition of D.O. and pH values increased the performance accuracy significantly for the best models for both the regions.

The comparison results indicate the advantage of the ANFIS model ANN for both Kinta and Agra. Generally, the best accuracy of ANFIS-III can be confirmed by the least values of RMSE, M.S.E., and MAPE and the high value of R^2 for both the two regions. However, Fig. 4(a and b) shows the scatter and time series plot of the observed versus computed three models in the verification phase for Kinta and Agra. From the graph, it is clear that ANFIS was more fitted and closer to the observed W.Q.I. than ANN and ARIMA models; this indicates that the ANFIS model outperformed ANN in terms of forecasting accuracy. Also, the result demonstrated that the models forecasted an outstanding accuracy for Kinta River in compare with the Yamuna River; this is attributed to the fact that Agra station received a huge amount of polluted river from Delhi while the considered stations for Kinta River are located at the upstream of the river, these outcomes are in line with the finding of [11]. A conventional ARIMA model was applied to modelling the linear interactions of the system. These models are often used as the reference comparison model with a non-linear model when combine. In the ARIMA model, the optimal parameters were chosen using different trial and error procedures, as in the other non-linear models. The development and the input combinations were carried out similarly to that of non-linear models. Table 3 shows clearly that, ARIMA-III model with five input combinations emerged to be in terms of performance criteria. According to Table 3, ARIMA-III increased the satisfactory average performance both for Kinta and Agra, respectively, with regards to ARIMA-I and ARIMA-II in the verification phase; this is due to the addition of significant parameters.

Among the linear and non-linear models, ANFIS revealed the optimal performance in forecasting the W.Q.I. for both regions. Table 3 confirmed that the non-linear models could averagely increase the performance accuracy of linear models up to 30% and 29% for Kinta and Yamuna (Agra) River in the verification phase. This moderate increase could be attributed to the great advantage and the ability of non-linear models to deal with the complex and dynamic nature of the problems; this conclusion is similar to that of [18, 19, 20]. A similar conclusion can be drawn from the box plots shown in Fig. 5 for Kinta and Agra, respectively, concerning predicted values in the verification phase. In comparison with all models, Figure 6 proved that ANFIS had the overall best fitting and can be applied as a superior alternative for the forecasting of W.Q.I. for both Kinta and Yamuna (Agra) River.

Table 3. Performance efficiency of linear and non-linear models (ANFIS, ANN, and ARIMA)

Regions	Models	R^2	RMSE	MSE	MAPE	R^2	RMSE	MSE	MAPE
Kinta River	ANFIS-I (2, trimf,2)	0.6513	0.2477	0.0229	0.0092	0.5984	0.1231	0.0139	0.0292
	ANFIS -II (3, trimf,2)	0.7643	0.1210	0.0218	0.0056	0.7162	0.0876	0.0090	0.0298
	ANFIS **-III (5,** trimf,2**)**	**0.9689**	**0.0767**	**0.0082**	**0.0033**	**0.9510**	**0.0456**	**0.0031**	**0.0219**
	ANN-I (2, 3, 1)	0.6189	0.1984	0.0437	0.0574	0.5100	0.1176	0.0127	0.0249
	ANN-II (3, 5, 1)	0.6971	0.3128	0.0543	0.0531	0.6503	0.1043	0.0205	0.0234
	ANN-III (5, 6, 1)	**0.9112**	**0.1073**	**0.0402**	**0.0218**	**0.8915**	**0.1003**	**0.0156**	**0.0125**
Yamuna (Agra)	ANFIS -I (2, trimf,2)	0.5804	0.2487	0.0539	0.0137	0.5003	0.1321	0.0134	0.0262
	ANFIS -II (3, trimf,2)	0.6643	0.3255	0.0561	0.0291	0.5347	0.1491	0.0102	0.0316
	ANFIS **-III (5,** trimf,2**)**	**0.9213**	**0.0782**	**0.0273**	**0.0069**	**0.9113**	**0.0822**	**0.0072**	**0.0203**
	ANN-I (2, 2, 1)	0.6361	0.2431	0.0164	0.0188	0.6145	0.1904	0.0145	0.0928
	ANN-II (3, 4, 1)	0.7423	0.3247	0.0262	0.0049	0.7231	0.1139	0.0132	0.0526
	ANN-III (5, 8, 1)	**0.8834**	**0.0892**	**0.0119**	**0.0033**	**0.8343**	**0.0839**	**0.0067**	**0.0314**
	ARIMA-I	0.5727	0.2307	0.0538	0.1464	0.5386	0.2073	0.0343	0.0701

Kinta River	ARIMA-II	0.6016	0.2345	0.0769	0.1315	0.5949	0.2566	0.0558	0.0321
	ARIMA-III	**0.6908**	**0.1321**	**0.0136**	**0.0129**	**0.6392**	**0.1716**	**0.0049**	**0.0272**
	ARIMA-I	0.4969	0.4985	0.1508	0.2416	0.4969	0.3843	0.1237	0.1839
Yamuna (Agra)	ARIMA-II	0.5233	0.3994	0.1144	0.1325	0.5171	0.2871	0.0593	0.1358
	ARIMA-III	**0.6571**	**0.3012**	**0.0682**	**0.0181**	**0.6394**	**0.1637**	**0.0293**	**0.2141**

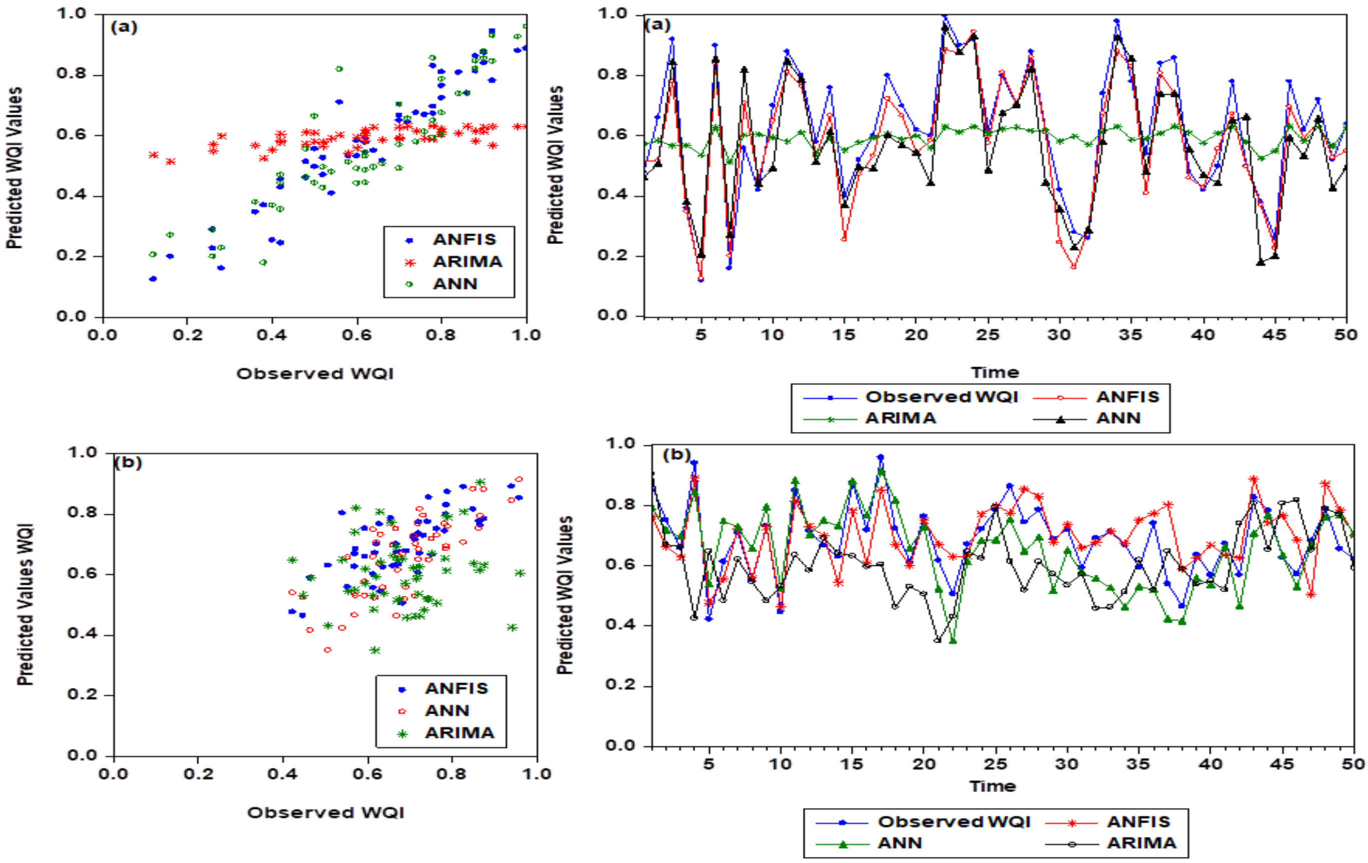

Fig. 5. Scatter and time series plots for observed vs computed W.Q.I. values (a) Kinta and (b) Agra rive

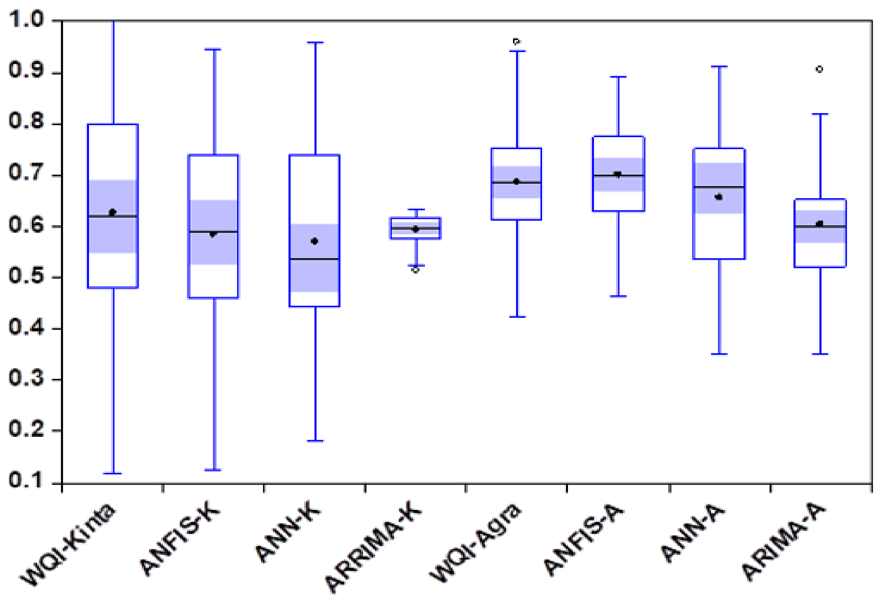

Figure 5. Box- plots for the predicted values for both Kinta and Agra stations

IV. CONCLUSION

The main target of this paper is to investigate and compare the performance of two non-linear (ANFIS and ANN) and a linear (ARIMA) models for forecasting the W.Q.I. with various water quality variables in two different regions. The performance efficiency criteria were compared in terms of Mean Square Error (M.S.E.), Root Mean Square Error (RMSE), Determination Coefficient (R^2), and Mean Absolute Percentage Error (MAPE). The obtained results of non-linear models indicated the superiority of the ANFIS model over ANN and ARIMA models for both the regions. The overall comparison of non-linear models for the two regions can averagely increase the performance accuracy of linear models up to 30% in the verification phase. To concludes, the ANFIS model proved high merit and, therefore, reliable in forecasting the W.Q.I. both at Kinta and Yamuna (Agra) River. To increase the accuracy and uncertainties problems of the models and to explore the contribution of each input combination, the results may also recommend that introducing other A.I. algorithms could lead to more precise and consistent prediction.

ACKNOWLEDGEMENTS

We want to acknowledge the help provided by the central pollution control board (CPCB) and the department of environment (D.O.E) Malaysia for providing the data used in this research work.

REFERENCES

[1] Sharma, D., & Kansal, A. (2011). *"Water quality analysis of River Yamuna using water quality index in the national capital territory"* , India (2000 – 2009), *0*, 147–157.

[2] Muhammad, S. Y., Makhtar, M., Rozaimee, A., Aziz, A. A., & Jamal, A. A. (2015*)." Classification model for water quality using machine learning techniques"*. International Journal of Software Engineering and Its Applications, *9*(6), 45-52.

[3] Abba, S. I., Said, Y. S., & Bashir, A. (2015). *"Assessment of Water Quality Changes at Two Location of Yamuna River Using the National Sanitation Foundation of Water Quality"* (NSFWQI). Journal of Civil Engineering and Environmental Technology, *2*(8), 730-33.

[4] Singh, P. (2017). *"Review on Data Mining Techniques for Prediction of Water Quality*", *8*(5), 396–401.

[5] Nourani, V., Khanghah, T. R., Sayyadi, M., Prof, A., Student, M. S., & Student, B. S. (2013). "*Application of the Artificial Neural Network to monitor the quality of treated water",* *3*(1), 38–45.

[6] Gaya, M. S., Abdul Wahab, N., & Samsudin, S. I. (2014). *"ANFIS modeling of carbon and nitrogen removal in a domestic wastewater treatment plant"*. J. Teknol, 67(5), 29-34.

[7] Nourani, V., Elkiran, G., & Abba, S. I. (2018). *"Wastewater treatment plant performance analysis using artificial intelligence–an ensemble approach"*. Water Science and Technology. DOI.org/10.2166/wst.2018.477.

[8] Singh, K. P., Basant, A., Malik, A., & Jain, G. (2009). *"Artificial neural network modeling of the river water quality",* — A case study, *220*, 888–895.

[9] Gazzaz, N. M., Yusoff, M. K., Aris, A. Z., Juahir, H., & Ramli, M. F. (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.

[10] Khan, Y., & Chai, S. S. (2017). *"Ensemble of ANN and ANFIS for Water Quality Prediction and Analysis-A Data Driven Approach"*. Journal of Telecommunication, Electronic and Computer Engineering (JTEC), 9(2-9), 117-122.

[11] Abba, S. I., Hadi, S. J., & Abdullahi, J. (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.

[12] Nourani, V., Hakimzadeh, H., & Amini, A. B. (2012). "*Implementation of artificial neural network technique in the simulation of dam breach hydrograph"*. Journal of HydroInformatics, 14(2), 478-496.

[13] K C Gouda, Libujashree R, Priyanka Kumari, Manisha Sharma, Ambili D Nair *"An Approach for Rainfall Prediction using Soft Computing"* International Journal of Engineering Trends and Technology 67.3 (2019): 158-164.

[14] Elkiran, G., Nourani, V., Abba, S. I., & Abdullahi, J. (2018). *"Artificial intelligence-based approaches for multi-station modelling of dissolve oxygen in river"*. Global Journal of Environmental Science and Management, *4*(4), 439-450.

[15] Kişi Ö, Nia AM, Gosheh MG, Tajabadi M.R.J., Ahmadi A (2012) *"Intermittent streamflow forecasting by using several data-driven techniques"*. Water Resour Manag 26:457–474.

[16] Legates, D. R., and McCabe, G. J. (1999*). "Evaluating the use of "goodness of fit" measures in hydrologic and hydroclimatic model validation*". Water Resources Research, 35(1), 233-241.

[17] Emamgholizadeh, S., Kashi, H., Marofpoor, I., & Zalaghi, E. (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.

[18] Sakizadeh, M. (2016). *"Artificial intelligence for the prediction of water quality index in groundwater systems"*. Modeling Earth Systems and Environment, *2*(1), 8.

[19] Al Suhili, R. H., & Mohammed, Z. J. (2014). *"Comparison between Linear and Non-linear ANN Models for Predicting Water Quality Parameters at the Tigris River"*. Journal of Engineering, 2*0*(10), 1-15.

[20] Hameed, M., Sharqi, S. S., Yaseen, Z. M., Afan, H. A., Hussain, A., & Elshafie, A. (2017). *"Application of artificial intelligence (A.I.) techniques in water quality index prediction: a case study in a tropical region, Malaysia"*. Neural Computing and Applications, *28*(1), 893-905.

[21] Chen, W., & Liu, W. (2015). *"Water Quality Modeling in Reservoirs Using Multivariate Linear Regression and Two Neural Network Models", 2015*.

[22] Kavasseri, R. G., and Seetharaman, K. (2009). *"Day-ahead wind speed forecasting using f-ARIMA models"*. *Renewable Energy*, *34*(5), 1388-1393.