DOI: https://dx.doi.org/10.4314/gjpas.v28i1.10

GLOBAL JOURNAL OF PURE AND APPLIED SCIENCES VOL. 28, 2022: 83-90
COPYRIGHT© BACHUDO SCIENCE CO. LTD PRINTED IN NIGERIA ISSN 1118-0579
www.globaljournalseries.com, Email: globaljournalseries@gmail.com

EVALUATION OF AUTO REGRESSIVE INTEGRATED MOVING AVERAGE (ARIMA) AND ARTIFICIAL NEURAL NETWORKS (ANN) IN THE PREDICTION OF EFFLUENT QUALITY OF A WASTEWATER TREATMENT SYSTEM.

HOWARD, C. C., ETUK, E. H. AND HOWARD, I. C.

(Received 10 January 2022; Revision Accepted 2 March 2022)

ABSTRACT

The main objective of wastewater treatment is to purify the water by degradation of organic matter in the water to an environmentally friendly status. To achieve this objective, some effluent (waste water) quality parameters such as Chemical oxygen demand (COD) and Biochemical oxygen demand (BOD_5) should be measured continuously in order to meet up with the said objective and regulatory demands. However, through the prediction on water quality parameters, effective guidance can be provided to comply with such demand without necessarily engaging in rigorous laboratory analysis. Box-Jenkin's Auto Regressive Integrated Moving Average (ARIMA) technique is one of the most refined extrapolation techniques for prediction while Artificial Neural Network (ANN) is a modern non-linear method also used for prediction. The Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Root Mean Square Error (RMSE) and Correlation coefficient (r) are used to evaluate the accuracy of the above-mentioned models. This paper examined the efficiency of ARIMA and ANN models in prediction of two major water quality parameters (COD and BOD_5) in a wastewater treatment plant. With the aid of R software, it was concluded that in all the error estimates, ANNs models performed better than the ARIMA model, hence it can be used in the operation of the treatment system.

KEYWORDS: prediction, artificial neural networks (ANNs), auto regressive integrated moving average (ARIMA).

INTRODUCTION

Inadequate management of a Waste Water Treatment Plant may cause serious environmental and public health problems, as its effluent when discharged into a receiving water body can cause or spread various diseases to human beings. Operation of a wastewater treatment plant is often affected by a variety of physical, chemical, and biological factors. In order to follow the treatment plant performance during the operation, effluent measurements would not be sufficient. Predicting any of these parameters, depending upon the influent water quality, will help the operator to control the system and to take necessary precautions before any problem comes up. Modeling a waste water treatment plant is considered a difficult task due to complexity of the treatment processes. A better control of a system can be achieved through developing a mathematical technique for predicting plant performance based on past observations of certain parameters.

The complex physical, biological and chemical processes involved in wastewater treatment may exhibit non-linear behaviors which cannot be described by linear mathematical models.

In recent years, several water quality models such as traditional mechanistic approaches have been developed in order to manage the best practices for conserving the quality of water. Most of these models need several different input data which are not easily accessible and make it a very expensive and time-consuming process. ANN has been observed to be a suitable approach for water quality modeling (Chen *et al*., 2003; Jan-Tai *et al*., 2006).

In addition, ANNs provide different advantages over traditional modeling approach of a wastewater treatment plant. For instance, when ANNs are applied to prediction of wastewater treatment plant performance task they will result in reduction of cost for undertaking laboratory tests. And also, parameter such as BOD_5 could be predicted from the model instead of waiting for five days

HOWARD, C. C., Department of Mathematics, Faculty of Science, University of Africa Toru-orua, Sagbama. Bayelsa State, Nigeria.
ETUK, E. H., Department of Mathematics, Faculty of Science, Rivers State University Port Harcourt, Nigeria.
HOWARD, I. C., Department of Chemistry/Biochemistry, School of Ind. and App. Sciences, Federal Polytechnic, Nekede, Owerri. Imo State, Nigeria

for its analysis, thereby saving time in addition to the advantage that efficiently predicted values of parameters will provide for proper operation and control of the waste water treatment plant. In a related study Areerachakul, (2012) opined that "several other modeling techniques such as; Fuzzy Inference System (FIS) and Neural Network (NN) be employed for the production of forecasting models to estimate water quality parameters". He compared the predictive power of the Adaptive Neuro-Fuzzy Inference System (ANFIS) model and the ANN model to estimate the BOD_5 on a data set from eleven sampling stations of the Saep channel in Bangkok, Thailand. Somvanshi *et al*. (2006) investigated two fundamentally different methods of ARIMA and ANN to design a model and predict behavior patterns in rainfall events based on past behavior. The study showed that the ANN model can be used as a predictive forecasting tool to estimate rainfall, which improves the ARIMA model. Sharma and Singh (2011) studied forecasting models to make comparisons between models to identify models suitable for forecasting rain, concluding that the ANN approach is better than other models. Rene and Saidutta (2008) developed several empirical relationships, between COD and BOD_5 with TOC (total organic carbon) using a combination of regression and artificial neural network analysis. The essence was to use TOC to estimate the accompanying BOD_5 or COD in water quality monitoring in a refinery's wastewater system. It was observed that the three models they developed gave accurate results, which indicates the versatility of the developed models.

This paper examined the efficiency of Auto regressive integrated moving average (ARIMA) and artificial neural network (ANN) models in prediction of two major water quality parameters (BOD_5 and COD) in a wastewater treatment plant

MATERIAL AND METHODS

Study area and data collection

The data for this study was generated by an oil and gas consultancy firm that actually carried out the field and laboratory work which involves collection of weekly waste water samples (Produced water) for the analysis of principal parameters BOD_5, COD, Conductivity, Temperature, pH, etc. using the Standard method for the examination of water and waste water (APHA, 1998) from the wastewater treatment plant located between Longitude 4^{o}34.276′ and Latitude 8^o 25.557′ at the Gulf of Guinea.

The water quality data (five years, total of 260 observations) were divided into two data sets. The first data set containing former 4-year records was used as the training data for model development; the second data set containing the remaining year's records was used as the testing and validating data to evaluate the performance of the established models. In this paper, only 52 data points from the test data set for forecasting is considered. The models were built using the Times Series Forecasting System tool of the R software packages.

Autoregressive Integrated Moving Average Model (ARIMA)

Box and Jenkins (1976) developed a practical approach to construct ARIMA models, which have fundamental effects on the applications of time series analysis and forecasting. Their methodology includes the following three iterations: model identification, parameter estimation and diagnostic testing. The autocorrelation function (ACF) and partial autocorrelation function (PACF) of the sampled data are used as the basic tools to identify the order of the best ARIMA model.

For the models obtained, diagnostics tests are performed using (a) Residual ACF (b) Ljung Box test. The residuals after fitting an ARIMA model should be a random noise. Therefore if the autocorrelations and partials of the residuals are obtained, there should not be any significant autocorrelation and partial autocorrelations. In this study, the various correlations for BOD_5 and COD are up to 41.6 lags each, were computed and their significance is tested by Box-Ljung test as shown in Table 1. Since the *p*-values for BOD_5 and COD are all above 0.05, indicating not significantly different from zero at a reasonable level, the selected ARIMA models are the appropriate models.

Table 1. Parameters estimation of ARIMA model.

	Model type	r^2	BIC	Ljung Box	Q(18)	
				Statistics	DF	*P*-value
BOD_5	(1, 0, 1)	0.098	2.74	29.6296	38.6	0.84973
COD	(2, 0, 0)	0.23	2.71	38.935	38.6	0.45463

Artificial neural network (ANN)

In this study, artificial neural network algorithm was applied to evaluate the effluent quality parameters (BOD_5 and COD).The ANN models are increasingly being used for prediction or simulating water resources variables because they are often capable to model complex systems with unknown or difficult behavioral rules or underlying physical processes. Most of these studies showed that ANNs performed better than traditional modeling techniques (Zhang *et al,* 2002). Artificial neural networks (ANNs) provide a method for solving many types of non-linear problems that are difficult to solve by traditional techniques (Leahy *et al*., 2008, Haciismailoglu *et al.,* 2009). All inputs and outputs are normalized between 0 and 1 in artificial neural network software. A proper process of data normalization and defacement is required before and after the execution of the program. The best and simplest way to normalize is to divide it by the maximum and after the execution of the program, multiply the result by the same amount. There are several neural network models for various applications available in research, but the ANN considered in this study is a fully connected multilayer perceptron (MLP), a network composed of neurons which consists of three layers of neurons: (1) an input layer; (2) a hidden layer, and (3) output layer. Each neuron has a number of inputs (from outside the network or the previous layer) and a number of outputs (leading to the subsequent layer or out of the network). Dawson *et al.*, (2006) in their study have averred that "a neuron computes its output response based on the weighted sum of all its inputs according to an activation function", hence the neuron structure for this study is shown in Figure 1 and 2 below. MLP neural networks were used to estimating BOD_5 and COD in the treated waste water.

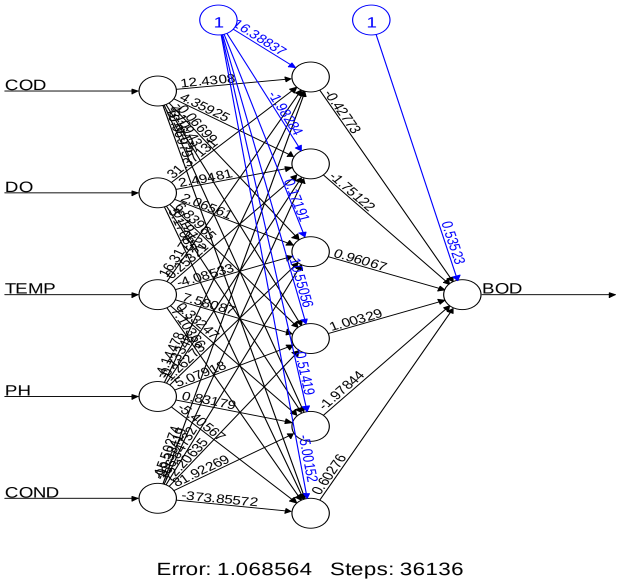

Figure 1: Artificial neural network structures for weekly BOD_5 prediction parameter

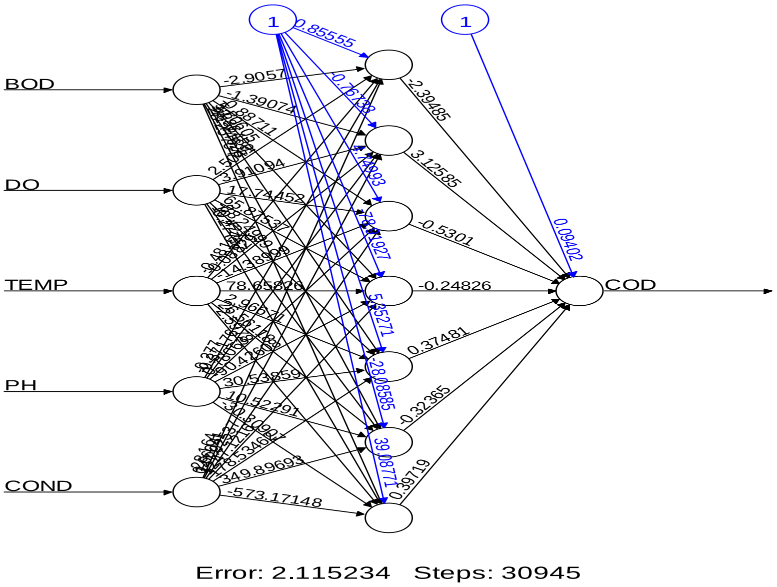

Figure 2: Artificial neural network structure for weekly COD prediction parameter.

The MLP-ANN proposed in this study had three layers: an input layer, a single hidden layer, and an output layer. The number of neurons in the input and output layers was determined by the number of input and output variables considered in the model, respectively. The number of neurons in the hidden layer was selected by analyzing the root mean squared error (RMSE) of the trained ANNs when a different number of hidden neurons in the ANN were used. To reduce the randomness of the training process, 100 repetitions of the experiment were performed for each number of hidden neurons. The activation function used in the proposed ANN is the sigmoid function (Equation 1) for the neurons of the hidden layer as well as the output layer. Because the input layer does not receive signals from other neurons, their neurons do not have an activation function because they only send input variables to neurons of the hidden layer.

$$Sig(x) = \frac{1}{1 + \exp(-x)} \quad (1)$$

ANN training was performed by applying the resilient back propagation with the backtracking algorithm employing the R software.

PREDICTION USING ANN

ANN model building process was performed using the significant water quality parameters. The best-suited architecture of Feed Forward Neural Network Model for our weekly BOD_5 and COD data were selected by comparing methods and changing the layer and number of neurons in each network. The BOD_5 model had an input environment with significant water quality parameters, one hidden layer with 6 neurons and one neuron in the output layer (see Figure 1 above) while COD model had an input environment with significant water quality parameters, one hidden layer with 7 neurons and one neuron in the output layer (see Figure 2 above). The number of neurons in the hidden layers was varied, as shown in Table 2 and 3 below. The number of hidden neurons (six and seven) for BOD and COD respectively with the smallest value of RMSE is the best fit. A resilient back propagation with weight backtracking algorithm was used for training the multilayer neurons until the best combination was achieved. A sigmoid activation function was used in the hidden layer and output layer. Set of random values distributed uniformly from 0 to 1 were utilized to start the weight of the neural network models. The best-fitting networks selected for BOD_5 and COD are N (5×6×1) and N (5×6×1) respectively.

Table 2: Neural network performance (BOD_5) using different number of hidden neurons

Number of hidden neurons	Training (RMSE)
4	28.09486
5	30.46437
6	26.81779
7	33.83007
8	31.51882
9	30.18726
10	30.93310

Table 3: Neural network performance (COD) using different number of hidden neurons

Number of hidden neurons	RMSE
4	24.29487
5	25.56701
6	23.62751
7	21.94817
8	24.95865
9	32.14235
10	32.14235

Evaluation Criteria for ANN and ARIMA Predictions

Four statistical criteria were applied to evaluate the performance of ANN and MLR models. These criteria were root mean square error (RMSE), mean absolute percentage error (MAPE), mean absolute error (MAE) and correlation coefficient (r). The expressions for these measures are as follows:

Mean absolute error (MAE)

$$\frac{1}{N}\sum_{i=1}^{N} O_i - P \qquad (2)$$

Root mean square error (RMSE):

$$\sqrt{\frac{\sum_{i=1}^{N}\left(O_i - P_i\right)}{O_i}} \qquad (3)$$

Mean absolute percentage error (MAPE):

$$\frac{1}{N}\sum_{i=1}^{N}\left|\frac{O_i - P_i}{O_i}\right| \qquad (4)$$

Correlation Coefficient (r) $\frac{\sum(O_i-\bar{\bar{O}})\,(P_i-\bar{\bar{P}})}{\sqrt{\sum(O_i-\bar{\bar{O}}\,)^2\,(P_i-\bar{\bar{P}})^2}}$

(5)

Where *N* is the number of data, O_i observed values, P_i predicted values at time *i* and the bar denotes the mean of the variable. For the best prediction, the MAE, RMSE and MAPE values should be small i.e., close to 0. The recital of water quality parameters forecasting models had been evaluated on the basis of R packages.

RESULTS AND DISCUSSION

The development of ARIMA and ANNs models were carried out to assess the predictive performance of the models. The water quality data (five years, a total of 260 observations) were divided into two data sets. The first data set containing former 4-year (2007-2010) records were used as the training data for model development; the second data set containing the remaining year (2011) records were used as the testing data to evaluate performance of the established models. All the models were built using the Times Series Forecasting System tool of the R software package

Auto Regressive Integrated Moving Average (ARIMA)

An ARIMA modeling of BOD_5 and COD data quality parameters of the wastewater were performed. The weekly time series of wastewater quality parameters of BOD_5 and COD were plotted which show that the time series data sets are stationary so they do not need any transformation in the data sets as shown in Figure 3 and 4 below. Then the autocorrelation function (ACF) and partial autocorrelation function (PACF) of the weekly BOD_5 and COD time series are utilized to quantify the values of p and q of the ARIMA models. As the previous section, we found the best fitting models for BOD_5 and COD are ARIMA (1, 0, 1) and (2, 0, 0) models, which are ARIMA model with autocorrelation 1, with integration of order 0 and moving average of order 1 and ARIMA model autocorrelation 2, with integration of order 0 and moving average of order 0, respectively, represented by Equation 6 and 7

$$y_t = 98.7543 + 0.79149y_{t-1} - 0.5907w_{t-1} + w_t \qquad (6)$$

$$y_t = 91.7946 + 0.3357y_{(t-1)} + 0.2267y_{(t-2)} + e_t \qquad (7)$$

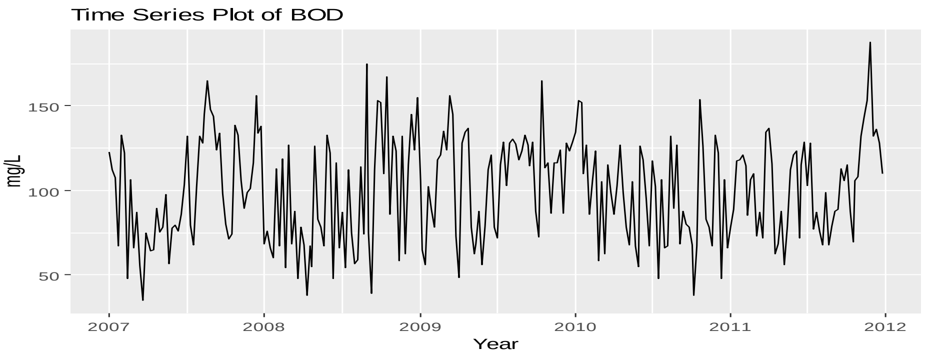

Figure 3: Time series plot of BOD_5

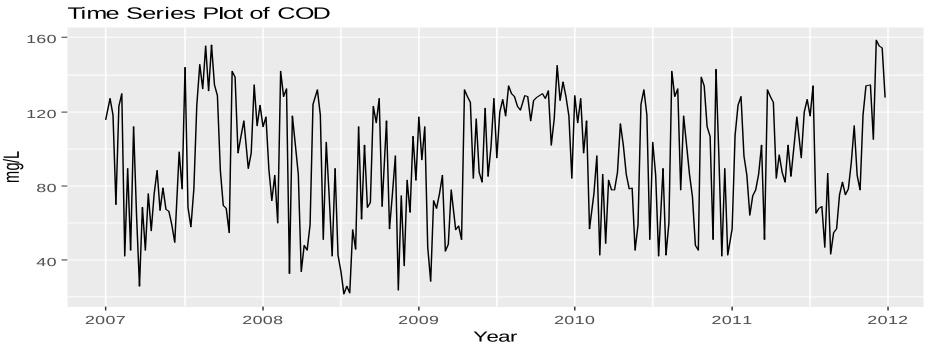

Figure 4: Time series plot of COD

A suitable model to predict water quality time series was built using ARIMA. As shown in Figure 3, although ARIMA models vary with the range, the model predictions are not adequate. This is due to the limitation of the linear modeling algorithm in the ARIMA model which is unsatisfactory in identifying and predicting nonlinear time series of water quality parameters.

Artificial Neural Network (ANN)

As described above, network training is done using a resilient back propagation with weight backtracking algorithm. A sigmoid function is used as the transfer function in both the hidden layer and output layer due to its suitable application. The best results were obtained for the ANN composed of six (6) and seven (7) neurons in the hidden layer for BOD_5 and COD respectively. The best-fitting networks selected for BOD_5 and COD are N (5×6×1) and N (5×6×1) respectively.

Comparative Performance of ARIMA and ANN

The performance of the ARIMA and ANN models are shown in terms of the correlation coefficient (r), i.e., the strength of the linear relationship between the observation and prediction for all parameters, as shown in Table 4 and 5 below. The r values of ARIMA is linear positive but weak for both BOD_5 and COD. These values are not satisfactory in common model applications. This is due to the limitation of the linear modeling algorithm in ARIMA model which is unsatisfactory in identifying and predicting nonlinear time series of water quality data. For the ANN model, the r is strong and positive. The results indicate that the neural network that was developed is able to detect the pattern in water quality parameters to provide prediction of the daily variation data. This means that they are satisfactory in identifying and predicting nonlinear time series of water quality data. In other words, the ANN model was able to detect and identify the pattern of water quality parameters to provide desired and valid predictions better than the ARIMA model.

Table 4: Comparative performance of ARIMA and ANNs models (BOD_5 data)

Model	RMSE	MAPE	MAE	r
ARIMA	27.1513	22.9013	22.7897	0.09
ANNs	23.0524	20.8972	19.1746	0.75

Table 5: Comparative performance of ARIMA and ANNs models (COD data)

Model	RMSE	MAPE	MAE	r
ARIMA	29.741441	27.01971	24.51272	0.23
ANNs	27.506245	24.415084	20.9486	0.65

Again comparison on the basis of error, the results revealed that the ANN has the lowest root mean square error (RMSE), mean absolute percentage error (MAPE) and mean absolute error (MAE) for predicting BOD_5 and COD as compared to ARIMA (Table 4 and 5 above). It is found that the ANN model could probably predict BOD_5 and COD with a better performance owing to their greater flexibility and capability to model linear/ nonlinear relationships. Figure 5 and 6 shows the observed and predicted values of ANN for BOD_5 and COD respectively and Figure 7 and 8 shows the observed and predicted values of ARIMA for BOD_5 and COD respectively. They show that ANN technique is more feasible as the observed and predicted values tend to be closer more in predicting BOD_5 and COD than the ARIMA technique. This study aimed at the comparison of ANN and ARIMA models for BOD_5 and COD predictions showed that ANN model proved to be a better technique for accurate prediction of BOD_5 and COD with minimum value of RMSE, MAPE and MAE rather than ARIMA.

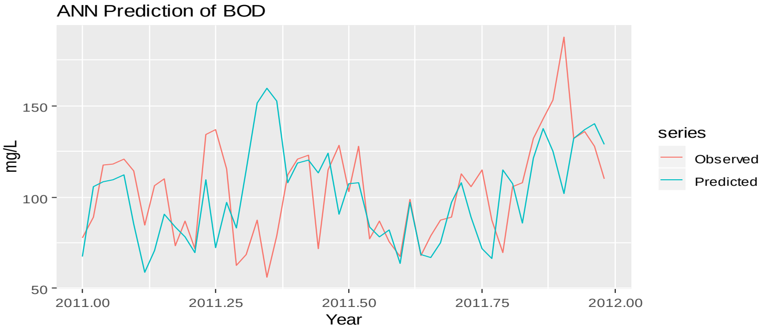

Figure 5: Comparison of the observed values and those predicted by the ANN model (BOD_5 data)

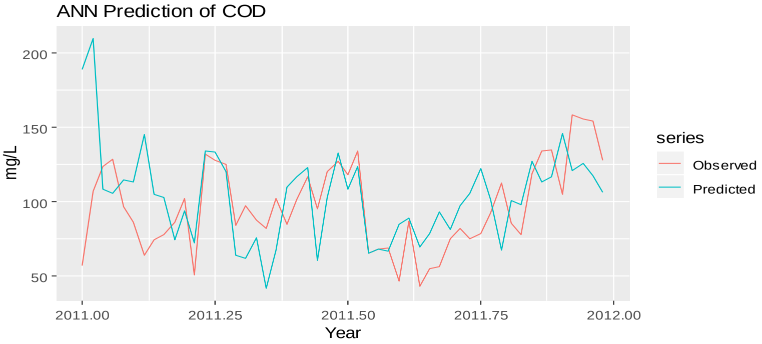

Figure 6: Comparison of the observed values and those predicted by the hybrid model (COD data)

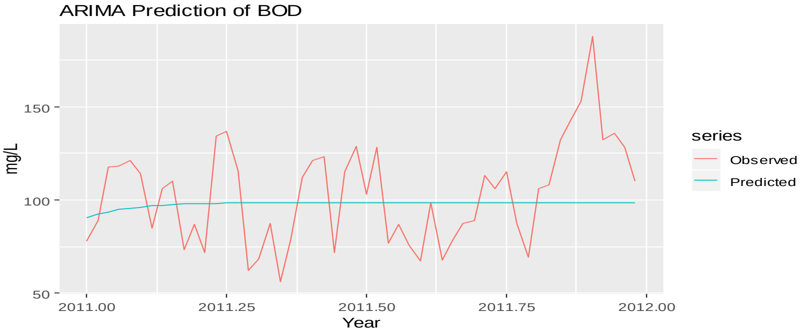

Figure 7: Comparison of the observed values and those predicted by the hybrid model (BOD_5 data)

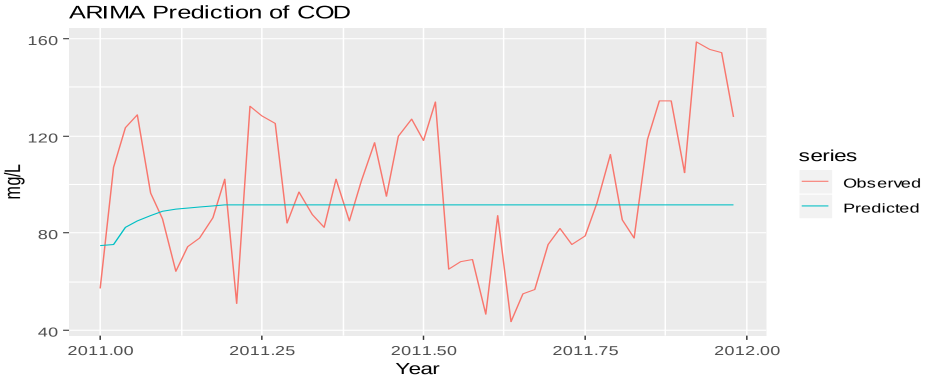

Figure 8: Comparison of the observed values and those predicted by the hybrid model (COD data)

CONCLUSION

This study used ARIMA and ANNs models to predict the water quality (BOD_5 and COD) time series data. The results obtained showed that ANN model is more reliable and suitable in predicting effluent quality. The ANN model developed in this study can be more useful in water quality management efforts to ensure that water resource is sustainable for the future. In this study, four accuracy measures, the RMSE, MAPE, MAE and r were formulated in order to demonstrate the performance of the developed models in predicting effluent quality time series. The ANN model performance as compared with ARIMA model gave the least values of MAE, MAPE and RMSE and also an improved performance in predicting effluent quality time series

REFERENCES

American Public Health Association (APHA), (1998). Standard Methods for the Examination of Water and Wastewater. (20th ed.) USA.: American Public Health Association.

Areerachakul, S, Junsawang, P, and Pomsathit. A., 2011. Prediction of Dissolved Oxygen Using Artificial Neural Network. Int Conf Comput Commun Manage 5:524–528

Box G.E.P. and Jenkins G.M., 1976.Time Series Analysis, Forecasting and Control. Revised . Toronto: Holden-Day,

Chen, J. C., Chang, N. B., and Shie, W. K., 2003. Assessing wastewater reclamation potential by neural network model. *Eng. Appl. Artif. Intell*; 16: 149–157.

Dawson C. W., Abrahart, R. J., Shamseldin, A. Y., and Wilby, R. L., 2006. Flood estimation at ungauged sites using artificial neural networks. *J Hydrol* 319:391–409.

Haciismailoglu M. C., Kucuk I, Derebasi N., 2009. Prediction of dynamic hysteresis loops of nano-crystalline cores. *Expert Syst Appl*, 36:2225–2227.

Jan-Tai, K.u, Ying-Yi, W, and Wu-Seng, L., 2006. A hybrid neural–genetic algorithm for reservoir water quality management. *Water Research*; 40:1367 – 1376.

Leahy P, Kiely G, Corcoran G., 2008. Structural optimisation and input selection of an artificial neural network for river level prediction. *J Hydrol*, 355:192–201.

Rene, E. R and Saidutta, M. B., 2008. Prediction of Water Quality Indices by Regression Analysis and Artificial Neural Networks. *Int J Environ Res* 2(2):183–188.

Sharma, M. A. and Singh, J. B., 2011. Comparative study of Rainfall forecasting models. *New York Science Journal,* Vol.4, No. 7, pp. 115-120.

Somvanshi, V. K., Pandey, O. P., Agarwal, P. K., Kalanker, N. V., Prakesh, M. R., and Chand, R., 2006. Modeling and prediction of rainfall using artificial neural network and ARIMA techniques. *Journal of Indian Geophysical Union* 10(2); 141-151.

Verma A. K. and Singh, T. N., 2013. Prediction of water quality from simple field parameters. *Environ Earth Sci*, 69(3):821–829.

Zhang, Y., Pulliainen, J., Koponen, S. and Hallikainen, M., 2002. Application of an empirical neural network to surface water quality estimation in the Gulf of Finland using combined optical data and microwave data. *Remote Sensing of Environment*; 81: 327– 336.