VOL. 23 NO. 2
ISSN: 2010-1086
DECEMBER, 2021

African Scholar Publications & Research International

African Scholar Journal of African Sustainable Development (JASD-2)

Application of a Hybrid Multiple Linear Regression (MLR) and Artificial Neural Network (ANN) Models to Produced Water Quality Prediction.

[1*]Howard, C.C., [2] Etuk, E. H. and [3]Howard, I. C.

[1]Department of Mathematics, Faculty of Science, University of Africa Toru-orua, Sagbama. Bayelsa State, Nigeria. [2]Department of Mathematics, Faculty of Science, Rivers State University Port Harcourt, Nigeria. [3]Department of Chemistry/Biochemistry Federal Polytechnic, Nekede, Owerri. Imo State, Nigeria

Abstract

In this paper the produced water quality prediction of a crude oil production facility (flow station) located at the Gulf of Guinea, Nigeria is presented. The time series data used were generated by a standard laboratory that actually carried out the field and laboratory analysis which involves weekly water quality data obtained directly from flow station for the period of five years. In order to predict water quality, hybrid models consisting of Multiple Linear Regression (MLR) models and Artificial Neural Network ANN) models were developed. In the first step, MLR models were first used to capture the linear component in the time series data and then the errors obtained, ANN were developed taking into account the non-linear pattern that MLR could not capture, in order to reduce potential errors. Once the hybrid was developed, 52 data points out of 260 data from the flow station were used for validation of the model and the result were compared with the MLR and ANN models built separately. These three different models are compared for their prediction abilities using statistical error measures viz. root mean square error (RMSE), mean

absolute error (MAE) and mean absolute percentage error (MAPE). Results showed that in these entire error estimates the hybrid MLR- ANN performance model was better than the MLR and ANN models in the examined flow station.

Keywords: *Water Quality prediction, Time series, Multiple Linear Regression (MLR), Artificial Neural Networks (ANN) and Hybrid model (MLR-ANN)*

Introduction

Water is the most important natural resource not only of a state or a country, but of the entire humanity. The prosperity of a nation depends primarily upon the judicious exploitation of this resource. Thus, it can be stated that the primary wealth of a nation is water, which flows in rivers, streams and beneath the earth alongside other mineral resources e.g. crude oil (Gajendran, 2011).

This all important resource is oftentimes misused and hence its quality characteristics are altered in such a way that they are not useful for the purpose for which they are intended. Water quality refers to the chemical, physical and biological characteristics of water. That is, it is a measure of condition of water relative to the requirements of one or more biotic species and or to any human need or purpose. (Boah *et al.,* 2011, Ehi-Douglas *et al.,* 2018).

On the other hand Produced water is a wastewater from crude oil processing in the oil and gas industry. It is mainly water that was mixed with crude oil and natural gas in an oil formation, which was drilled out, mixed again with several chemicals to aid its separation from the crude oil and natural gas. The volume of such water is usually very high. For instance Howard *et al.,* (2012) in their study gave a daily estimate of 95000liters of produced water from a given flow station in the Niger delta. However, this depends on the nature and number of the wells, age of the wells, etc. This water is filled with lots of chemicals that if not treated properly to meet required standards could destroy both fauna and flora of its immediate place of disposal, which of course will lead to economic loss. Hence the management of produced water is a major environmental

challenge in the world (EGASPIN 2018, Howard *et al.*, 2017). The management of produced water requires an integrative and multifaceted approach so as to protect the environment. One of such approach is the modeling of the produced water quality parameters from data generated from standard laboratory analysis of such produced water over time to forecast future trend so as to establish when things are going wrong in the operation or system. Various techniques have been adapted by different researchers to predict water quality parameters of which time-series analysis has been adjured.

The Multiple Linear Regression (MLR) method is commonly used techniques to obtain a linear input output model for a given dataset (Sahoo *et al*, 2009, Torres *et al*. 2005). However, this model will face some difficulties, especially when the independent variables are following certain distribution. Thus, Artificial Neural Network (ANN) was adopted as an approach to extracting information, required no priori assumptions about the model in terms of mathematical relationships or distribution of data and it is a well suited method with self-adaptive, self-organizing and error tolerance (Gonzalez *et al*., 2011 and Vilas *et al*., 2010). Efficiency and precision in prediction of water quality parameters using the hybrid MLR-ANN model is still a pandemic among researchers, due to the natural conditions in ocean water systems itself, which involved chemical, biological and physical processes and interaction among them may affect the model performance drastically. Thus, to overcome this problem as well as to improve the strength of MLR, we proposed a hybrid approach, i.e. - Multiple Linear Regression MLR to Artificial Neural Network ANN coined as Multiple Linear Regression - Artificial Neural Network (MLR-ANN).

Muhamad *et al*., (2016) examined the capabilities of the MLR-ANN model and compared it with the individual models MLR and ANN.. The results indicated that the model improved the capabilities of the MLR-ANN model compared to the single model MLR and ANN. Thus, they discovered that MLR-ANN model is efficient and accurate. So, with regard to the importance of prediction of water quality parameters, the focus of this paper was to develop a hybrid model to predict water quality time series data and assess its performance relative to MLR and ANN models.

Material and methods

Study area and scope

The data for this study was generated by a standard laboratory that actually carried out the field and laboratory work which involves collection of weekly effluent samples (Produced water) for the analysis of principal parameters BOD_5, COD, DO, etc. using the standard method for the examination of water and waste water (APHA, 1998) from a flow station located between Longitude 4°34.276′ and Latitude 8° 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 the first four years record was used as the training data for model development; the second data set containing the remaining year's records was used as the testing data to evaluate the performance of the established models. In this study, only 52 data points from the test data set for forecasting was considered. All the models were built using the Times Series Forecasting System tool of the R software package. The data compositions for the two water quality parameter were given in the table 1 below.

Table 1 Sample composition for two water quality parameters

Parameters	Sample size (2007-2011)	Training set (size) (2007-2011)	Test set (size) (2011)
COD	260	208	52
DO	260	208	52

MULTIPLE LINEAR REGRESSION (MLR)

Multiple Linear Regression expresses the relation between dependent variable y and more independent variable $(x_1, x_2, x_3, ..., x_p)$. Linear regression simply has one dependent variable which varies with one independent variable. However, when we need to explain about the dependent variable with two or more independent variables we need to use multiple linear regression. The multiple linear regression models as in Equation (1) is as follow:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p + \varepsilon \quad (1)$$

Where, β is the coefficient of parameters, y and x are dependent and independent variables respectively, while ε i.e. error term.

Artificial Neural Network Model

Artificial Neural Networks are massively parallel adaptive networks of simple non-linear computing elements called neurons which are intended to abstract and model some of the functionality of the human nervous system in an attempt to partially capture some of its computational strengths. Artificial Neural Network (ANN) is loosely based on biological neural systems, in that; they are made up of an interconnected system of neurons. Also, a neural network can identify patterns adaptively between input and output data set in a somewhat analogous fashion to the learning process. Neural networks are highly robust with respect to underlying data distributions and no assumptions are made about relationships between parameters.

Artificial Neural Networks (ANNs) provide a methodology for solving many types of non-linear problems that are difficult to solve by traditional techniques. In Artificial Neural Network Software all inputs and outputs are normalized between 0 and 1. Appropriate process of normalization and denormalization of data is needed before and after the program execution. The best and the simplest way is to divide it by the maximum for normalization and after the program execution the result is to be multiplied by the same amount. There are many neural network models, but the ANN considered in our work was a totally connected multilayer perceptron (MLP), which is a network composed by neurons called perceptrons, whose structure is shown in Figure 1.

They have a set of m inputs ($x_1, x_2, \ldots, x_m$) multiplied by their respective weights ($w_1, w_2, \ldots, w_m$) and combined with a bias (b) to generate the signal:

$$V = \sum_{i=1}^{m} w_i \times x_i + b \qquad (2)$$

The activation function φ(.) is then applied to generate the output:

$$Y = \varphi\left(\sum_{i=1}^{m} w_i \times x_i + b\right) \qquad (3)$$

MLPs have connections with all of the perceptrons of the previous layer and their outputs serve as input of all of the neurons of the next layer. Figure (2) shows one example of a totally connected MLP with four inputs, one output, and one hidden layer with five neurons.

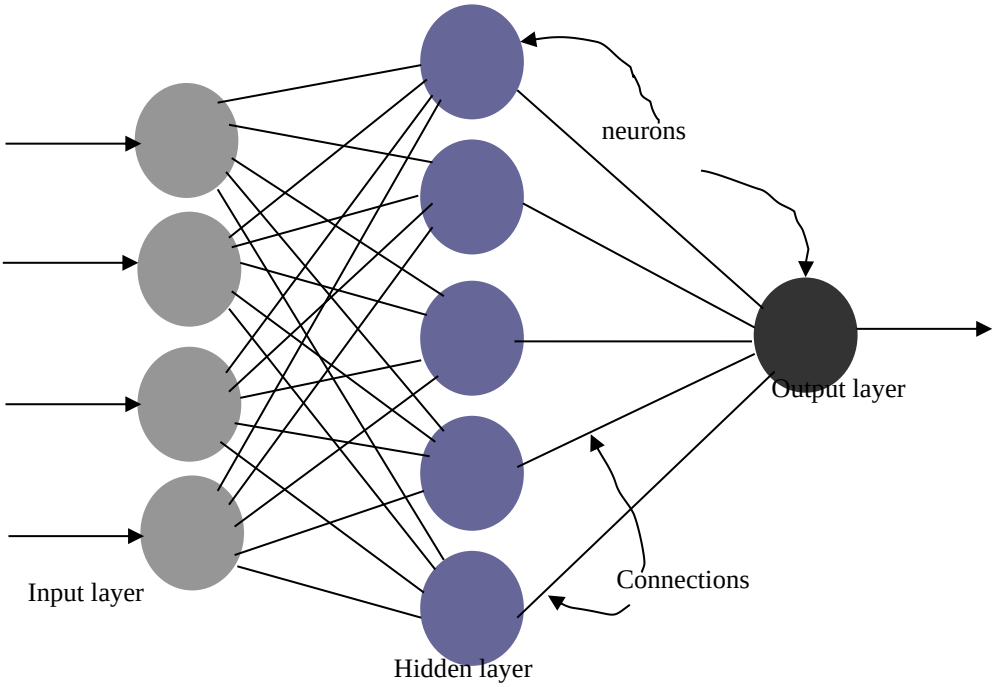

Figure 2: Diagram showing *totally connected feed forward* ANN.

The MLP-ANNs proposed in this work had three layers: the input layer, one single hidden layer, and one output layer. The number of neurons in the input and output layers were determined respectively by the number of input and output variables considered in the model. The number of neurons in the hidden layer was chosen analyzing the root mean squared error (RMSE) of the trained ANN when a different number of hidden neurons were used in the ANN. To minimize the randomness of the training process, 100 repetitions of the experiment for each number of hidden neurons were performed. The activation function used in the proposed ANN is the sigmoid function (Eq. 4) for the neurons of the hidden layer and 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 the input variables to the neurons of the hidden layer

$$Sig(x) = \frac{1}{1 + \exp(-x)} \qquad (4)$$

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

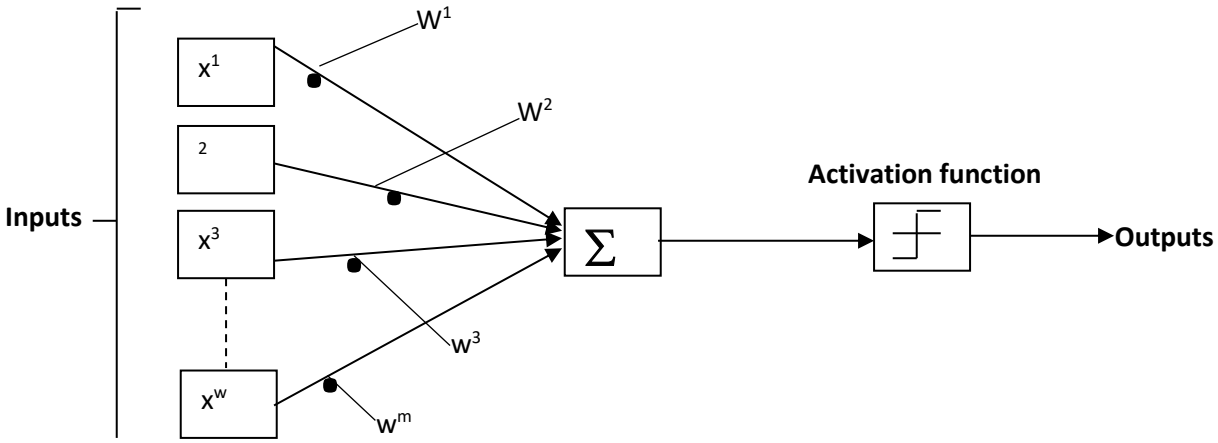

Figure 1 General design of neural network (N^{p-q-1})

Hybrid model (MLR-ANN)

Some researchers in hybrid linear and nonlinear models believe that it may be reasonable to consider a time series to be composed of a linear autocorrelation structure and a nonlinear component (Zhang, 2003). That is,

$$Y_t = N_t + L_t \quad (5)$$

Where L_t denotes the linear component and N_t denotes the nonlinear component. These two components have to be estimated from the data. First, we let MLR to model the linear component, and then the residuals from the linear model will contain only the nonlinear relationship (Zhang, 2003). Let e_t denote the residual at time *t* from the linear model, then

$$e_t = y_t - \hat{L}_t \quad (6)$$

Where $\hat{L}_t$ is the forecast value for time *t* from the estimated relationship (1). By modeling residuals using ANNs, nonlinear relationships can be discovered (Zhang, 2003). With *n* input nodes, the ANN model for the residuals will be:

$$e_t = f(e_{t-1}, e_{t-2}, \ldots, e_{t-n}) + e_t \quad (7)$$

Where *f* is a nonlinear function determined by the neural network and e_t is the random error. Note that if the model *f* is not an appropriate one, the error term

is not necessarily random (Zhang, 2003). Therefore, the correct model identification is critical. Denote the forecast from (7) as $\hat{N}_t$, the combined forecast will be

$$\hat{y}_t = \hat{L}_t + \hat{N}_t \tag{8}$$

The hybrid model exploits the unique feature and strength of MLR model as well as ANN model in determining different patterns. Thus, it could be advantageous to model linear and nonlinear patterns separately by using different models and then combine the forecasts to improve the overall modeling and forecasting performance (Zhang, 2003).

Performance Evaluation Criteria of Comparison

In this paper, both linear and nonlinear models were used in the data sets, and also error of estimation method was used to determine the accuracy of the data as the smaller the error, the higher the accuracy of the data. The performance criteria evaluation model to measure the error of data as well as error reduction is as follows

Root mean square error (RMSE):

$$\sqrt{\frac{\sum_{i=1}^{N}(O_i - P_i)}{O_i}} \tag{9}$$

Mean absolute percentage error (MAPE):

$$\frac{1}{N}\sum_{i=1}^{N}\left|\frac{O_i - P_i}{O_i}\right| \tag{10}$$

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 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 COD model results

In the first step, data set of COD time series data was fitted with multiple regression models to predict the weekly chemical oxygen demand (COD) as

dependent parameter taking the other weekly independent parameter as BOD_5. The most significantly contributed parameter (BOD_5) is selected using forward stepwise regression analysis as the best subset having the smallest BIC value. The best fit multiple regression models is given below:

$$COD = 27.1735 + 0.65481BOD_5 + \mu_i \quad (11)$$

In the second step, the residuals from the MLR model were modeled by using the ANN model. A resilient back propagation with weight backtracking algorithm was used for training the multilayer perceptron until the optimum network architecture was achieved. The best fitting network selected, is composed of five inputs, seven hidden and one output neurons (in abbreviated form, ***N*** (5×7×1)), with five input nodes. The outputs of the neural network can be used as predictions of the error terms in the MLR model.

In the final step, if the forecast in step one and two are denoted by $\hat{L}_t$ and $\hat{N}_t$ respectively then the combine forecast will be;

$$\hat{y}_t = \hat{L}_t + \hat{N}_t \quad (13)$$

In the hybrid model algorithm, the input and output of the optimal parameter for COD were normalized between 0 and 1.

The comparative performance of MLR, ANN and hybrid model for predicting COD data is given in table 2. The comparison of the observed values and those estimated by the hybrid model for COD data set were plotted in figure 3. The figure shows the predictions and observations of the models for the testing and validation period. The results indicated the model prediction reasonably match the observed parameters.

Table 2 Comparative performance of MLR, ANN and hybrid model (COD data)

Model	RMSE	MAPE
MLR	18.25344	19.03417
ANN	41.72618	36.07570
MLR-ANN	18.07286	15.65634

.

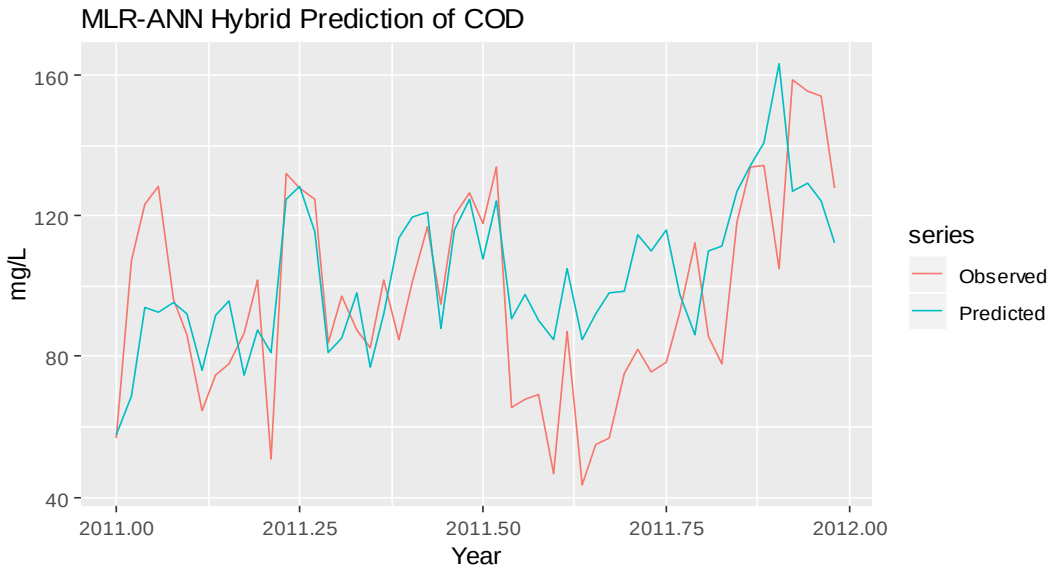

Figure 3. Comparison of the observed values and those estimated by the hybrid model (COD data)

The DO model results

In the first step, as in the previous section, data set of DO time series data is fitted with multiple regression models fitted to predict the weekly average dissolved oxygen (DO) as dependent parameter taking the weekly independent parameters as biochemical oxygen demand (BOD_5). BOD_5 was selected as the most significantly contributed parameters using stepwise forward regression analysis as the best subset having the smallest BIC value. The best fit multiple regression models is given below:

$$DO = 3.86043 - 0.014018BOD_5 + \mu_t \qquad (14)$$

In the second step, an ANN model building process was performed using resilient back propagation with weight backtracking algorithm for training the multilayer perceptron until the optimum network architecture was achieved The best fitting network selected, is composed of five inputs, six hidden and one output neurons (in abbreviated form, ***N*** (5×6×1)), with five input nodes. The outputs of the neural network can be used as predictions of the error terms in the MLR model.

In the final step, if the forecast in step one and two are denoted by $\hat{L}_t$ and $\hat{N}_t$ respectively then the combine forecast will be;

$$\hat{y}_t = \hat{L}_t + \hat{N}_t \quad (15)$$

In the hybrid model algorithm, the input and output of the optimal parameter for DO were normalized between 0 and 1.

The comparative performance of MLR, ANN and hybrid model for predicting DO data is given in table 3. The comparison of the observed values and those estimated by the hybrid model for DO data set were plotted in figure 4. The figure shows the predictions and observations of the models for the testing and validation period. The results indicated the model prediction reasonably match the observed parameters.

Table 3 Comparative performance of MLR, ANN and hybrid model (DO data)

Model	**RMSE**	**MAPE**
MLR	1.046285	38.581010
ANN	1.178251	43.340881
MLR-ANN	1.011545	37.328600

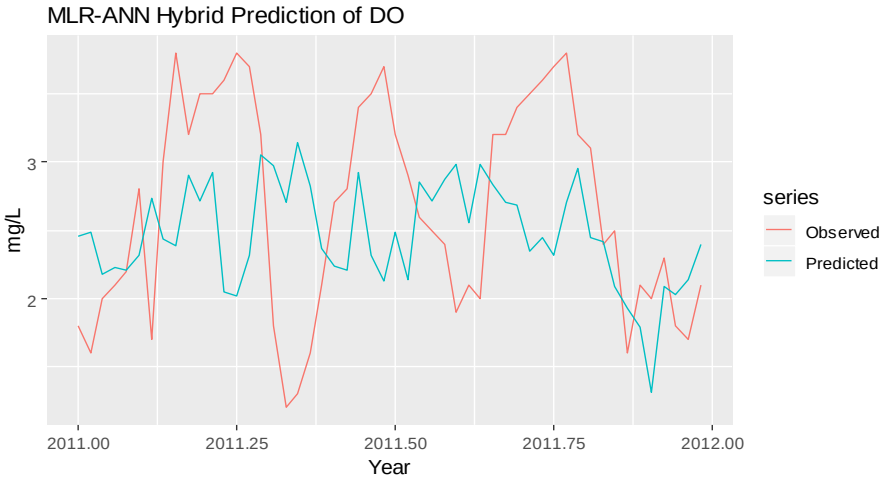

Figure 4. Comparison of the observed values and those estimated by the hybrid model (DO data)

Comparison with other models

In order to assess the ability of the hybrid model relative to MLR and ANN model, the MLR and ANN models were constructed for the same time series data (COD and DO). From the MLR models, which were examined with the

aim of describing the time series and producing a prediction, COD = 27.1735 + 0.65481BOD_5 + μ_i model was found to be appropriate for the COD data and DO = 3.86043 - 0.014018BOD_5 + μ_t model was found to be appropriate for DO data. These models present good forecasting accuracy, with regards to the MAPE and RMSE values for the test data set: COD = 27.1735 + 0.65481BOD_5 + μ_i model for COD data— MAPE 19. 034% and RMSE 18.2534 and DO = 3.86043 - 0.014018BOD_5 + μ_t model for DO data- MAPE 38.58% and RMSE 1.046 The MLR used in this case showed good adaptability for the same time series. The ANN model for COD data consists of five inputs and seven hidden and one neurons (*N* (5×7×1)) and for DO data consists of five inputs, six hidden and one output neuron (5×6×1)). The performances of MLR and ANN for DO and COD data in terms of the performance indices are presented in table 4 and table 5.

From Table 4 and 5, one can see that the hybrid model yielded more accurate results than both MLR and ANN models used separately. For COD data, the percentage improvements of the hybrid model over MLR in terms of MAPE and RMSE were 17.75% and 0.987% respectively, while compared to the ANN, the corresponding percentage improvements were 56.60% and 56.69% respectively. And for DO data, the percentage improvements of the hybrid model over MLR in terms of MAPE and RMSE were 3.246% and 2.98% respectively, while compared to the ANN, the corresponding percentage improvements were 13.87% and 14.14% respectively. The overall forecasting results of above mentioned models and improvement percentage of the hybrid model in comparison with those models were summarized in tables 4 and 5 respectively.

Table 4. Percentage improvement of the hybrid model incomparison with MLR and ANN (COD data).

Model	MAPE%	RMSE%
MLR	17.75	0.987
ANN	56.60	56.69

Table 5. Percentage improvement of the hybrid model in comparison with MLR and ANN (DO data).

Model	MAPE%	RMSE%
MLR	3.246	2.98
ANN	13.87	14.14

The results indicated that the hybrid model performed well for predicting COD and DO. It is clearly known that the MLR-ANN is able to predict the parameters with a high degree of accuracy as compared to the MLR and ANN models. In conclusion, MLR-ANN approach can produce the better prediction of COD and DO in the produced water than the MLR and ANN modeling approach.

Conclusion

This study used MLR, artificial neural network (ANN) and hybrid MLR-ANN models to predict the produced water quality time series. A new hybrid approach MLR-ANN model was developed to predict dissolved oxygen (DO) and chemical oxygen demand (COD) using continuous weekly data of water quality parameter. The results obtained showed that MLR model is more reliable and suitable when coupled with ANN model in predicting produced water quality time series. The hybrid model developed in this study can be used in predicting produced water quality of the flow station, which if tested in other flow stations; can be useful in the oil and gas industry as a working model, which will save cost of analyses, chemical, exposure and other environmental hazards. It can also be more useful in water quality management efforts to ensure that water resource is sustainable for the future. To examine the MLR-ANN model performance compared to MLR and ANN, statistical error measurements such as MAPE and RMSE were used. The hybrid model performance was compared relatively to the single models MLR and ANN. The least values of MAPE and RMSE give an improved performance in predicting produce water quality time series

References

Boah, D, K., Twum, S. B Khashei M.& Bijari M. (2011). A novel hybridization of artificial neural networks and ARIMA models for time series forecasting; *Applied Soft Computing* 11; 2664–2675.

EGASPIN (Environmental Guidelines and Standards for the Petroleum Industry in Nigeria) (2018). **(**Revised Edition). Lagos: Department of Petroleum Resources DPR).

Ehi-Douglas, O. M., Briggs, A. O. Ehimiaghe M., Howard, I. C.(2018). A New Approach to 'ECM' reporting through Quality Indexing. Being a paper presented at the International HSE Biennial Conference on the Oil and Gas Industry, 26 – 28th Nov. held at Eko Hotels, Lagos.

Gajendran, C. (2011) Water quality assessment and Prediction modelling of Nambiyar River basin, Tamil Nadu, India. A Ph. D thesis in Civil Engineering Anna University Chennai 600 025 India

Gonzalez Vilas, L., Spyrakos, E. and Torres Palenzuela, J.M. (2011). Neural Network

Estimation of Chlorophyll a from MERIS Full Resolution Data for the Coastal Waters of Galician rias (NW Spain). *Remote Sensing of Environment* , 115, 524-535.

Howard, I. C., Briggs A. O. and Muritala, I. K. (2012). Quality assessment of the surface water of a near-shore oilfield in the Niger Delta, Nigeria- 35th .*Nig.Jou*. *Contemp. Dev. Studies* 1(2) :1-5

Howard, I.C., Azuatola, O.D. & Abiodun, I.K. (2017) Investigation on impacts of artisanal refining of crude oil on river bed sediments. *Our Nature* 15 (1-2): 34-43

Muhamad, S. L., Mohd, N. A.R., Sugan. G. G., Nurul, H. Z., Razak. Z., MdSuffian. I., Idham, K., (2016). Improved the Prediction of Multiple Linear Regression Model Performance Using the Hybrid Approach: A Case Study of Chlorophyll-a at the Offshore Kuala Terengganu, *Terengganu. Open Journal of Statistics*, 6, 789-804.

Sahoo, G., Schladow, S. and Reuter, J. (2009) Forecasting Stream Water Temperature Using Regression Analysis, Artificial Neural Network, and Chaotic Non-Linear Dynamic Models. *Journal of Hydrology*, 378, 325-342.

Torres-Palenzuela, J.M., Vilas-Gonzalez, L. and Mosqura-Gimenez, A. (2005) Correlation between MERIS and in-Situ Data for Study of Pseudo-nitzschia spp. Toxic Blooms in Galician Coastal Area. Millpress, Rotterdam, 497-507.

Vilas, L. G., Spyrakos, E. and Palenzuela, J.M.T. (2010) Neural Network Estimation of Chlorophyll-a from MERIS Full Resolution Data for the Coastal Waters of Galician rias (NW Spain). *Journal Remote Sensing of Environmen*t, 115, 524-535.

Zhang, G. P., (2003). Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing*, 50, 159–175.