Journal of Sustainability Science and Management
Volume 13 Number 1, June 2018 : 27-37
ISSN: 1823-8556

IMPROVING THE PERFORMANCE OF ANN-ARIMA MODELS FOR PREDICTING WATER QUALITY IN THE OFFSHORE AREA OF KUALA TERENGGANU, TERENGGANU, MALAYSIA

MUHAMAD SAFIIH LOLA[1,4*], NURUL HILA ZAINUDDIN[2], MOHD TAJUDDIN ABDULLAH[3,4], VIGNESWARY PONNIAH[1], MOHD NOOR AFIQ RAMLEE[4], RAZAK ZAKARIYA[3], MD SUFFIAN IDRIS[3] AND IDHAM KHALILI[3]

[1]School of Informatics and Applied Mathematics, [3]School of Science Marine and Environment, [4]Kenyir Research Institute, Universiti Malaysia Terengganu, 21030 Kuala Nerus, Terengganu, Malaysia.
[2]Mathematics Department, Faculty of Science and Mathematics, Universiti Pendidikan Sultan Idris, 53900 Tanjong Malim, Perak, Malaysia.

**Corresponding Author: safiihmd@umt.edu.my*

Abstract: Developing a high degree of accuracy of time series forecasting model in sea water quality resources is very important. However, it is not easy due to the time series data of sea water quality that are complex and difficult to predict. An existing ARIMA time series model or an artificial neural network cannot solve this problem because of the linear and nonlinear relationships. Therefore, the hybrid model of artificial neural network and ARIMA is proposed and found effective to predict water quality data. In this study, the performance of the proposed models are investigated using water quality parameters such as water temperature, pH, salinity and dissolved oxygen in offshore of Kuala Terengganu. The results reveal that our proposed models perform much better in terms of producing smaller MAE and RMSE values, high correlation coefficients and also reduced error percentage for all parameters up to the maximum of 87.87% in both MAE and RMSE as compared to ARIMA and ANN models. Therefore, the proposed models can be used as the best alternative model for forecasting activities, particularly when higher degrees of accuracy in forecasting become a priority.

Keywords: Prediction, Artifical Neural Network, ARIMA, water quality, accuracy, sustainability

Introduction

A large amount of research has been done using time series models such as Multi Linear Regression (MLR), Principle Component Analysis (PCA), Autoregressive Moving Average (ARMA), Autoregressive Integrated Moving Average (ARIMA) and Seasonal Autoregressive Integrated Moving Average (SARIMA) (see, McKenzie, 1984; Hipel & Mcleod, 1994; Cornillon *et al.,* 2008; Mohd Zamri *et al.,* 2009; Rita *et al.,* 2013; Ibrahim *et al.,* 2010; Muhamad Safiih *et al.,* 2017a; Syerrina *et al.,* 2017; Samsuri *et al.,* 2017; Muhamad Safiih *et al.,* 2017b). However, the major weakness of these models is generated from a linear component which has difficulties in capturing the nonlinear component. On the other hand, Artificial Neural Network (ANN) is nonlinear in nature and is influenced by the behaviour of neurons in them. It can approximate the function to a satisfying level of accuracy. Hence, a hybrid of ARIMA and artificial neural network back propagation model is proposed. The use of hybrid models has the advantage of capturing patterns of data sets as well as improving the prediction accuracy (Luxhoj, *et al.,* 1996; Balkin & Ord 2000; Medeiros & Veiga, 2000; Tseng *et al.,* 2002; Zhang, 2003; Armano *et al.,* 2005; Taskaya & Casey, 2005; Chen & Wang 2007; Kim & Shin, 2007; Isinkaye *et al.,* 2015; Qiu & Song, 2016). The reason for using this hybrid approach is mainly based on the complexity of water quality real data sets and single model approaches would not be sufficient to determine the patterns very well. In this study, the hybrid NNARIMA models were developed in order to predict the water quality time series and evaluate its performance (Zhang, 2003).

Materials and Methods

Study Area

This research was conducted in the offshore around Kuala Terengganu i.e. Kuala Terengganu,

Kg. Marang, Kg. Setiu and Kuala Besut (Figure 1). The data were collected from 30th April until 3rd May, 2015, and 126 observation were obtained from 26 sampling stations at different depths.

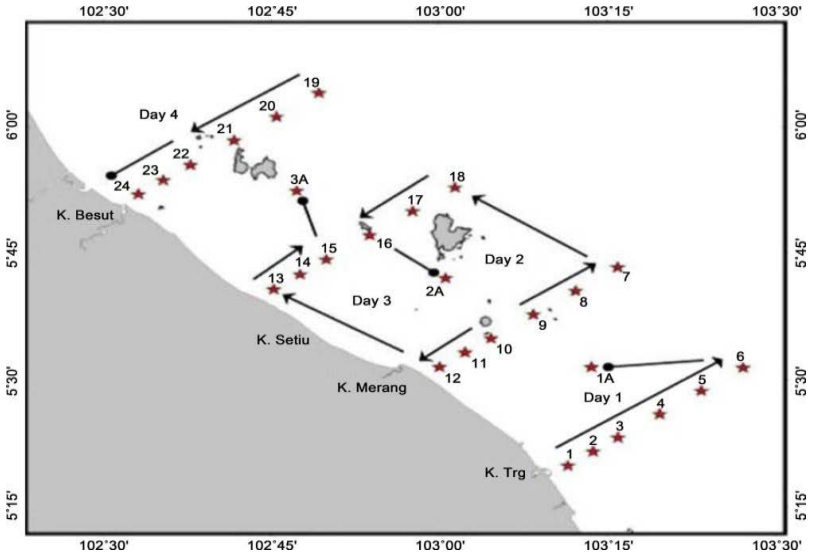


Figure 1: Study area in the offshore at Kuala Terengganu, Terengganu, Malaysia

ARIMA Modelling Approach

ARIMA model is represented by a general term of ARIMA *(p, d, q)* as follows:

$$y_t = \mu + \emptyset_1 y_{t-1} + \emptyset_2 y_{t-2} + \cdots + \emptyset_p y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \cdots - \theta_q \varepsilon_{t-q} \quad (1)$$

where *p* and *q* are the number of autoregressive terms and the number of lagged forecast errors in the prediction equation, respectively. The number of *p, d* and *q* are obtained by looking at the autocorrelation function (ACF) and partial autocorrelation function (PACF) plots. The ARIMA modelling approach consists of three steps: model identification, parameter estimation and diagnostic checking. (see, McKenzie, 1984; Hipel & Mcleod, 1994; Cornillon *et al.,* 2008; Mohd Zamri *et al.,* 2009; Rita *et al.,* 2013; Ibrahim *et al.,* 2010; Muhamad Safiih *et al.,* 2017a; Syerrina *et al.,* 2017; Samsuri *et al.,* 2017; Muhamad Safiih *et al.,* 2017b). Model identification consists of two steps: determining whether the series are stationary and examining the ACF and PACF functions. The model with the minimum Akaike's Crriterion is chosen as the best fit model (Qiu & Song, 2016).

Artificial Neural Network Modelling

One advantage of the neural network model compared to other nonlinear models is the universal estimators that can emulate the class of functions with a high level of accuracy (Zhang *et al.,* 1998). The strength of their estimation is parallel with the information from the processing data. The initial assumption is not necessary to establish this model during the model building process. Instead, the network model is largely determined by the characteristics of the data. A feed forward circuit (feed forward) is among the hidden layer that is widely used to model time series and forecasting. This model is characterized by three mobile network layers of simple processing units linked by a series of a cycles. The relationship between output (y_t) and input ($y_{t-1} \ldots\ldots y_{t-p}$) has the following mathematical representation (Khashei & Bijari, 2011):

$$y_t = w_o + \sum_{j=1}^{q} w_j . g(w_o + \sum_{i=1}^{p} w_{ij} . x_{t,i}) + \varepsilon_t, \quad (2)$$

whereby w_{ij} (i=0,1,2,…,p) and w_j (j=0,1,2,…,q) are parameters of the model which are also called the connection weights. The terms p, q, ε_t, w_0 and $w_{0,j}$ and g are the number of input nodes, the number of hidden nodes, error term and weights of the arcs leaving from the bias terms and sigmoid equation, respectively. Activation functions consist of a few forms and represented by the condition of neurons in the network (Khashei & Bijari, 2011) as follows:

$$g(x) = \frac{1}{1+e^{(-x)}} \quad (3)$$

$$Tanh(x) = \frac{1 - e^{(-2x)}}{1 + e^{(-2x)}} \quad (4)$$

Normally, an artificial neural network model in Eq. (1) can be written in terms of nonlinear functional mapping from past observations i.e., $y_{t-1}, y_{t-2}, \ldots\ldots y_{t-p}$ to the future values of y_t, which is:

$$y_t = f\left(y_{t-1}, y_{t-2} \cdots y_{t-p}, \alpha\right) + \epsilon_t \quad (5)$$

where α is a vector for all parameters, $f(\square)$ is a function determined by the network structure and connection weights and ϵ_t is the error term. The general structure of neural network is shown in Figure 2.

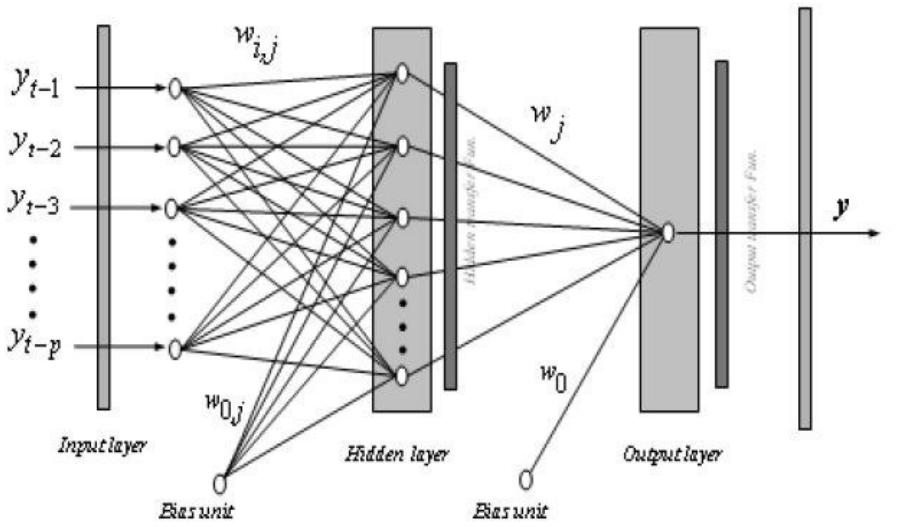


Figure 2: General structure of neural network structure and its connection weight

Hybrid Model of NNARIMA

In this study, a hybrid between linear and nonlinear models was proposed in order to yield an accuracy of the forecasting time series results named as neural network autoregressive integrated moving average (NNARIMA) model as in Eq. (6). In Eq. (6), we consider the time series model as a function for both linear and non-linear components.

$$y_t^* = l_t^* + nl_t^* \quad (6)$$

where l_t^* is linear while nl_t^* is nonlinear. Components of Eq. (6) involved 2 phases, i.e., phase 1 and phase 2. Phase 1 is to obtain a linear model component of ARIMA model. The error of the phase 1 contains nonlinear relationship or could also be linear relationships which linear models could not capture or solve (Kashei & Bijari, 2011). Therefore, the error at time *t* is represented by:

$$\epsilon_t = y_t^* - \hat{y}_t \quad (7)$$

The phase 2 is nonlinear component where in this phase, the predictable and the error values in phase 1 were used. In this phase, one of the artificial neural networks which are the multilayer perceptron (MLP) was used to model nonlinear relationship. In order to produce a hybrid results, a simultaneous linear model was used, i.e., remains error linear models as well as the relationship of linear and nonlinear of the original data. Therefore, errors can be modelled using neural network to identify the nonlinear relationship. With *n* input nodes, the neural network model for the error is as follows:

$$\epsilon_t = f(\epsilon_{t-1}, \epsilon_{t-2}, \cdots, \epsilon_{t-n}) + \omega_t \quad (8)$$

where *f* is a nonlinear function which is dependent on the neural network and ω_t is a random variable. Then, the combined forecasting model would be written as:

$$yF_t = lF_t + nF_t \quad (9)$$

Comparison of ARIMA, ANN and the hybrid NNARIMA Models

Both linear and nonlinear models were used to analyse the data, although linearity has been found in this series. Only one step forward predictions are considered. Two key performance indicators, the mean absolute error (MAE) and mean square error (RMSE), which are calculated from the following equation, will be used to measure the performance of the predicted models (Rahimi 2016; Mohd Zamri *et al.,* 2009; Muhamad Safiih *et al.,* 2009; Ibrahim *et al.,* 2010; Muhamad Safiih 2013; Nurul Hila & Muhamad Safiih, 2016; Nurul Hila *et al.,* 2016; Muhamad Safiih *et al.,* 2017a; Muhamad Safiih *et al.,* 2017b). RMSE investigates the overall performance of the models while MAE evaluates the models.

$$\text{MAE} = {|\sum_{i=1}^{n} y_i - yF_t|}/{n} \quad (10)$$

$$\text{RMSE} = \sqrt{{\sum_{i=1}^{n} (y_t - yF_t)^2}/{n}} \quad (11)$$

Results and Discussion

ARIMA Modelling

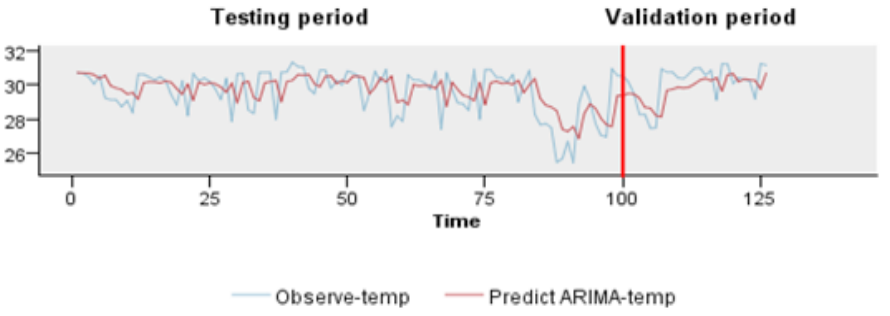
In this study, several steps were made to choose the ideal ARIMA model parameters that satisfy diagnostic checking of the residuals. In the identification stage, the autocorrelation function (ACF) and partial autocorrelation function (PACF) were used to study the stationary nature of the data and to determine the possible best fit models. The best fit model was then determined by using the Akaike's Criterion (AIC) for all the parameters including water temperature, pH, salinity and dissolved oxygen. The models were then checked for adequacy by analysing the independence of the residuals.

Table 1: Best fit model for all parameters

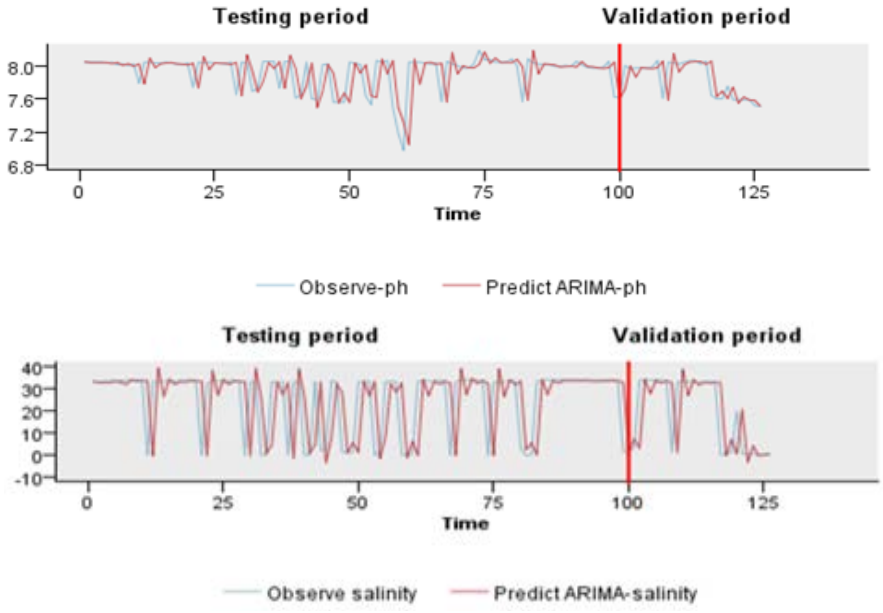
Parameters	Type of Model	MSE	AIC
Water Temperature	ARIMA(1,1,1)	1.364	0.3422
pH	ARIMA(2,1,2)	0.03571	-3.2688
Salinity	ARIMA(0,1,2)	189.2	5.2749
Dissolved Oxygen (DO)	ARIMA(1,1,1)	19.20	2.9871

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 modelling algorithm in the ARIMA model which is unsatisfactory in identifying and predicting nonlinear time series of water quality parameters.

(i) Temperature



(ii) pH



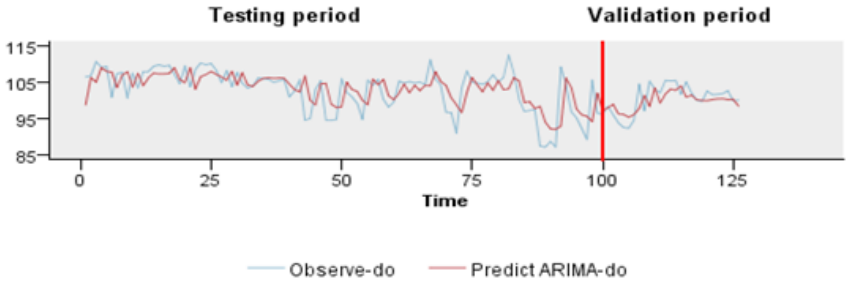
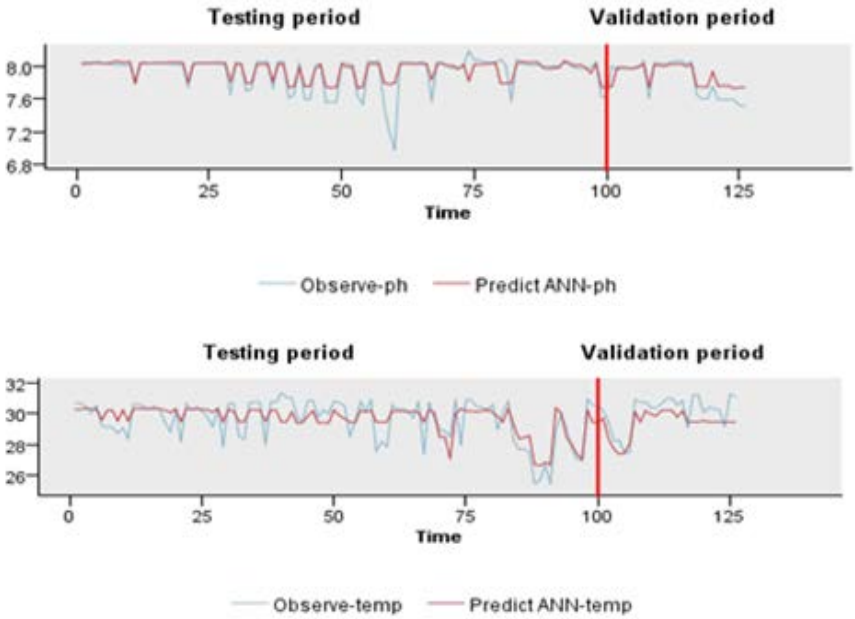


Figure 3: The ARIMA model for water quality parameters:-
(i) temperature, (ii) pH, (iii) salinity, (iv) dissolved oxygen (DO)

ANN Modelling Approach

A neural network was developed to predict the optimal model to forecast water quality time series. In this approach, water temperature, pH, salinity and dissolved oxygen were used as the input data. It is important to emphasize that the target would be changed simultaneously according to the input data. For example, if we need to predict water temperature, water temperature is the target and the other parameters would be set as the input data. There are 3 partitions in the neural network model which comprises of training, testing and validation. During training, the inputted data will be selected in the network and it will customize itself based on the error contained in the model. Next was testing, which is an independent step to test the viability of models. Finally, validation is used to measure the networks ability to generalize and evaluated as stopping criteria for training sample. The total data used in this study is 70% for training, 20% for testing and 10% for validation purposes. The results shown in Figure 4 indicate that the neural network developed was able to detect the pattern in water quality parameters. This result provides good prediction of the daily variations in the data because the predicted graph follows closely the observed graph.

(i) Temperature



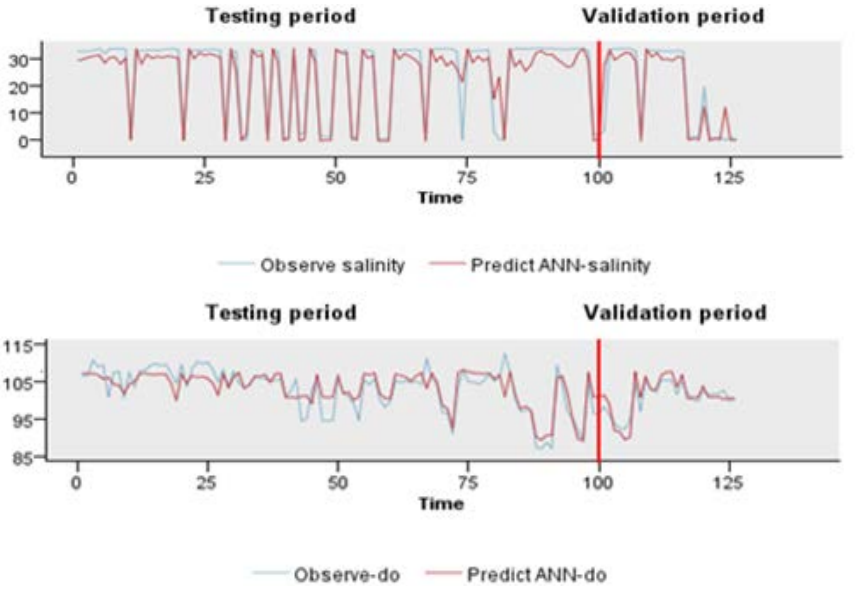
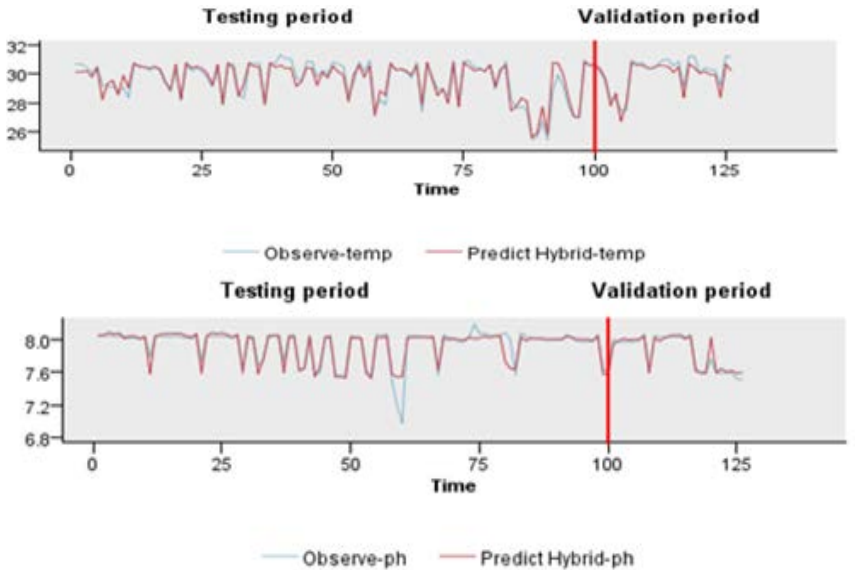


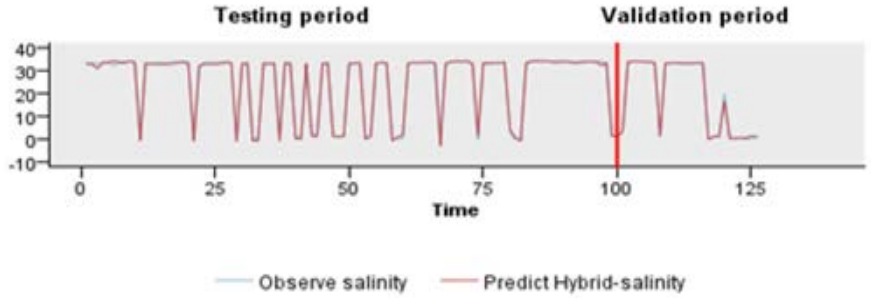
Figure 4: The ANN model for water quality parameters:-
(i) temperature, (ii) pH, (iii) salinity, (iv) dissolved oxygen

The Hybrid Modelling Approach

The testing and validation period for all parameters based on the hybrid model are shown in Figure 5 (i),(ii), (iii) and (iv). The figures show that the predicted data follow closely the observed data for all water quality parameters. The predicted data were able to identify the pattern of the input data to provide desired and valid predictions better than the ARIMA and ANN models. Hence, hybrid models provide the most reliable prediction when compared to single models.

(i) Temperature





(iv) Dissolved Oxygen (DO)

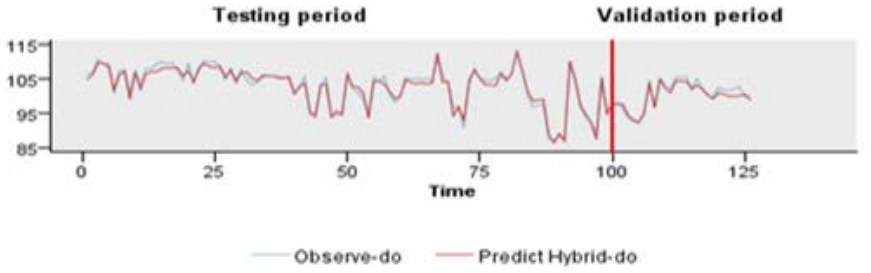


Figure 5: Hybrid model for water quality parameters:-
(i) temperature, (ii) pH, (iii) salinity, (iv) dissolved oxygen (DO)

Comparative Performance of the Models

In order to evaluate the performance of the developed models, this study use three methods namely, (i) statistical performance evaluation criteria which are Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) as in Eqs. (10) and (11), (ii) the correlation coefficient and, (iii) MAE and RMSE, reduced percentage error. For the statistical performance evaluation criteria, the comparative performance of ARIMA, ANN and NNARIMA models for all parameters are shown in Table 2. From Table 2, it is proven that the hybrid model of NNARIMA has the lowest MAE and RMSE for all parameters compared to ARIMA and ANN models. The results indicate that our proposed model produces highly accurate forecasting time series water quality data as compared to the single ANN and ARIMA models.

Table 2: Model performance using MAE and RMSE

	Temperature °C		pH		Salinity (ppt)		DO (ppm)	
Models	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
ARIMA	0.9100	1.1585	0.1303	0.2134	9.5102	18.9201	3.1811	4.2905
ANN	0.7031	3.6207	0.07206	0.1323	3.4192	6.0801	2.2211	2.8948
NNARIMA	0.3275	0.2935	0.0431	0.0889	3.0437	4.8789	1.7780	0.5204

The performance of the ARIMA, ANN and the proposed models are shown in terms of the correlation coefficient, i.e., the strength of the linear relationship between the observation and prediction for all parameters, as shown in Table 3. The correlation coefficient values of ARIMA are linear positive but weak and moderate for temperature (°C), pH, salinity (ppt) and DO (ppm), respectively. This values are not satisfactory in common model applications. This is due to the limitation of the linear modelling algorithm in ARIMA model which is unsatisfactory in identifying and predicting nonlinear time series of water quality data. For

the ANN model, the correlation coefficients are 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. All the correlation values are above 0.75 which means that they are satisfactory in identifying and predicting nonlinear time series of water quality data. However, the correlation coefficient values of the proposed model are very strong, linear and positive. These indicate that the hybrid linear and nonlinear model is satisfactory in common model applications. In other words, the proposed model was able to detect and identify the pattern of water quality parameters to provide desired and valid predictions better than the ARIMA and ANN models.

Table 3: The performance of models using correlation coefficients

Models	Correlation Coefficient			
	Temperature (°C)	pH	Salinity (ppt)	DO (ppm)
ARIMA	0.483	0.457	0.392	0.631
ANN	0.759	0.846	0.947	0.858
NNARIMA	0.944	0.906	0.999	0.980

Meanwhile, Table 4 shows that the MAE reduced error percentage decreased by 22.74%, 44.68%, 64.05% and 30.18% for water temperature, pH, salinity and DO when ANN models were used. Applying the hybrid models, the MAE reduced error percentage decreased by 53.42%, 40.18%, 10.89% and 19.94% in the MAE values when hybrid model is used for water temperature, pH, salinity and DO, respectively. Comparatively, the RMSE reduced error percentage decreased by 10.52%, 38.03%, 61.81%, and 32.55% for water temperature, pH, salinity and DO when the ANN models were used. When the hybrid models were used, the RMSE reduced error percentage decreased by 74.67%, 58.34%, 74.21%, and 87.87% for the parameters.

Table 4: MAE and RMSE Reduced Percentage Error for all parameters

Parameters	MAE Reduced Error (%)		RMSE Reduced Error (%)	
	ARIMA-ANN	ARIMA-HYBRID	ARIMA-ANN	ARIMA-HYBRID
Temperature (oC)	22.74	64.01	10.52	74.67
pH	44.68	66.92	38.02	58.34
Salinity(ppt)	64.05	68.00	61.81	74.21
DO (ppm)	30.18	44.11	32.53	87.87

Conclusion

This study used ARIMA, neural network and hybrid NNARIMA models to predict the water quality time series. The hybrid model developed would be able to utilize the benefits of both the traditional methods and ANN. The results obtained show that ANN model is more reliable and suitable when coupled with ARIMA model in predicting water quality time series. The hybrid 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, two accuracy measures, the RMSE and MAE, were formulated in order to demonstrate the performance of the developed models in predicting water quality time series. The hybrid model performance was compared relatively to the single models ANN and ARIMA. The least values of MAE and RMSE give an improved performance in predicting water quality time series.

Acknowledgements

Authors express gratitude to all participants who are involved in this study from data gathering to the completion especially to the School of Informatics and Applied Mathematics (SIAM) and School of Marine Science and Environment

(SSME), University Malaysia Terengganu (UMT), also to Kenyir Research Institute, University Malaysia Terengganu (UMT) for their sponsorship. The author claims no conflicts of interest in this study.

References

Armano, G., Marchesi, M., & Murru, A. (2005). A hybrid genetic-neural architecture for stock indexes forecasting. *Information Sciences,* 170: 3–33.

Balkin, S. D., & Ord, J. K. (2000). Automatic neural network modelling for univariate time series. *International Journal of Forecasting,* 16: 509–515.

Chen, K. Y., & Wang, C. H. (2007). A hybrid SARIMA and support vector machines in forecasting the production values of the machinery industry in Taiwan. *Expert Systems with Applications,* 32: 54–264.

Cornillon, P., Imam, W., & Matzner, E. (2008). Forecasting time series using principal component analysis with respect to instrumental variables. *Computational Statistics and Data Analysis,* 52: 1269–1280.

Nurul Hila, Z. and Muhamad Safiih, L. (2016). The performance of BB-MCEWMA model: Case study on sukuk Rantau Abang Capital Berhad Malaysia. *International Journal of Applied, Business and Economic Research,* 14(2): 639-653.

Nurul Hila, Z., Muhamad Safiih, L. and Nur Shazrahanim, K. (2016). Modelling moving centerline exponentially weighted moving average (MCEMA) with bootstrap approach: Case study on sukuk musyarakah of Rantau Abang Capital Berhad, Malaysia. *International Journal of Applied, Business and Economic Research,* 14(2): 621-638.

Hipel, K. W., & McLeod, A. I. (1994). *Time Series Modelling of Water Resources and Environmental Systems*. Amsterdam: Elsevier.

Ibrahim, M.Z., Zailan, R., Ismail, M., Lola, M.S., (2010). Time-series Analysis of Pollutants in East Coast Peninsular Malaysia. *Journal of Sustainability Science and Management,* 5(1): 57-65.

Isinkaye, F. O., Folajimi, Y. O., & Ojokoh, B. A. (2015). Recommendation systems: Principles, methods and evaluation. *Egyptian Informatics Journal,* 16(3): 261-273.

Kim, H., & Shin, K. (2007). A hybrid approach based on neural networks and genetic algorithms for detecting temporal patterns in stock markets. *Applied Soft Computing,* 7: 569–576.

Khashei, M. & Bijari, M. (2011). An artificial neural network (p,d,q) model for time series forecasting. *Expert Systems with Applications*. 37: 479-489

Luxhoj, J. T., Riis, J. O., & Stensballe, B. (1996). A hybrid econometric-neural network modelling approach for sales forecasting. *International Journal of Production Economics,* 43: 175–192.

McKenzie, E.D. (1984). General exponential smoothing and the equivalent ARMA process, *J. Forecasting* 3: 333–344.

Medeiros, M. C., & Veiga, A. (2000). A hybrid linear-neural model for time series forecasting. *IEEE Transaction on Neural Networks,* 11(6): 1402–1412.

Mohd Zamri, I. Roziah, Z., Marzuki, I., Muhamad Safiih, L. (2009), Forecasting and Time Series Analysis of Air Pollutants in Several Area of Malaysia. *American Journal of Environmental Sciences,* 5(5): 625-632.

Muhamad Safiih, L. Abu Osman, M.T. and Anton, A. K. (2009), Semi-Parametric of Sample Selection Model Using Fuzzy Concepts. *Journal of Modern Applied Statistical Methods*. 8(2): 547-559.

Muhamad Safiih, L (2013), Fuzzy Parametric

Sample Selection Model: Monte Carlo Simulation Approach. *Journal of Statistical Computation and Simulation;* 83(6): 992-1006.

Muhamad Safiih, L, Nurul Hila, Z. Mohd Noor Afiq, R, Muhamad Na'eim, A.R, Mohd Tajuddin A. (2017a), Improvement of Estimation Based on Small Number of Events Per Variable (EPV) using Bootstrap Logistics Regression Model. *Malaysian Journal of Fundamental and Applied Sciences,* 13(4): 693-704.

Muhamad Safiih, L, Nurul Hila, Z. Mohd Noor Afiq, R, Hizir, S. (2017b). Double Bootstrap Control Chart for Monitoring SUKUK Volatility at Bursa Malaysia. *Jurnal Teknologi,* 79 (6): 149-157.

Qiu, M., & Song, Y. (2016). Predicting the Direction of Stock Market Index Movement Using an Optimized Artificial Neural Network Model. *PLoS ONE,* 11(3): 1-11.

Rahimi, A. (2016). A methodology approach to urban land-use change modelling using infill development pattern-a case study in Tabriz, Iran. *Ecological Process,* 5(1): 3-15.

Rita, S., Yony, H., Rubiyanto, Fakhri, A.M., Madzlan, A., (2013). Multiple Linear Regression (MLR) Modeling of Wastewater in Urban Region of Southern Malaysia. *Journal of Sustainability Science and Management,* 8(1): 93-102.

Samsuri, A. Marzuki, I. Si, Y.F. (2017), Multiple Linear Regression (MLR) models for long term Pm10 concentration forecasting during different monsoon seasons. *Journal of Sustainability Science and Management,* 12(1): 60-69.

Syerrina, Z. Naeim, A.R. Muhamad Safiih L.and Nuredayu, Z. (2017). Explorative Spatial Analysis of Coastal Community Incomes in Setiu Wetlands: Geographically Weighted Regression. *International Journal of Applied Engineering Research.* 12 (18): 7392-7396

Taskaya, T., & Casey, M. C. (2005). A comparative study of autoregressive neural network hybrids. *Neural Networks,* 18: 781–789.

Tseng, F. M., Yu, H. C., & Tzeng, G. H. (2002). Combining neural network model with seasonal time series ARIMA model. *Technological Forecasting and Social Change,* 69: 71–87.

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

Zhang, G., Patuwo, B. E., & Hu, M. Y. (1998). Forecasting with artificial neural networks: The state of the art. *International Journal of Forecasting,* 14: 35–62.