academicJournals

Vol. 9(9), pp. 191-200, September 2017
DOI: 10.5897/IJWREE2017.0740
Article Number: 414D80C65672
ISSN 2141-6613
Copyright © 2017
Author(s) retain the copyright of this article
http://www.academicjournals.org/IJWREE

International Journal of Water Resources and Environmental Engineering

Full Length Research Paper

Application of stochastic models in predicting Lake Malawi water levels

Rodgers Makwinja[1,3]*, Titus Phiri[2,3], Ishmael B. M. Kosamu[1] and Chikumbusko C. Kaonga[1]

[1]Department of Physics and Biochemical Sciences, University of Malawi, The Polytechnic, Private Bag 303, Chichiri, Blantyre 3, Malawi.
[2]Senga Bay Fisheries Research Unit, P. O. Box 316, Salima, Malawi.
[3]Department of Fisheries, P. O. Box 593, Lilongwe, Malawi.

Received 15 July, 2017; Accepted 3 August, 2017

Stochastic models have proven to be practically fundamental in fields such as science, economics, and business, among others. In Malawi, stochastic models have been used in fisheries to forecast fish catches. Nevertheless, forecasting water levels in major lakes and rivers in Malawi has been given little attention despite the availability of ample historical data. Although previous multichannel seismic surveys revealed the presence of low stands (sediment bypass zone) in Lake Malawi indicating that since the beginning of its formation, important water level fluctuations have been occurring, these previous surveys failed to predict and highlight much more clearly the status of these levels in the future. Therefore, the main objective of the study was to fill these research gaps. The study used *Autoregressive* (AR), *Moving Average* (MA), Autoregressive Moving Average (ARMA) and *Autoregressive Integrated Moving Average* (ARIMA) processes to select the appropriate stochastic model. Based on lowest *Normalized Bayesian Information Criterion* (NBIC), *Root Mean Square Error* (RMSE), *Mean Absolute Percentage Error* (MAPE), *Mean Forecast Error* (MFE), *Maximum Absolute Percentage Error* (MAXAPE), *Maximum Absolute Error (MAXAE*), and *Mean Absolute Error* (MAE) - ARIMA (0,1,1) model is found suitable for forecasting Lake Malawi water levels which shows negative trend up to 2035. The study further predicted that Lake Malawi water levels will decrease from the current average level of 472.97 m to an average of 468.63 m for the next 18 years (up to 2035).

Key words: Forecasting, Lake Malawi, modelling, stochastic, time series, water levels.

INTRODUCTION

Time series stochastic process is a set of random variables $\{z_t\}$ where the index t takes values in a certain set C (Alonso and Garcia-Martos, 2012). The process provides attractive modeling techniques for forecasting and planning because historical data can be used to a reasonable level of certainty (Box et al., 2015). The model deals with a sequential set of data points, measured typically over successive times. It is

*Corresponding author. Email: makwinjarodgers@gmail.com.

mathematically defined as a set of vectors $x(t)t = 0,1,2$ ...where t represents the time elapsed (Hipel and McLeod, 1994). The variable *x(t)* is treated as a random variable and the measurements taken during an event in a time series are arranged in a proper chronological order. The principles of stochastic process are to describe and summarize time series data, fit low-dimensional models and make forecast (Box et al., 2015). Time series data have many forms and represent different stochastic processes. According to literature, *Autoregressive (AR)* and *Moving Average (MA)* models have been widely and commonly used in different fields (Box and Jenkins, 1970; Hipel and McLeod, 1994). The combination of *AR* and *MA* models forms *Autoregressive Moving Average* (ARMA). However, ARMA model only works with stationary time series data. Thus, from application viewpoint, ARMA models are inadequate to properly describe non-stationary time series, frequently encountered in practice. For this reason, the *Autoregressive Integrated Moving Average (ARIMA*) model (Box and Jenkins, 1970) is proposed. The ARIMA model is a generalisation of an ARMA model which includes the case of non-stationarity as well (Chang et al., 2012). The model was first proposed by Box and Jenkins in the early 1970s and was often termed as Box-Jenkins models (Stuffer and Dhumway, 2010). Because ARIMA model is relatively systematic, flexible and can grasp more original time series information, it is widely used in meteorology, engineering technology, marine, economic statistics, prediction technology, hydrology and water resources studies (Yevjevich, 1972; Aksoy et al., 2013; Cryer and Chan, 2008; Kantz and Schreiber, 2004).

In Malawi, ARIMA model has been commonly used in fisheries to forecast fish catches (Zindi et al., 2016; Lazaro and Jere, 2013; Singini et al., 2012; Mulumpwa et al., 2016). Nevertheless, forecasting water levels in major lakes and rivers in Malawi has been given little attention despite the availability of ample historical data. On the same note, although previous multichannel seismic surveys (Scholz and Rozendahl, 1988; Johnson and Davis, 1989; De Vas, 1994) revealed the presence of low stands in Lake Malawi indicating that since the beginning of its formation, important water level fluctuations have been occurring, these previous surveys failed to predict and highlight much more clearly the status of these levels in the future. Consequently, the present study was designed to fill these research gaps.

MATERIALS AND METHODS

Study area and physiography

The study was conducted in Lake Malawi, located at the southern end of the Great Rift Valley region. It is an elongated lake surrounded by mountains with highest elevations to the north. Figure 1 shows that the boundaries of Lake Malawi cross Mozambique and Tanzania with an outlet in the southern end. The lake is ranked as the ninth largest and third deepest freshwater lake in the world with an estimated total area of 28,750 km^2 and a volume of about 7725 km^3. The Shire River is the outlet of Lake Malawi and flows approximately 410 km from Mangochi to Ziu Ziu in Mozambique, where it drains into Zambezi River (Shela, 2000). According to Shela (2000), the physiography of upper Shire has offered opportunities for regulating river flows and subsequently lake levels, with possible expansion. The middle section of Shire River is estimated to be 80 km and is very steep characterised by rock bars and outcrops with water falls of about 370 m.

Data collection and time series model description

Lake Malawi has been there over the years. Literature has shown that in early 1924, Dixey attempted to understand the hydrology of Lake Malawi (Dixey, 1924). However, he failed due to lack of hydrological data (Dixey, 1924). Later in the years, the fear of period of no outflow by authorities greatly forced them to seriously monitor the Lake levels (Drayton, 1984). Department of Water Resources seriously embarked on collection of water levels data later in the years; however, the data collected from 1950s to somewhere around 1980s were too complex and the quality was too inconsistent. Similar observation was reported by Kaunda (2015). Because of these past data anomalies, the present study analysed the univariate time series data of Lake Malawi water levels from 1985 to 2016 period. Figure 1 shows that the Department of Water Resources collects water levels data from three stations along the lake shore (Chilumba, Nkhatabay and Monkey Bay). The water level is normally the average of three records ignoring the water level gradient which is between the north and south tip of the lake (Kumambala, 2010).

Application of stochastic models

The study used two linear time series models known as *Autoregressive* (AR) (Box and Jenkins, 1970) and *Moving Average* (MA) (Zhang, 2003) models. These models were combined to form *Autoregressive Moving Average* (ARMA) (Cochrane, 1997). The combination of these two models were based on famous Box-Jenkins principle (Box and Jenkins, 1970) also known as the Box-Jenkins models.

Autoregressive Moving Average (ARMA) Model

An ARMA (p,q) model which is a combination of AR(*p*) and MA (*q*) models was developed. In an AR (*p*) model, the future value of a variable was assumed to be a linear combination of (*p*) past observations and a random error together with a constant term. Mathematically, the AR (*p)* model (Hipel and McLeod, 1994) is expressed as

$$\boldsymbol{\gamma_t = c + \sum_{i=1}^{p} \varphi_i \gamma_{t-i} + \varepsilon_t = c + \varphi_1 \gamma_{t-1} + \varphi_2 \gamma_{t-2} .. \varphi_p \gamma_{t-p} + \varepsilon_t} \qquad (1)$$

where γ_t and ε_t are the actual value and random error at time period *t*, respectively,φ_i (i=1, 2 ... *p)* are model parameters and *c* is a constant. Just as an AR (*p*) model regress against past values of the series, an MA (*q*) model uses past errors as the explanatory variables. The MA (*q*) is given by $\boldsymbol{\gamma_t}$ (Hipel and McLeod, 1994) and is expressed as:

$$\boldsymbol{\gamma_t = \mu + \sum_{j-1}^{q} \theta_j \varepsilon_{t-j} + \varepsilon_t = \mu + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} .. + \theta_q \varepsilon_{t-q} + \varepsilon_t} \ldots \qquad (2)$$

Here, μ is the mean of the series, *j =1, 2….* are model parameters

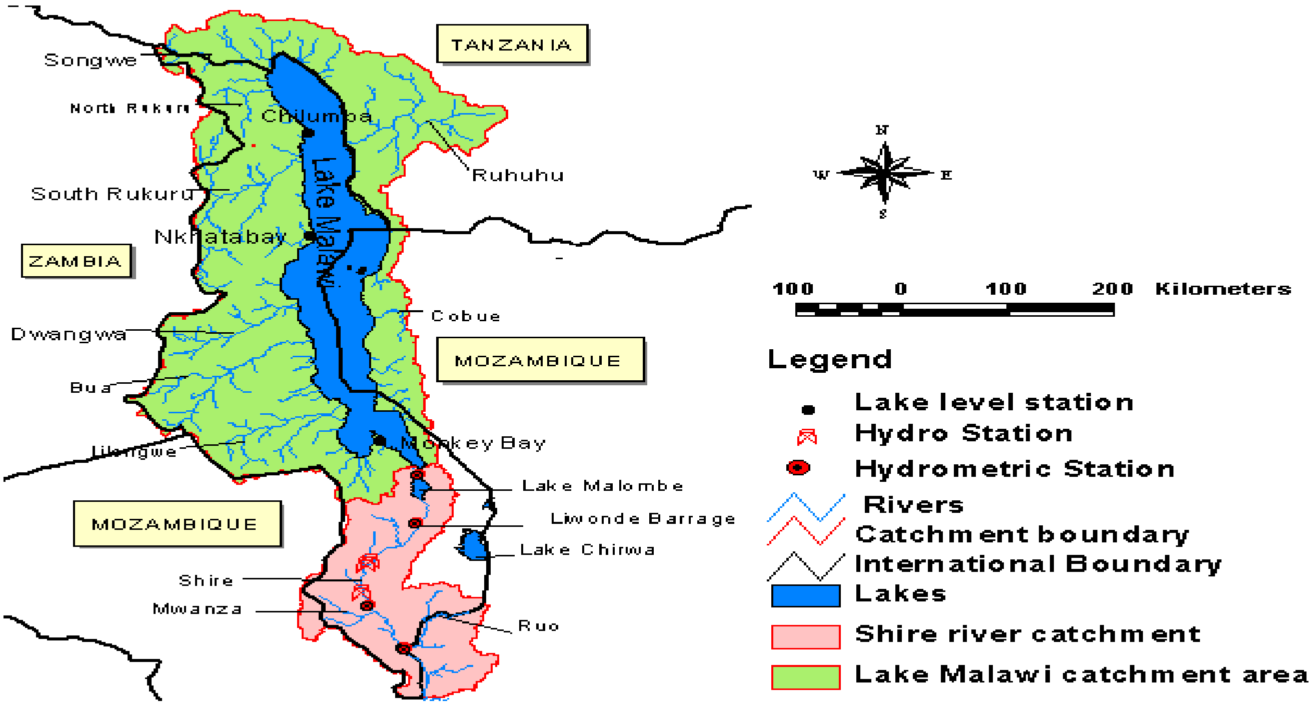

Figure 1. Map of Malawi showing Lake Malawi-Shire River system (GoM, 2005).

and q being the order of the model. The random shocks are assumed to be white noise (Hipel and McLeod, 1994) process. *Autoregressive* (AR) and *Moving Average* (MA) models were combined together to form a general and useful class of time series models known as the ARMA model. Mathematically, an ARMA (p, q) model is presented as (Cochrane, 1997):

$$\boldsymbol{\gamma_t = c + \varepsilon_t \sum_{i=1}^{p} \varphi_i \gamma_{t-i} + \sum_{j-1}^{q} \theta_j \varepsilon_{t-j}} \tag{3}$$

where the model orders p, q refers to p *autoregressive* and q *moving average* terms. Usually, ARMA models are manipulated using the lag operator notion. The lag operator is defined as $Ly_t = Y_{t-1}$. Polynomial of lag operators are used to represent ARMA models as follows:

AR(*p)* model: $\varepsilon_t = \varphi(L)\gamma_t$ (4)

MA(*q*) model:$\gamma_t = \theta(L)\varepsilon_t$ (5)

ARMA (*p*, *q*) model: $\varphi(L)\gamma_t = \theta(L)\varepsilon_t$ (6)

where

$$\varphi(L) = 1 - \sum_{i=1}^{p} \varphi_i L^i \text{ and } \emptyset(L) = 1 + \sum_{j=1}^{q} \theta_j L_j \tag{7}$$

Stationary analysis

When an AR (*p*) process is presented as:$\varepsilon_t = \varphi(L)\gamma_t$, the $\varphi(L) = 0$ is known as the characteristic equation for the process. Box and Jenkins (1970), proved that a necessary and sufficient condition for the AR (p) process to be stationary is that all roots of the characteristic equation must fall outside the unit circle. It is very important to note that ARMA models can only be used for stationary time series data. The fact that Lake Malawi water levels data was non-stationary, led to proposition of the *Autoregressive Integrated Moving Average* (ARIMA) model which is a generalization of ARMA model. In ARIMA model, non-stationary time series data is made stationary by applying finite differencing of data points (Cochrane, 1997). The mathematical formulation of the ARIMA (*p, d, q*) using lag polynomials is given below (Lombardo and Flaherty, 2000).

$$\varphi(L)(1-L)^d\gamma_t = \theta(L)\varepsilon_t, \tag{8}$$

$$\left(1 - \sum_{i=1}^{p} \varphi_i L^i\right)(1-L)^d\gamma_t = \left(1 + \sum_{j=1}^{q} \theta_j L^i\right)\varepsilon_t \tag{9}$$

here *p, d* and *q* are integers greater than or equal to zero and refer to the order of the autoregressive, integrated and moving average parts of the model, respectively. The integer *d* controls the level of differencing.

Autocorrelation (ACF) and Partial Autocorrelation (PACF)

To determine a proper model for fitting time series data, ACF and PACF analysis was carried out. These statistical measures reflected how observations in a time series data are age-related to each other. For modelling and forecasting purposes, ACF and PACF against consecutive time lags were plotted. These plots helped to determine the order of AR and MA terms. Below are the mathematical models: For a time, series $\{x(t), t = 0,1,2 \ldots.\}$ the autocovariance at lag k is defined as:

$$\gamma_k = Cov(x_t x_{t+k}) = E[(x_t - \mu)(x_{t+k} - \mu)] \tag{10}$$

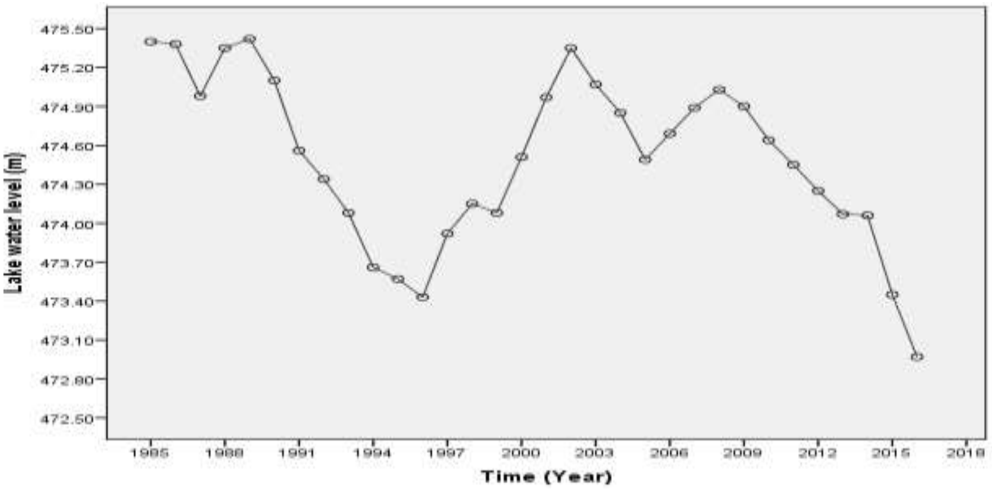

Figure 2. Water levels of Lake Malawi from the period of 1985 to 2016.

The autocorrelation coefficient at lag *k* is defined:

$$\rho_k = \frac{\gamma_k}{\gamma_0} \qquad (11)$$

where μ is the mean of the time series, that is, $\mu = E[x_t]$. The autocovariance at lag zero, that is, γ_0 is the variance of the time series. Another measure, known as the *Partial Autocorrelation Function* (PACF) is described by Box and Jenkins (1970). It is used to measure the correlation between an observation *k* past period and present observation after controlling observations at intermediate lags.

Trend model fitting

Conducting various diagnostic tests is an important step in time series modeling (Chung, 2009). The famous Box-Ljung Q-statistics as described by Box and Jenkins (1970) was used to transform the non-stationary data into stationary and to check adequacy for the residuals. In practice, the Box-Ljung Q-statistics was computed (Ljung and Box, 1978) as

$$Q = n(n+2)\sum_{k=1}^{m}\frac{\hat{r}_k^2}{n-k} \qquad (12)$$

where $\hat{r}_k$ is the estimated autocorrelation of the series at lag *k* and *m* is the number of lags being tested. Box and Jenkins (1970) developed a practical approach to build ARIMA model, which best fit a given time series and also satisfy the parsimony principle. According to Box and Jenkins (1970), the three-step approach of *model identification*, *parameter estimation* and *diagnoststic checking* to determine the best persimonious model from general class of ARIMA models (Zhang, 2003) were applied. The three-step process was repeated several times until a satisfactory model was finally selected. The appropriate model selection step is very critical. It is based on the fact that sample ACF and PACF, calculated from the training data should match with the corresponding theoretical or actual values (Chatfield, 1996). In this case, various model fitting statistics like Root Mean Square Error (RMSE), Maximum Absolute Percentage Error (MAXAPE) and Maximum Absolute Error (MAXAE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Mean Forecast Error (MFE) and Bayesian Information Criterion (BIC) were employed to evaluate the adequacy of AR, MA and ARIMA processes. Based on Normalized BIC, the principle is that the lower the value, the better the model. Fit statistics such as MAPE, MAE, MFE, BIC and RMSE were calculated as shown below:

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\gamma_i - \bar{y}_i}{\gamma_i}\right| \qquad (13)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\gamma_i - \bar{y}_i| \qquad (14)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\gamma_i + \bar{y}_i)^2} \qquad (15)$$

$$MFE = \frac{1}{n}\sum_{t=1}^{n} e_t \qquad (16)$$

$$BIC(p) = n\ln\left(\hat{\sigma}_e^2/n\right) + P + P\ln(n) \qquad (17)$$

where, γ_i and $\bar{y}_i$ are actual observed and predicted values respectively, while *n* is number of predicted values. In BIC model, *n* is the number of effective observations used to fit the model, *p* is the number of parameters in the model and $\hat{\sigma}_e^2$ is the sum of sample squared residuals. Upon identification of optimum model, forecast of the Lake Malawi water levels from 2017 to 2035 were made.

All inferential and descriptive statistics were performed using International Business Management Statistical Package for Social Scientists software (IBM SPSS 20) (IBM Corp, 2011).

RESULTS AND DISCUSSION

Model selection

The stationarity of a stochastic process was visualized in form of a data plot as shown in Figure 2. According to Hipel and McLeod (1994), identification of stationarity in time series data is a necessary condition for building a time series model that is useful for forecasting. Sankar (2011) defined time series stationarity as a set of values that vary over time around a constant mean and constant variance.

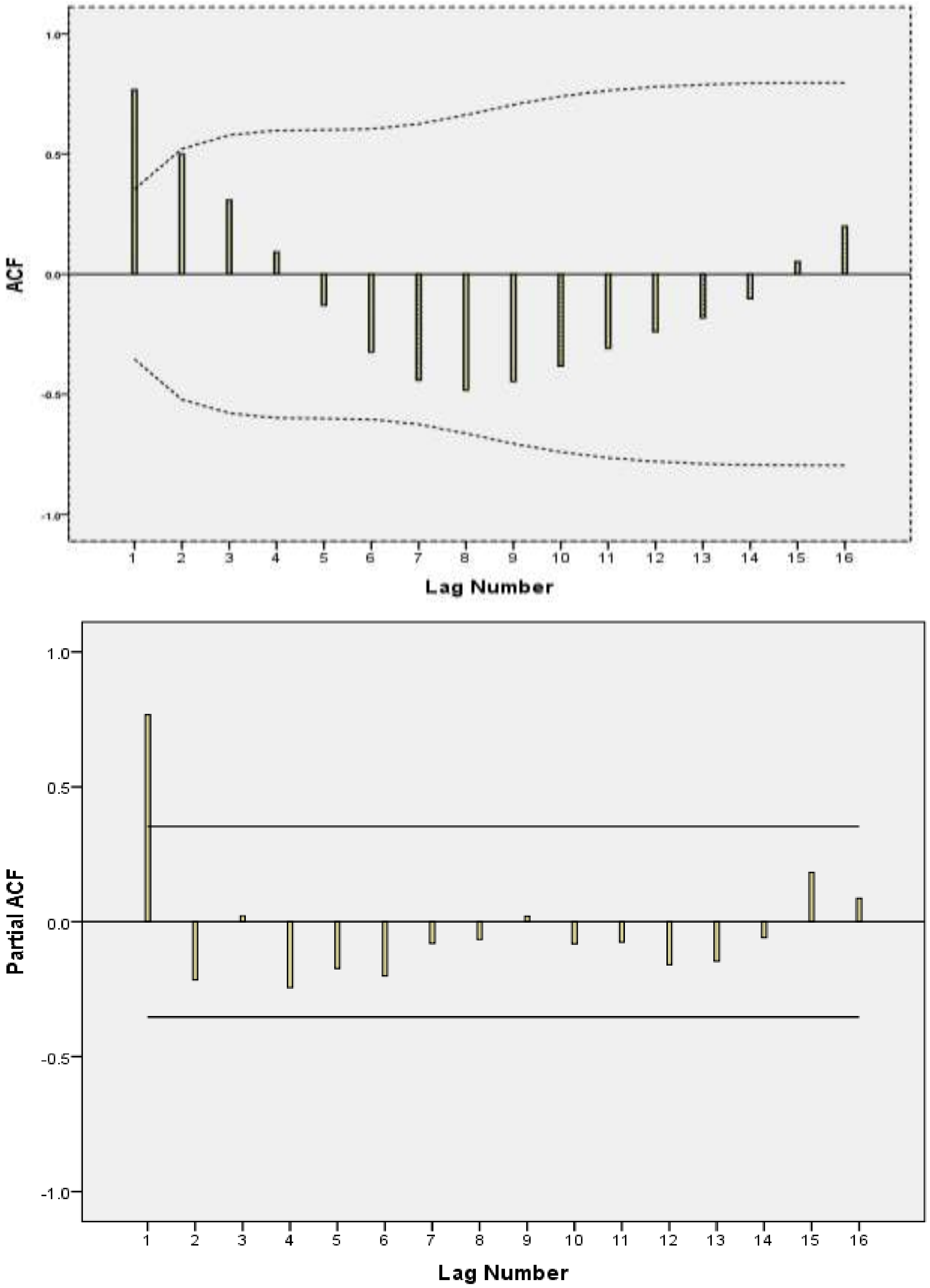

Figure 3. Autocorrelograms and partial autocorrelograms of first order differenced data.

According to Hipel and McLeod (1994), time series data showing seasonal patterns are usually non-stationary in nature. From Figure 2, it is very apparent that the time series data from Lake Malawi water levels is non-stationary due to unstable means which increase and decrease at some points throughout 1985 to 2016. Similar observation was reported by several authors in Lake Malawi (Lazaro and Jere, 2013; Singini et al., 2012; Mulupwa et al., 2016; Zindi et al., 2016). Given these difficulties in Lake Malawi water levels time series data, first order differencing of the data and stationary test were conducted on the newly constructed series of the data. Since the newly constructed data was stationary in mean, the next issue was how to select an appropriate model that can produce accurate forecast based on the description of historical pattern in the data and how to determine the optimal model order. In this case, the values of p and q in the ARIMA model were identified by plotting autocorrelogram and partial autocorrelogram presented in Figure 3.

Figure 3 illustrated that autoregressive model of order p(AR (q)) was stationary and moving average model of

Table 1. ACF and PACF for time series data of Lake Malawi water levels.

Lag	ACF	Std Error	Box-Ljung Statistic			PACF	Std Error
			Value	df	Sig		
1	0.767	0.177	20.647	1	0.000*	0.767	0.177
2	0.499	0.261	29.694	2	0.000*	-0.216	0.177
3	0.309	0.289	33.267	3	0.000*	0.021	0.177
4	0.092	0.299	33.598	4	0.000*	-0.244	0.177
5	-0.130	0.300	34.276	5	0.000*	-0.173	0.177
6	-0.323	0.302	38.646	6	0.000*	-0.200	0.177
7	-0.441	0.313	47.112	7	0.000*	-0.080	0.177
8	-0.483	0.331	57.673	8	0.000*	-0.066	0.177
9	-0.446	0.353	67.094	9	0.000*	0.020	0.177
10	-0.383	0.370	74.331	10	0.000*	-0.082	0.177
11	-0.308	0.382	79.239	11	0.000*	-0.076	0.177
12	-0.239	0.390	82.349	12	0.000*	-0.159	0.177
13	-0.181	0.394	84.235	13	0.000*	-0.147	0.177
14	-0.102	0.397	84.869	14	0.000*	-0.059	0.177
15	0.052	0.398	85.044	15	0.000*	0.182	0.177
16	0.200	0.398	87.753	16	0.000*	0.086	0.177

^{ns}: Non-significant, *, **: Significant at P<0.01, and P < 0.05, respectively.

Table 2. Fit statistics for various competing ARIMA models.

ARIMA (p,d,q)	RMSE	MAPE	MAXAPE	MAE	MAXAE	MFE	NBIC
ARIMA (1,1,0)	0.28	0.04	0.12	0.19	0.53	0.38	-0.21
ARIMA (1,1,2)	0.29	0.05	0.13	0.21	0.63	0.52	-1.87
ARIMA (1,1,1)	0.29	0.05	0.12	0.22	0.58	0.46	-2.00
ARIMA (0,1,1)	0.29	0.05	0.12	0.22	0.59	0.54	-2.12
ARIMA (2,1,2)	0.31	0.05	0.13	0.22	0.13	0.49	-1.70

order q(MA (q)) was good. Guti´errez-Estradade et al. (2004) explained that a good autoregressive model of order p(AR (q)) has to be stationary and a good moving average model of order q(MA (q)) has to be invertible. The invertibility and stationarity gives a constant mean, variance and covariance which is a necessary condition for forecasting (Singini et al., 2012). Following Hipel and McLeod (1994), autocorrelation and partial autocorrelation coefficients (ACF and PACF) of up to 16 lag were considered. The type and order of the adequate model required to fit the series was determined. As the ACF values diminished rapidly with increasing lags, it was assumed that lynx series was stationary. The autocorrelation and partial autocorrelation coefficients (ACF and PACF) of various orders of differenced series of data were computed and presented in Table 1. The basic principle of model parsimony states that the model with smallest number of parameters is to be selected so as to provide an adequate representation of the underlying time series (Chatfield, 1996). In other words, out of a number of suitable models, it is very important to consider the simplest model while still upholding an accurate description of inherent properties of the time series (Zhang, 2007).

As discussed by Hipel and McLeod (1994), a number of ARIMA models were competed in order to select the simplest one as shown in Table 2. Hipel and McLeod (1994) observed that the more complicated the model, the more possibilities will arise for departure from actual model assumptions. In other words, with the increase of model parameters, the risk of model overfitting also subsequently increases. Although over fitted time series models describe the data very well, it may not be suitable for future forecasting. Therefore, genuine attention was given to select the most parsimonious model among all other possibilities. Using the coefficients in Table 1, various ARIMA models were identified and the models together with their corresponding fit statistics are presented in Table 2. The Root Mean Square Error (RMSE) which measured how much dependent series varies from its model-predicted level was lowest (0.28) in ARIMA (1,1,0) model which according to Cao and Francis

Table 3. Lake Malawi water levels estimated ARIMA model.

Parameter	Estimate	Std Error	t-value	p-value
Constant	10.56	0.41	0.59	0.56^{ns}
AR	0.38	0.18	2.13	0.04**

^{ns}: Non-significant, *, **: Significant at P<0.01 and P < 0.05, respectively.

(2003), indicated a good forecast of the model. Similarly, Mean Absolute Error (MAE) also known as Mean Absolute Deviation was lowest (0.19) in ARIMA (1,1,0) model which indicated a good forecast of the model. In other words, the magnitude of overall error occurring due to forecasting was very small.

It was further noted that Mean Absolute Percentage Error (MAPE) was lowest and smallest in ARIMA (1,1,0) meaning that the percentage of average absolute error occurring was very small. In other words, the opposite signed errors did not offset each other. It was further interesting to note that Maximum Absolute Percentage Error (MAXAPE) and Maximum Absolute Error (MAXAE) expressed as percentage was very small in ARIMA (1,1,0) model indicating overall good model fit. According to Czerwinski et al. (2007), the best model should have adequate accuracy measures (RMSE, MAE) and lowest Normalised BIC for it to have accurate forecasts. Therefore, ARIMA (1,1,0) model was selected because it had lowest RMSE, MAE, MFE and Normalized Bayesian Information Criterion (NBIC). It was further observed that the coefficients of the parameters of ARIMA (1,1,0) model were significant. According to Czerwinski et al. (2007), the model which indicate lowest normalized BIC and is significant ($p<0.05$) is a better model in terms of forecasting performance than with large normalized BIC. Estimates of the selected ARIMA (1,1,0) model are presented in Table 3.

Based on the study findings, the most suitable model for forecasting Lake Malawi water levels was confirmed to be ARIMA (1,1,0).

Model systematic checks

The basic model verification is concerned with checking the residues to see if they contain any systematic pattern which could still be eliminated to improve the performance of the selected model. Therefore, the selected ARIMA (1,1,0) model was subjected to autocorrelations and partial autocorrelations of residues of various orders. Various autocorrelations of up to 24 lags were computed and plotted as shown in Figure 4. The results showed that none of the autocorrelation was significantly different from zero at any reasonable level. This implied that the selected ARIMA (1,1,0) model was an appropriate model for forecasting Lake Malawi water levels.

It is very apparent from Figure 4 that autocorrelations of the coefficients are within 95% confidence interval, suggesting that the selected model was well fitted in time series model and had an accurate forecast.

Forecasting

Using the selected ARIMA (1,1,0) model, the forecast of Lake Malawi water levels was made from 1985 to 2035.

For preciseness and accurateness sake, only observations from 2005 to 2016 were compared with the forecasted values as shown in Table 4. Figure 5 on the other hand, indicates the forecasted value from 1985 to 2035. Czerwinski et al. (2007) explained that the forecasted and actual values need to be very close, meaning that the forecasting error must be very low for the model to qualify as good. As observed in Table 4, the noise residues were a combination of positive and negative errors indicating that the model had a good performance of forecasting. It was further interesting to note that the magnitude of the difference between the forecasted and actual values were very low indicating a good forecasting performance. In Figure 5, it is very apparent that Lake Malawi water levels are fluctuating with a negative trend. Such negative trend will continue up to 2035.

Figure 5 further indicated that values for water levels increased during 2006 to 2010 and decreased up to 2015 when compared to values of 2005. However, the trend declined continuously up to 2035. The basic principle of ARIMA model assumes that time series data is linear and follows a particular known statistical distribution such as normal distribution (Cochrane, 1997). Therefore, it may be concluded that the trend in this study behaved in a manner consistent with ARIMA principle which is assumed to follow a certain probability model described by joint distribution of random variable. It is also interesting to note that time series is non-deterministic in nature such that it cannot predict with certainty what will occur in the future. Based on this observation, the study indicated that there is high probability that Lake Malawi water levels will decrease as far as up to 468.63 m by 2035. Kidd (1983) had similar observation in 1915 and recorded the lowest lake level of 469 m above sea level. Drayton (1984) in the 1980s reported that Lake Malawi water levels have been unstable over the years with notable events occurring in 1890s where unusual low water levels (112 m) were recorded. He further noted that the Lake water levels were near cessation of outflows for more than 20 years (from 1890s to 1935) and experienced high levels and outflows in 1970s and 1980s which caused flooding of lakeshore communities and areas immediately downstream. Kidd (1983) earlier noted that a small decrease in the ratio resulted in the basin being closed with no outflow as occurred between 1915 and 1937. Recently, Shela (2000) observed unusual low level (115 m) and outflows in the 1990s which was further

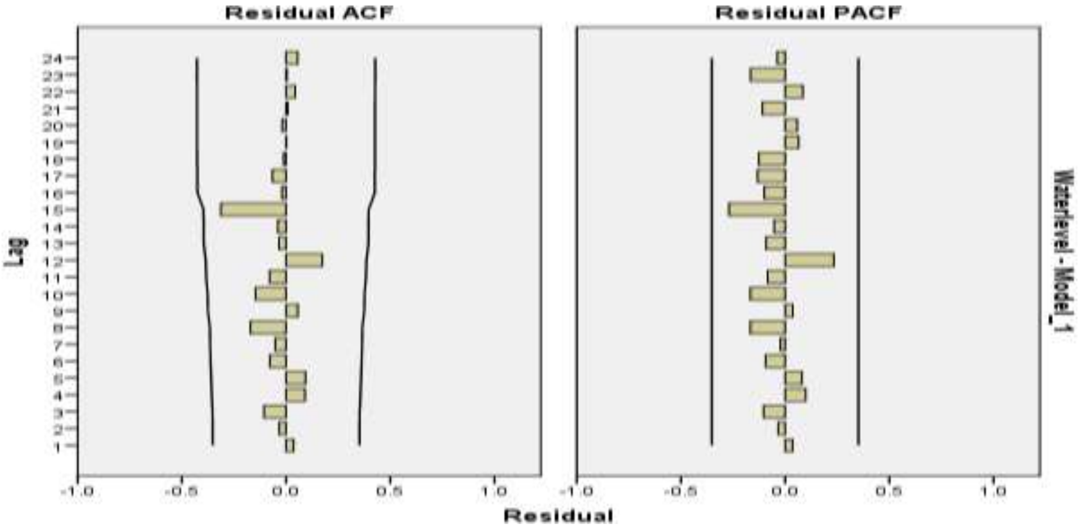

Figure 4. ACF and PACF residue.

Table 4. Forecasted Lake Malawi water levels.

Year	Actual water level (m)	Predicted water level (m)	95% confidence interval
2005	474.49	474.70	(476.0, -475.6)
2006	474.69	474.28	(-475.6, 474.4.4)
2007	474.89	474.69	(-474.5, 373.4)
2008	475.03	474.89	(-134.31, 124.32)
2009	474.90	475.00	(-402.1, 470.29)
2010	474.64	474.77	(-373.69, 374.88)
2011	474.45	474.45	(-171.1, 170.29)
2012	474.25	474.29	(-804.29, 815.48)
2013	474.07	474.08	(102.41, 105.6)
2014	474.06	473.90	(-604.17, 575.05)
2015	473.45	473.96	(-073.86, 075.05)
2016	472.97	473.11	(-408.69, 414.88)
2017		472.68	(-106.48, 124.68)
2018		472.46	(-102.31, 114.5)
2019		472.26	(-401.36, 404.55)
2020		472.07	(-102.52, 203.71)
2021		471.87	(-108.09, 103.27)
2022		471.68	(-071.45, 073.47)
2023		471.47	(-570.9, 573.62)
2024		471.27	(-010.42, 013.78)
2025		471.05	(-069.97, 073.780
2026		470.84	(-409.55, 407.81)
2027		470.61	(-468.14, 473.81)
2028		470.38	(-1068.75, 1473.8)
2029		470.15	(-106.36, 107.76)
2030		469.91	(-401.98, 423.71)
2031		469.67	(-132.08, 132.4)
2032		469.42	(-246.21, 246.2)
2033		469.16	(-187.63, 179.2)
2034		468.90	(-465.03, 472.7)
2035		468.63	(-1464.6, 1473.7)

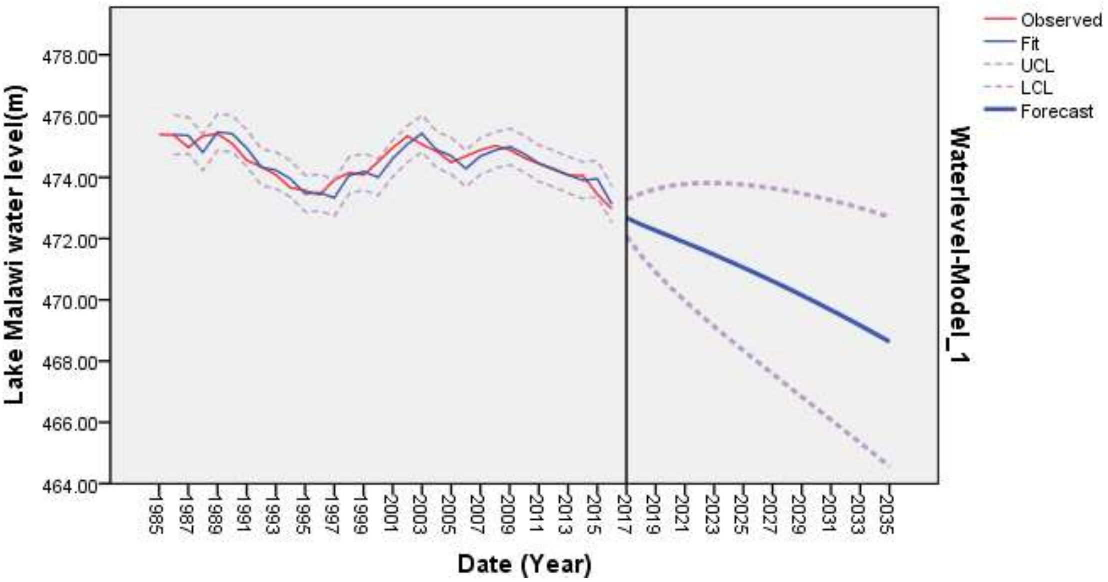

Figure 5. Actual and forecasted Lake Malawi water level.

associated with a widespread regional drought. The study by Neuland (1984) also revealed that there is little risk of the lake level exceeding 477.8 m above mean sea level. Using the most recent observed climatic parameters of the lake, the predicted level by Neuland (1984) remains below 477 m and further indicated a high probability of negative trend of future water levels as reported in the present study. Kumambala and Ervine (2010) further added that it is very unlikely for the water level to increase to a maximum height of 477 m as it was in 1980. Recent prediction by Kaunda (2015) indicated that near future and far future projects show that water yield will decrease by 8.84% and therefore Lake Malawi water level is expected to drop. However, Kaunda findings were thus on short term from 2017 to 2020. Following the dramatic rise in lake level in 1979, Drayton (1979) made a statistical analysis of lake levels and recommended a “safe" static level of 477.6 m ASVD for the next 30 years. Nonetheless, the negative trend of Lake Malawi water levels predicted in the present study is worrisome. With such future prediction, deliberate effort has to be made to find appropriate policy options and strategies for sustaining Lake Malawi water levels.

Conclusion

The study selected the ARIMA (0, 1, 1) model for forecasting Lake Malawi water levels. The ARIMA (0, 1, 1) had lowest Normalized Bayesian Information Criterion (NBIC), Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE), Mean Forecast Error (MFE) and Mean Absolute Error (MAE) which indicated a good forecast of the model. Based on the selected model, it is very apparent that Lake Malawi water levels fluctuation is showing a negative trend. Such negative trend is predicted to continue up to 2035. The model further predicted that Lake Malawi water levels will decrease up to 468.63 m by 2035. This study provides critical information for future policy making and formulation of intervation strategies for sustaining Lake Malawi water levels.

RECOMMENDATION

The major limitation of ARMA and ARIMA models in this study was that they only capture short-range dependence (SRD). In other words, they belong to the conventional integer models. In practice, several time series exhibit long range dependence (LRD) in their observations. To overcome this difficulty, it is recommended that a similar study should be conducted using *Autoregressive Fractionally Integrated Moving Average* (ARFIMA) model with ability to capture long range property of the fraction system accordingly and project extended period of more than 18 years.

CONFLICT OF INTERESTS

The authors have not declared any conflict of interests.

ACKNOWLEDGEMENT

The authors wish to thank Department of Water Resources of the Ministry of Agriculture, Irrigation and

Water Development for the kind support in providing historical data for Lake Malawi water levels. Many thanks to Mayamiko Chaweza for assisting in data analysis.

REFERENCES

Alonso A, Garcia-Martos C (2012). Time series and stochastic processes. Madrid: Universidad Carlos III de Madrid. Available online at www.etsii.upm.es/ingor/estadistica/Carol/TSAtema3petten.pdf

Aksoy H, Unal N, Eris E, Yuce M (2013). Stochastic modeling of Lake Van water level time series with jumps and multiple trends. Hydrol. Earth Syst.Sci. 17:2297-2303.

Box G, Jenkins G (1970). Time Series Analysis, Forecasting and Control. San Francisco: Holden-Da. Available online at www.library.wiley.com/doi/10.1002/9780470316566.refs/pdf

Box G, Jenkins G, Reinsel G, Ljung G (2015). Time Series Analysis: Forecasting and Control. Hoboken, NJ, USA: John Wiley and Sons.

Cao L, Francis E (2003). Support Vector Machine with Adaptive Parameters in Financial Time Series Forecasting. IEEE Trans. Neural Networks 14(6):1506-1518.

Chang X, Gao M, Wang Y, Hou X (2012). Seasonal Autoregressive Integrated Moving Average Model For Precipitation Time Series. J. Maths. Stat. 8(4):500-505.

Chatfield C (1996). Model uncertainty and forecast accuracy. J. Forecasting pp. 495-508.

Chung M (2009). Lecture on time series diagnostic test. Taipei 115, Taiwan: Institute of Economics, Academia Sinica. Available online at www.ntu.edu.tw/~ckuan/pdf/Lec-DiagTest_0902.pdf

Cochrane H (1997). Time Series for Macroeconomics and Finance. Chicago: Graduate School of Business, University of Chicago, Spring. Available online at www.bseu.by/russian/faculty5/stat/docs/4/Cochran,TimeSeries.pdf

Cryer J, Chan K. (2008). Time Series Analysis with Application in R 2nd ED. New Yolk: Springer.

Czerwinski I, Guti´errez-Estrada J, Hernando-Casal J (2007). Short-term forecasting of halibut CPUE: Linear and non-linear univariate approaches. Fisheries Res. 86:120-128.

De Vas A (1994). Pliocene en Kwartaire evolutie van het Livingstone bekken (Malawi rift, Tanzanie) afgeleid uithoge-resolutie reflectieseismische profielen.Licenciaat theis. Belgium: s, Universiteit Gent. Avaliable online at https://www.gadventures.com/trips/victoria-falls-serengeti-adventure.../DVN/

Dixey F (1924). Lake level in relation to rainfall and sunspots. Nature 114:659-661.

Drayton R (1979). A study of the causes of the abnormally high levels of Lake Malawi. Lilongwe: Wat. Resour. Div. Tech. Pap.No 5.

Drayton R (1984). Variation in the level of Lake Malawi. J.Hydrol.Sci. 29:1-12.

GoM (2005). National Spatial Data Center, Ministry Lands Housing physical Planning and Surveys, Lilongwe, Malawi. Available online at www.lands.gov.mw/index.php/contacts/physical-planning.ht

Guti´errez-Estradade J, L´opez-Luque E, Pulido-Calvo I (2004). Comparison between traditional methods and artificial neural networks for ammonia concentration forecasting in an eel (Anguilla Anguilla L.) Intensive rearing system. Aquat. Eng. 31:183-203.

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

IBM Corp (2011). IBM SPSS Statistics for Windows, Version 20.0. Armonk, NY: IBM Corp. Available online at www.sciepub.com/reference/159832

Johnson T, Davis T (1989). High resolution profiles from Lake Malawi, Africa. J. Afr. Earth. Sci. 8:383-392.

Kantz H, Schreiber T (2004). Nonlinear Time Series Analysis 2nd Edn. Cambridge: Cambridge University Press.

Kaunda P (2015). Investigating the Impacts of Cliamte Change on the Levels of lake Malwi Thesis, Department of meteology, university of nairobi, Kenya.

Kidd C (1983). A water resources evaluation of Lake Malawi and the Shire River. Geneva:UNDP project MLW/77/012, World Meteological Organisation.

Kumambala P, Ervine A (2010). Water Balance Model of Lake Malawi and its Sensitivity to Climate Change. Open. Hydrol. J. 4:152-162.

Kumambala P (2010). Sustainability of water resources development for Malawi with particular emphasis on North and Central Malawi. PhD thesis. Glasgrow, UK: University of Glasgrow. Available on line at https//theses.gla.ac.uk/1801/1/2010Kumambalaphd.pdf

Lazaro M, Jere W (2013). The Status of the Commercial Chambo (Oreochromis Species) fishery in Malaŵi: A Time Series Approach. Intl. J. Sci. Technol. 3:1-6.

Lombardo R, Flaherty J (2000). Modelling Private New Housing Starts In Australia. Pacific-Rim Real Estate Society Conference Sydney: University of Technology Sydney (UTS) pp. 24-27.

Ljung G, Box G (1978). On a measure of lack of fit in time series models. J. Bio. 65:297-303.

Mulumpwa M, Jere W, Mtethiwa A, Kakota T, Kang'ombe J (2016). Application Of Forecasting In Determining Efficiency Of Fisheries Management Strategies Of Artisanal Labeo mesops Fishery Of Lake Malawi. Int. J. sci. Tech. Resea 5(09):28-40.

Neuland H (1984). Abnormal high water levels of Lake Malawi? -An attempt to assess the future behaviour of the lake water levels. Geo. J. 9:323-334.

Sankar I (2011). Stochastic Modelling Approach for forecasting fish product export in Tamilnadu. J. R .Resea. Sci.Technol. 3(7):104-108.

Scholz E, Rozendahl B (1988). Low lake stands in Lakes Malawi and Tanganyika, East Africa, delineated from multifold seismic data. J. Sci. 240:1645-1648.

Shela O (2000). Naturalisation of Lake Malawi Levels and Shire River Flows: Challenges of Water Resources Research and Sustainable Utilisation of the Lake Malawi-Shire River System. Sustainable Use of Water Resources Maputo: 1st WARFSA/WaterNet Symposium, pp. 1-12. Available online at https://floodobservatory.colorado.edu/SiteDisplays/Shire2SHELA.PDF

Singini W, Kaunda E, Kasulo V, Jere W (2012). Modelling and Forecasting Small Haplochromine Species (Kambuzi) Production in Malaŵi-A Stochastic Model Approach. Intl. J. Sci. Technol. Resea 1:69-73.

Stuffer D, Dhumway R (2010). Time series Analysis and its Application 3rd Ed. New Yolk: Springer.

Yevjevich V (1972). Probability and Statistics in Hydrology. Colorado: Water Resources Pub. Available online at https://www.abebooks.com/.../probability-and-statistics-in-hydrology/.../yevjevich-vu

Zhang G (2007). A neural network ensemble method with jittered training data for time series forecasting. J. Infor. Sci. 177:5329-5346.

Zhang G (2003). Time series forecasting using a hybrid ARIMA and neural network model. J. Neuroc. 50:159-175.

Zindi I, Singini W, Mzengeleza K (2016). Forecasting Copadichromis (Utaka) Production for Lake Malaŵi, Nkhatabay Fishery- A Stochastic Model Approach. J. Fish. Livest. Prod.4:162.