International Journal of Scientific and Innovative Mathematical Research(IJSIMR)
Volume 8, Issue 4, 2020, PP 14-27
ISSN No. (Online) 2347-3142
DOI: https://doi.org/10.20431/2347-3142.0804002
www.arcjournals.org

Forecasting Rainfall in Tanzania Using Time Series Approach Case Study: Dar es Salaam

Paul Andrew Panga[1*], Shaban Nyimvua[2], Isambi Mbalawata[3]

[1]*National Institute of Transport (NIT)*

[2]*Department of Mathematics University of Dar es salaam*

[3]*African Institute of Mathematical Sciences (AIMS)*

Corresponding Author:** ***Paul Andrew Panga, *National Institute of Transport (NIT), Tanzania*

Abstract: *The prediction of rainfall on monthly time scale has been attempted by a number of researchers by using different time series techniques at different time periods around the world. It is challenging to forecast rainfall at monthly time scale because of spatial and temporal random variation caused by a numbers of dynamic and environmental factors. In this paper, an attempt has been made to develop a Seasonal Autoregressive Integrated Moving Average (SARIMA) Model to analyze long term monthly rainfall data of Dar es Salaam region in Tanzania for the period of fifty three years (1961 to 2014). Rainfall observations were discovered to have Seasonality and also non-stationarity and hence differencing and Seasonal differencing was used to attain stationarity. Rainfall data were found to have two seasons namely October to December (OND) and March to May (MAM) .The analysis exhibited that the Seasonal ARIMA model which is satisfactory in describing the monthly rainfall data in Dar es Salaam Tanzania is SARIMA $(2, 1, 1)(1, 1, 1)_{12}$. The model was then used for predictions of monthly rainfall values from January 2015 to December 2024. The forecasting results showed that monthly rainfall values have a decreasing trends, hence that may be a threat to agriculturists and water managers in the region. The study will be useful to decision makers for the region of Dar es Salaam to establish priorities and strategies based on the impacts posed by the variation of rainfall.*
Keywords: *Seasonal ARIMA (SARIMA) models, Forecasting, time series, MAPE, RMSE*

1. INTRODUCTION

For a decades now, the trends of rainfall instability and its impacts has been a crucial climatic problems facing different nations. The scenario has been linked directly or indirectly with global warming, which pose its impacts to a number of sectors particularly agriculture and tourism whose contribution is vital to any countries economy. Vital sectors of the Tanzania economy such as agriculture, fishing, to mention few, depend on rainfall. Considering this case for the survival and growth of the plants, rainfall is required, though too much or too little is still a problem (Shahin et al, 2016). On the other hand, heavy precipitation is associated with floods, loss of people's lives and outbreak of diseases (Ojija et al, 2017). So it is unavoidable to have untimely effects of dynamics of rainfall patterns. The United Republic of Tanzania (URT, 2007) and (Orindi V and Murray L, 2005) alludes that the impacts of rainfall instability will persist to torture agriculture, biodiversity, livelihood, health and other sectors. Hence, early indication may help to solve a number of problems associated with dynamic of rainfall trends.

Variation of climate has been a topic in many parts of the world due to its immediate effects on people's lives (Ghahraman, 2007). Over the past decades Tanzania has witnessed the increase of climatic events such as floods, which are linked with grievous ecological and socio-economic intimidation like loss of lives and destruction of structural design (Kijazi A and Reason C, 2009). Serious floods that have recently tortured many parts of the Tanzania include that of 2006, 2009, 2010, 2011, 2012, 2014, 2016 and 2017 (Chang'a et al, 2007). It is evident that the coastal areas of Tanzania including Dar es Salaam region will encounter a number of damages due to increasing trend of temperature plus instability in precipitation. The trends of precipitation in the region is dynamic and there is variability in rainfall caused by a number of different time scales from daily to decadal (UARK, 2017). It has been noticed that the trend and variability of rainfall will continue at a longer timescale (IPCC, 2012). Therefore, there is a need for a suitable prediction method to be applied in forecasting precipitation patterns.

Recently, Time series analysis and forecasting was observed to be an important tool when applied in studying the variations and trends of different hydo-meteorological variables such as precipitation, humidity, temperature, streamflow and many other environmental parameters (Nury et al, 2013). Various published papers have analyzed precipitation by using Time series Box and Jenkins ARIMA and SARIMA approaches, which gives the usefulness of modelling rainfall from different parts of the world. Libya (Tariq et al, (2014), El-Mahal et al, (2016)), Nigeria (Osarumwense et al, (2013), Eni (2015)), Sri lanka (Alibuhtto et al, (2019)), Iran (Machekposhti et al, (2018)), Jordan (Momani et al, (2009)),Iraq (Chawsheen et al, (2017)), Baghdad (Ali (2013)), Greece (Karavities et al, (2015)) and Bangladesh (Nury et al, (2013), Sultana et al. (2015)). Most of the observations and time series modelling results of the mentioned studies have declared projected instability in rainfall patterns. However, there are limited or no published papers that have attempted to understand, analyze, model and predict rainfall by using Box and Jenkins ARIMA approach in Tanzania particularly Dar es salaam. Therefore, this paper would seem to be the first application of the Box and Jenkins ARIMA approach for rainfall in Dar es Salaam, Tanzania.

In this study, we first check if the condition of stationarity in the time series data is attained, then followed by finding the appropriate time series model for monthly rainfall data by using previous available data from January 1961 to December 2014 of Dar es Salaam region in Tanzania. Second, we will check if the parameters are sensitive to the time series models and finally, we will predict the future trends of rainfall values by using the time series model developed. Box and Jenkins methodology will be used in developing the time series model. The approach flows through identification of the model, estimation of the model parameters, diagnostic checking of the selected model and lastly the use of the model in forecasting purposes (Box GE and Jenkins GM, 1976).

Different researchers allude that socio economic development of the developing countries like Tanzania are hindered by the trends and patterns of climatic extremes (Chang'a et al, 2017). Efforts like achieving Millennium Development Goals (MDG), Sustainable Development Goals (SDG) and National Developmental Vision (Visions 2025) which are associated with reducing poverty, hunger and promoting food security are hampered by floods and natural disasters like drought, thus if not managed properly the prolonged impacts will continue in the future. Hence this study is of practical importance for providing information to decision makers, planners, climatologist, meteorologist and others in predicting the future rainfall. The paper is organized as follows. Section 2, describes study area, data and methodology for fitting time series models. Finally, the results of the appropriate time series model and their prediction are discussed in Section 3 and finally we give conclusion.

2. Materials and Methods

2.1. Data Collection

The Tanzania Meteorological Agency (TMA), Tanzania, is the responsible organization for the collection and publication of meteorological data. It is a government agency responsible for meteorology issues in the country. The data used in this paper are completely secondary in nature and they are collected from TMA. The daily rainfall data from the period of January 1961 - December 2014 of Dar es Salaam region of Tanzania are used. The given daily rainfall data was converted into monthly data by using the totaling approach. Since rainfall data are time dependent data then monthly data was converted into time series data and then smoothing moving average of five data points was applied.

2.2. Study Area

Dar es Salaam is one among the thirty regions of Tanzania, lying at the latitudes of 6°52' South and longitudes of 39°12' East. It is among highly populated coastal regions with population of 6,368,272 covering the area of 1,393km2 (WPR, 2019). The region constitutes of five districts which are Kinondoni, Ubungo, Kigamboni, Ilala and Temeke.

Dar es Salaam region is characterized by tropical type of climate with higher degree of hotness, high humidity and average annual precipitation of over 1000 mm (UARK, 2017). The region is characterized by bimodal rainy seasons. The longer rain falls from March to May (MAM) and shorter rains fall from October to December (OND). The map of Tanzania and the extract of Dar es Salaam region from the map are exhibited in Figure 1.

Figure1. *A map of Tanzania and the extracted map of Dar es Salaam*

2.3. Decomposition of the Time series

The decomposition of the time series patterns are categorized into two namely additive and multiplicative hypothesis. In this study decomposition by additive method is used and this can be expressed mathematically as:

$$Y_t = T_t + S_t + C_t + R_t,$$

where Y_t, T_t, S_t, C_t, R_t and t stands for time series, trends, seasonal, cyclic, random components and time respectively.

2.4. Moving Average Smoother

A smoother is one of the most useful tools for expressing the trends of the dependent variable as the function of one or more regressors. Smoothers is used when it happens the amount of horizontal scatter in data places difficulties in seeing the trends. The methods is commonly used in univariate time series analysis.

3. METHODOLOGY

According to (Box GE and Jenkins GM, 1976), Time series is a sequence of data arranged in chronological order. Fundamental principal of the time series is to understand the historical trends of the data at a particular time. Normally if the previous values are well described then they can be used to forecast the future values of the series. Time series data can either be discrete if it is recorded at a discrete time point or continuous if it is recorded at every instance of time. Time series observations for rainfall and temperature which are recorded continuously can be converted to discrete time points. Mathematical technique for modelling precipitation values is a stochastic process. A number of probability models have been developed to understand weekly, monthly and annual precipitation, however nowadays monthly rainfall are analyzed by using time series models. Time series models have been thoroughly studied by Box and Jenkins (1976). Till now, a number of researchers still use this method due to its effectiveness in forecasting purposes (Mishra U and Jain V, 2010).

3.1. Autoregressive Integrated Moving Average (ARIMA) Model

ARIMA model is the generalization of the Autoregressive Moving Average (ARMA) Model. In ARMA model the present values of the time series is expressed as the linear combination of p past values and q previous error which is calculated by subtracting the fitted values of the previous error, also the random error is included (Mishra A and Desai V, 2005). However, the model is only used when the data are stationary. It is important to note that the series is said to be stationary when the observations has constant mean and variance. The differencing can be incorporated in the ARMA model when the observation are non-stationary. These model is now called Autoregressive Integrated Moving Average

(ARIMA) model. It is referred to as ARIMA (p, d, q) model where p, d and q are order of autoregressive components, order of differencing used and order of moving average components of the model respectively. Differencing a series involves subtracting its current and previous values d times. The non-seasonal ARIMA (p, d, q) model can be expressed in a compact way as:

$$\varphi(B)\Delta^d y_t = c + \theta(B)\varepsilon_t,$$

here ε_t is independent and identically distributed with mean zero and constant variance. y_t is the time series observation at time t. Also φ and θ are the coefficient of AR and MA process respectively. $(1-B)^d = \nabla^d$ is the non-seasonal difference operator of order d $(d = 0, 1, 2)$, $\varphi(B)$: is Non-seasonal AR operator(Its polynomial $\varphi(B)$= 1 - $\varphi_1(B) - \varphi_2(B^2) - \cdots - \varphi_p(B^p)$,note that $\varphi_1, \varphi_2, \ldots, \varphi_p$, are non-seasonal AR parameters and when $\varphi(B)$ = 0 means that the root of the polynomials lies outside the unit circle. $\theta(B)$ is the non-seasonal moving average parameter (polynomial $\theta(B)$ = 1 - $\theta_1(B) - \theta_2(B^2) - \cdots - \theta_q(B^q)$), note that: $\theta_1, \theta_2, \ldots, \theta_q$ are non-seasonal MA parameters and when $\theta(B)$ = 0 means that the roots of the polynomials lies outside the circle. The weakness of this model is that it does not fit for time series observation with seasonal effects. To deal with this (Box GE and Jenkins GM, 1976) came with the following model.

3.2.Multiplicative Seasonal ARIMA (SARIMA) Model

Some of the time series has the character of seasonal components that re occur in every S observations. For monthly data S is 12 (12 *in one year*) and for quarterly data S *is* 4 (4 *in one year*). One among the useful method which incorporate seasonality is the theorem propounded by Box and Jenkins (1970) called Seasonal Autoregressive Integrated Moving Average (SARIMA). It is the mostly used technique for forecasting which is in need of a long seasonal time series data. This model disintegrate the previous data into Autoregressive process, which take into account the memory of the previous values in Integrated process, which deals with stabilizing or making the data stationary and Moving average process, which includes previous error terms. The model is useful in capturing non seasonal and seasonal behavior of the time series observations. Let $x_i(i = 1, 2, \ldots, t)$ be a series under consideration. The multiplicative Seasonal ARIMA (SARIMA) model for the series is given by (Box GE and Jenkins GM, 1976)

$$\phi(B)\Phi(B^S)[(1-B)^d(1-B^S)^D X_t] - \mu = \theta(B)\Theta(B^S)a_t$$

Or

$$\phi(B)\Phi(B^S)(W_t - \mu) = \theta(B)\Theta(B^S)a_t$$

Where, X_t is the time series observations at time t, t: is the discrete time, S is the seasonal length, μ: is the mean level of the time series process (Usually computed as average of W_t), note when $d + D > 0$ implies $\mu \equiv 0$, at: residual of the series, $NID(0;\delta^2)$, $\Phi(B^S)$: is the seasonal AR operator (polynomial $\Phi(B) = 1 - \Phi_1(B) - \Phi_2(B^2) - \cdots - \Phi_p(B^p)$, $(1-B)^D = \nabla_s^D$: is the seasonal difference operator of order D $(D = 0, 1, 2)$, $W_t = \nabla^d \nabla_s^D X_t$: is the stationary series formed after differencing X_t number of terms of W_t series are computed by $n = N - d - SD$, $\Theta(B^S)$: is the seasonal MA operator of order Q (polynomials $\Theta(B^S)$ = 1 - $\Theta_1(B^S)$ - $\Theta_2(B^{2S})$, note that $\Theta_1, \Theta_2, \ldots, \Theta_Q$ are the seasonal MA parameters and when $\Theta(B^S)$ = 0 means the root of the polynomials lies outside the circle.

SARIMA model is represented as $SARIMA(p;\ d;\ q)(P;D;Q)$, where: $(p;\ d;\ q)$ are the non-seasonal operator and $(P;D;Q)$ are the seasonal operator. Note: If the model is non seasonal, then only $(p;\ d;\ q)$ is required and if the model is seasonal then only $(P;D;Q)$ are needed.

3.3.Box and Jenkins Algorithms

Back in 1976, Box and Jenkins give a methodology (Presented in the Table below) in time series analysis to find the best fit model using the previous values to give the predicted values. One of the advantages of Box and Jenkins ARIMA time series model is that, it has an ability to generate the sequence of historical data and produce mathematical formula which will then be used to generate forecasted values. Also some articles have approved Box and Jenkins methodology as a very strong tools for giving solution of the prediction problems due to its ability to provide very tremendous correct prediction of the time series and also it yields a framework to develop the model and do analysis (Montgomery D and Johnson L, 1967). The aims of using Box and Jenkins Prediction approach are to

look for suitable formula that will force the error term to show no change in pattern and must be as small as possible. In this study the approach is used to develop the model and do prediction of rainfall values. The Conceptual framework of Box and Jenkins modelling approach is given in the Table below (Box GE and Jenkins GM, 1976)

(Erik. Erhardt)

	Plot the Series
1	Is variance stable?
2	No, apply transformation, go to 1
3(a)	Yes, continue
3(b)	Obtain ACFs and PACFs
4	Is mean stationary?
5	No, apply regular and seasonal differencing
6(a)	Yes, continue
	Model selection
6(b)	Estimate parameter values
7	Are the residual uncorrelated?
8	No, modify the model, go to step 5
9	Yes, continue
10(a)	Forecast
10(b)	
11	

3.4.Stationarity Check

Stationary time series is attained when the probability distribution properties such as mean and variance remain fixed at a time, and they must not be uncorrelated in the series. It should be noted that, before starting any time series modelling, you must check whether or not the series is stationary and uncorrelated. A number of methods can be used to convert non-stationary time series into stationary time series. These methods are Augmented Dickey Fuller Tests (ADF), Phillips-Peron (PP) and Kwiatkowski-Phillips-Schmidt- Shin (KPSS) test, ACF and PACF plots.

3.5. Augmented Dickey Fuller Test

H0: There exists a unit root (non-stationarity)

H1: There exists a trend-stationarity.

The obtained test statistic is then compared against the critical value of the DF test given by the formula:

$$Dft = \frac{\hat{\gamma}}{SE(\hat{\gamma})},$$

where, SE is the standard error and $\hat{\gamma}$ is the estimates of least square for γ. Note that if the test statistic is less than the critical value then the null hypothesis is rejected.

Kwiatkowski-Phillip-Schmidt-Shin (KPSS) Test

The fundamental aim of this test statistic is to test the null hypothesis that is stationarity against the alternative hypothesis that is non stationary.

H0: That the series is stationary.

H1: That the series is non stationary.

The formula for KPSS is given as:

$$KPSS = T^{-2} \sum_{i=1}^{t} {S_t^2}/{\delta_T(L)},$$

Where

$S_t = \sum_{i=1}^{t} e_i,$

here e is the residual which is given by $e = [e_1, e_2, \dots, e_t]'$ and $\delta_T(L)$ is the term used to predict the variance over the long run of the random shock (residuals).

When the value of the KPSS is large, the null hypothesis is rejected, that is there is no enough evidence to support the fact that the time series move slowly along the threshold mean.

Box and Jenkins Model Identification

The model specification and selection is important when using Box and Jenkins ARIMA method. To get the suitable model of the time series observations, the analysis of autocorrelation function (ACF) and partial autocorrelation function (PACF) needs to be effectively performed. The two line graphs will show how the sequence of data in the time series are related to one another. Using the ACF gives a huge benefits of measuring the amount of linear dependence between that data in a time series that are separated by a lag k. Also the plots of PACF is used to make a decision of how many autoregressive terms are important to expose one or more of the time lags where most of the correlation appears, Seasonality of the series, trends either in mean or variance of the series (Pankratz A, 1983). Normally the graphs of ACF and PACF are plotted against the successive time lag in order to do prediction. Through these two plots the coefficients of AR and MA can be easily obtained. They are used not only to guess the form of the models but also not obtain suitable estimates for the model parameters (Box GE and Jenkins GM, 1976). In order to get acceptable model, the (AR(p)) has to be stationary and (MA(q)) has to be invertible (Czerwinski et al, 2007). At the preceding stage, the values of $p;\ q;\ P; Q$ are obtained by studying ACF and PACF plots. The identification tools for SARIMA (p; d; q)(P;D;Q)S are presented in the Table below

Table1. *Characteristics of the ACF and PACF for Pure Seasonal AR, MA and ARMA models (Shumway et al. 2006)*

	AR(P)s	MA(Q)s	ARMA(P,Q)s
ACF	Dies off at lags k's, k=1,2…	Cut off after lag Q's	Dies off at lag k's, k=1,2,…
PACF	Cut off after lag P's	Dies off at lag k's, k=1,2,…	Dies off at lag k's, k=1,2,…

Parameter Estimation

After completing identifying the tentative model, the next step is to estimates the parameters of the model. The model parameters are estimated by using the maximum likelihood estimates (MLE). Normally MLE is used in finding the parameters that maximize the probability of observations (Zakaria et al, 2012).

Model Selection

The selection of the model to be used is based on the criterion test. The Autocorrelation and Partial autocorrelation function have shown the usefulness in identifying the orders of the model. They also help in proposing where the model comes from (AIDOO, 2010), although the required model is chosen based on the test statistic such as Bayesian Information Criterion (BIC) and Akaike Information criterion (AIC). Burnham et al, (1998) assert that the crucial idea in performing the test criterion is to identify whether the required model over-fit or under-fit the observational values.

Akaike's Information Criterion

AIC (Akaike, 1976) is a tool for estimating the parameter of the likelihood function in order to find the better model which will be used for the forecasting of future observation. The best fit of the model is checked by using the AIC. Various models are tested and classified based on the number of AIC and the model with the lower number of AIC is considered the better model. The theory of assessing the model based on AIC is that the fitted data should approach the observational data. AIC does not allow additional parameter to be included in the model that is over-fitting is not given priority here. AIC is defined mathematically as:

$AIC = 2K + nlog(\frac{rss}{n}),$

where n is the number of sample size, k is the parameter of the model $(p + q + P + Q + 1)$ and rss is the residual sum of squares of the model estimated. It is important to note that if the models have equal number of AIC, then the one with a minimum number of parameters is chosen (Yurekli et al, 2005). When the sample size decreases, AIC could not give the better model, so Sugiura (1978) came with an extended model for AIC that can handle small sample observations. The model is known as Corrected AIC (AICc) and its mathematically defined as:

$$AICc = AIC + \frac{2K^2 + 2K}{n-k-1},$$

where n is the sample size for the data given and k is the number of parameters. The advantage of $AICc$ over AIC is that it gives the correct model when there are small number of sample sizes.

Bayesian Information Criterion

Bayesian Information Criterion (BIC) works very closely to AIC. It is simply defined as the method for estimating the parameter of the specified model under limited number of models. When using the BIC, it allows additional parameter to be included in the likelihood function that may give the model over-fitting. This problem can be fixed by introducing the penalized coefficient in the value of model parameter which becomes stronger than for AIC.

The BIC takes the form:

$$BIC = klog(n) + nlog(\frac{rss}{n}),$$

Where k represent the number of parameter included in the model, rss is the sum of error variance and n is the sample size. It is evident that BIC gives accurate answer when there are large number of sample and AIC is much stronger when the sample size decreases and at the same time there are larger number of parameters (McQuarrie A and Tsai C, 1998).

Diagnostic checking

The selected models must have as smaller errors as possible. Normally the model diagnostic checking is accomplished by carefully analysis of the residual series, histogram of the residual, normal QQ plots and diagnostic test (Ljung et al, 1978). So in order to check whether the model residuals follows a white noise property (that is the process is stationary and independent) then the following test is used.

Ljung - Box Test

The method is also known as the modified Box pierce statistics. This tool is used as diagnostic approach to check if there is lack of fit in the time series model. Univariate Ljung - Box test is based on the assumptions that there is no autocorrelation up to the specified lags. Ljung - Box is used to check if the residuals of the time series correlate with the white noise under the hypothesis that:

$H0$: The model does not reveal lack of fit.

$H1$: The model reveals lack of fit.

The model is defined as:

$$Q = N(N+2)\sum_{k=1}^{m} r_k^2 \Big/ (N-k),$$

where r_k^2 is the predicted time series autocorrelation at $lag\ k$, m represent the number of lags to be check and N represent available time series data. Since the model follows a chi-square distribution then it is easy to indicate the significant relationship based on the criteria for the test. In this case the chi square value needs to be compared with the tabulated values in order to evaluate the valid model otherwise the model will be rejected. The relationship that is used for statistical test is:

$$Q > \chi^2_{1-\alpha,m},$$

Here, α is the level of significance and m is the degree of freedom. When the value of Q is higher it implies that there is significant autocorrelation in the random shock of the time series and that the null hypothesis is rejected, that is residuals have no autocorrelation at the significance level (Yusof et al, 2012).

Forecasting

In order to make the correct decision about the time series prediction, suitable forecasting tools are required. The suitable selected model is not the criteria that the model is the best for prediction purposes. So, in order to get the appropriate forecasting model, measures of errors such as Mean Absolute Error (MAE), Mean Square Error (MSE) and Root Mean Square Error (RMSE) must be performed well to be assured that the obtained model is exactly the required model for forecasting the time series observations.

Mean Absolute Error

As the name explain, MAE is just the mean of the absolute errors. It is the simplest measure of prediction accuracy. The MAE tells us the deviation of the prediction from the average. The model is defined as (Park, 1999):

$$MAE = \frac{1}{2}\sum_{t=1}^{m}|\varepsilon_t|,$$

where ε_t is defined by:

$$\varepsilon_t = y_t - f_t.$$

here, ε_t stands for error term, y_t stands for observational values, f_t stands for forecasting values, t is the time and m is the total observational data (Spyros et al. 1998).

Mean Square Error

This method is sometimes termed as a good measure of overall forecasting error. It is used to measure the deviation of the squared errors for the prediction values. The MSE is defined as:

$$MSE = {}^{1}/_{n}\sum_{t=1}^{m}\varepsilon_t^2,$$

where, ε_t is defined by

$$\varepsilon_t = y_t - f_t.$$

Where, ε_t stands for error term, y_t stands for observational values, f_t stands for forecasting values, t is the time and m is the total observational data (Spyros et al. 1998).

Root Mean Square Error

The Root Mean Square Error is the known as Root Mean Squared Deviation. It is used to compute how the approximated value diverges from the actual value of the specified model. The RMSE is defined as (Park, 1999):

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(X_{predicted} - X_{actual})^2}{n}},$$

where, $X_{predicted}$ are the predicted values of the observations, X_{actual} are the actual values of the observations and n gives the total number of observations (Spyros et al. 1998).

When the value of measure of error is around zero, then it means that model has the perfect skills for forecasting or in other words we say the models has no errors (Mc- Quarrie et al. 1998). It is important to note that when the values of measures of errors are smaller, it indicates that the model is the best for forecasting purposes.

4. Results and Discussions

Preliminary Analysis

The descriptive statistics of the rainfall series recorded from January 1961 to December 2014 are illustrated in the Table 2. The mean and standard deviation of the rainfall values were 92.81 and 98.57003 while maximum and minimum values of the rainfall were 569.40 and 0.00 respectively. In Table 2, reveals that the lowest monthly rainfall in Dar es Salaam is 0.00 mm and was recorded in January 1962, June 1980, February 1984, 1989 and December 1998, and the highest rainfall is 569.40 and was recorded in April 2002. The lowest rainfall values were seen most frequently in dry seasons like January, February and June, and highest rainfall are mostly observed in rainy seasons like April. It was observed that the rainfall values were more varying (with standard deviation of 98.57003) from their mean.

Table 2. *Summary statistics of rainfall values (1st January1961 – 31th December 2014)*

Rainfall(mm)	Range	Maximum Value	Minimum Value	Mean	Standard Deviation	Variance
Values	569.40	569.40	0.000	92.81	98.57003	9716.052

Time Plot for Rainfall Data

Variation in time of the rainfall values are presented in the Time Plot. The Plot exhibited a set of values taken at different time points and graphed in a time series (Figure 2). The Smoothed plot for rainfall data in Dar es Salaam region was shown in Figure 1. From the plot, we can observe that there is seasonal cycle in the series at the same time the variance were observed to be more varying from the mean, hence this indicates that the rainfall time series is not stationary. However, for the case of trends, it is not easy to depict it. Clearly the rainfall plot seems to have the strong yearly circle.

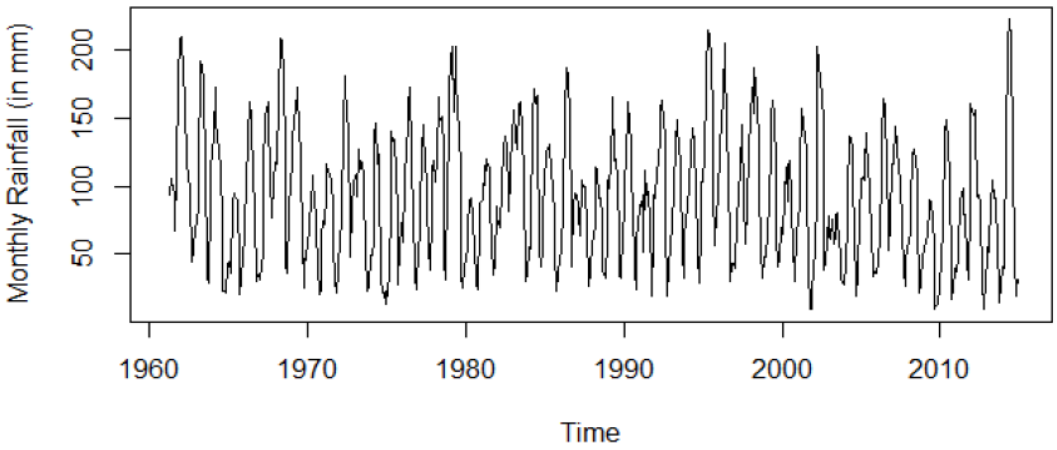

Figure2. *Smoothed Time Series Plot of Monthly Rainfall from January 1961 to December 2014*

Seasonal ARIMA Modelling

The decomposition of monthly rainfall time series data was plotted in order to see whether the time series has trends, seasonal, cyclic and random components. We plotted Year on X-axis and the observed monthly rainfall on Y-axis (Figure 2). However it has been observed that it is difficult to interpret the trends based on visual inspection technique. So in order to scrutinize the trends, we decompose the average monthly rainfall data by additive decomposition approach by the statistical software R as observed in Figure 3. From the decomposition upward trends are depicted in some years, for instance from 1961 to 1970 and 1976 to 1980. In Figure 3, it seems that there is sturdy seasonal cycle in the monthly rainfall data set.

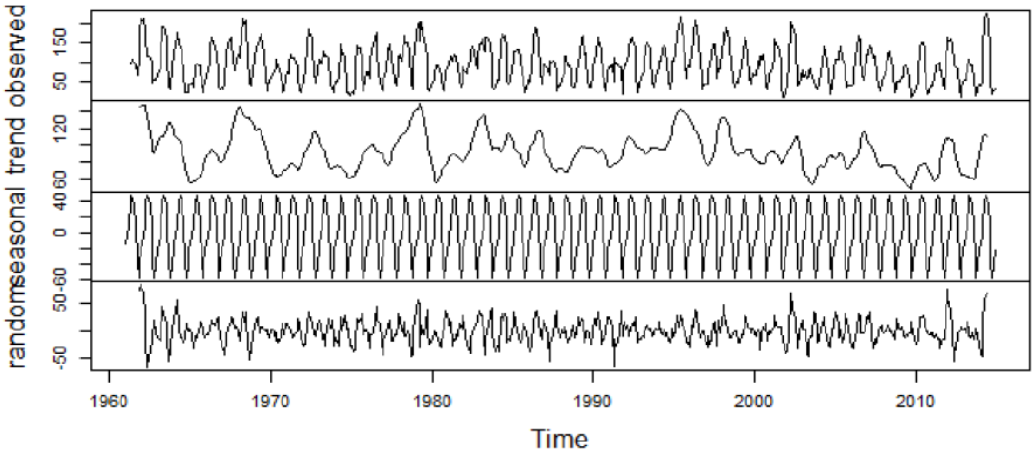

Figure3. *Decomposition of Smoothed Monthly Rainfall Series*

Ordinarily, the Box and Jenkins methodology works under assumptions that the time series are stationary and serially correlated. In this study, graphical inspection and unit root tests were the most selected techniques for stationarity test of monthly rainfall data. In Figure 3, the ACF and PACF plots are drawn using 60 lags on X-axis and values for autocorrelation on the Y-axis. The seasonal autocorrelation relationships dominate the two plots. Hence due to the appearance of strapping

seasonality and upward trends, we conclude that the average monthly rainfall series is not stationary. This result was supported by ADF test statistics.

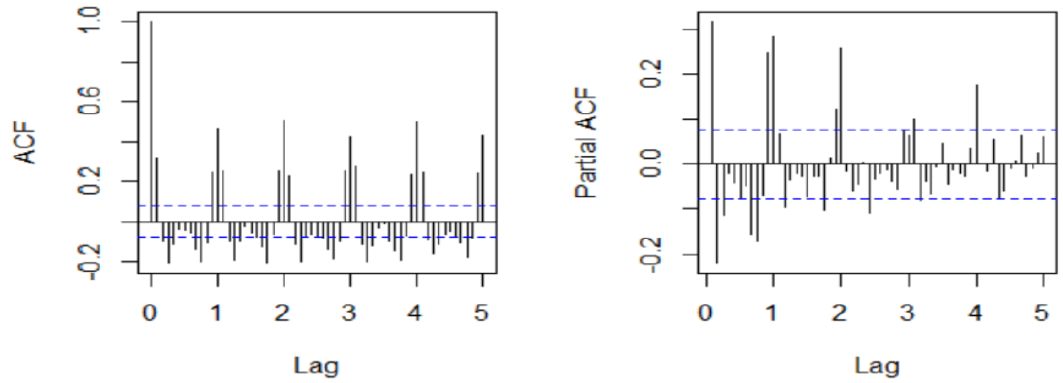

Figure4. *Autocorrelation and Partial autocorrelation function of Monthly Rainfall*

Based on Box and Jenkins approach, before fitting the model either ARIMA or SARIMA, we must make sure that the stationarity condition is attained. This is accomplished by performing seasonal differencing of the average monthly rainfall data, so as to eliminate seasonal characteristics. After conducting seasonal differencing series we use Graphical approach (ACF and PACF) and Unit root test (ADF and KPSS tests), in order to check if the stationarity condition is achieved. For the ACF and PACF plots shows that as the number of lags increases there is slow decays of ACF and PACF plots also most of the lags are outside the 95 percent confidence limits, which is the confirmation that the series is not stationary. Also the unit roots test (ADF and KPSS) gives p-values of 0.07 and 0.01. The interpretation for these two test is that, for ADF the p-value is greater than 0.05 which means the null hypothesis should not be rejected, also for the KPSS test the p-values is less than 0.05 which means that the null hypothesis should be rejected. Hence it shows that the series are not stationary. Thus after performing seasonal differencing the seasonal characteristics is not observed any more (Figure 5). Most of the spikes lies within the confidence limits except very few individual correlations appear larger compared with the confidence limits. At the non-seasonal level, results show that ACF spikes at lag 1, and goes off after lag 1 while PACF exhibit a significant spike at lag 1 and cut off at lag 1. At the seasonal level, ACF are observed to have spikes at lag 12 and also cut o_ at lag 12 while PACF goes off after lag 12. However little number of lags are faintly observed outside the confidence limits. Also the statistical tests for stationarity results show that P-value for ADF test is 0.01 and P-value for KPSS is 0.1 which is greater than 0.05 (level of significance), so we cannot reject the null hypothesis of trends-stationarity. The results from these tests concludes that the monthly rainfall differenced series is stationary.

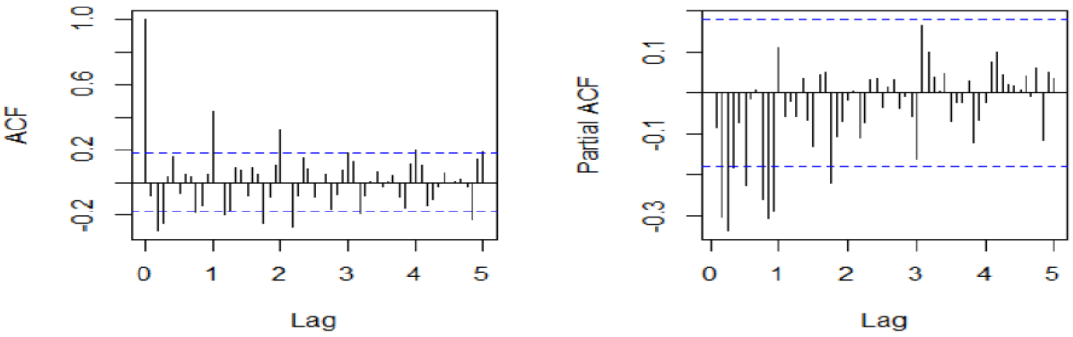

Figure5. *ACF and PACF plot for seasonal differenced Monthly Rainfall series*

Next, our goal is to look for suitable SARIMA model of monthly rainfall data from ACF and PACF plots shown in Figure 3, 4, 5 and 6. The ACF plot reveals significant spikes at lags 1, which indicates a non-seasonal $MA\,(1)$ component. Also from the same plot significant spikes are observed at Seasonal $lag\ 12$, which tells us a seasonal $MA\,(1)$ component. Moreover, in PACF plot a significant spike at $lags\ 12$, entails seasonal $AR\,(1)$. Therefore after critical examination of the two plots (ACF and PACF), the number of models has been identified and the following models were the most competing ones. $SARIMA\,(1,0,0)\times(1,0,1)_{12}$, $SARIMA\,(1,0,1)\times(1,0,0)_{12}$, $SARIMA\,(2,1,1)\times(1,1,1)_{12}$, $SARIMA\,(1,1,1)\times(1,0,1)_{12}$, $SARIMA\,(0,0,1)\,(1,0,1)_{12}\ and\ SARIMA\,(1,0,0)\times(2,0,0)_{12}$.

The parameters of the models were estimated by the maximum likelihood estimator obtained from R soft- ware and the results are presented in Table 3.

Table3. *Summary of Parameter Estimates and Selection Criteria for Rainfall Models*

Model	Parameter	Estimate	SE	t-value	p-value	Criteria
$SARIMA(1,0,0)(1,0,1)_{12}$	AR1 (ϕ_1)	0.1088	0.0392	2.7733	0.0057	AIC = 9.518676
	SAR1 (ψ_1)	0.9996	0.0009	1104.8896	0.0000	AICc = 9.521907
	SMA1 (Θ_1)	-0.9736	0.0293	-33.2478	0.0000	BIC = 8.546293
	Constant	91.7832	21.3745	4.2941	0.0000	
$SARIMA(1,0,1)(1,0,0)_{12}$	AR1 (ϕ_1)	-0.0038	0.1226	-0.0311	0.9752	AIC = 9.888768
	MA1 (θ_1)	0.2076	01255	1.6543	0.0986	AICc = 9.891999
	SAR1 (ψ_1)	0.4359	0.0404	10.7786	0.0000	BIC = 8.916385
	Constant	93.7516	6.9907	13.4109	0.0000	
$SARIMA(0,0,1)(1,0,1)_{12}$	MA1 (θ_1)	0.1116	0.0394	2.8295	0.0048	AIC = 9.518528
	SAR1 (ψ_1)	0.9996	0.0009	1104.7636	0.0000	AICc = 9.521759
	SMA1 (Θ_1)	-0.9737	0.0291	-33.4950	0.0000	BIC = 8.546145
	Constant	91.2708	21.0857	4.3286	0.0000	
$SARIMA(1,0,0)(2,0,0)_{12}$	AR1 (ϕ_1)	0.1580	0.0397	3.9803	0.0001	AIC = 9.739141
	SAR1 (ψ_1)	0.2794	0.0375	7.4415	0.0000	AICc = 9.742371
	SAR2 (ψ_2)	0.3836	0.0376	10.1881	0.0000	BIC = 8.766757
	Constant	94.0425	10.2948	9.1349	0.0000	
$SARIMA(1,1,1)(1,0,1)_{12}$	AR1 (ϕ_1)	0.10850	0.0314	3.3451	0.0009	AIC = 9.516392
	MA1 (θ_1)	-1.0000	0.0014	-704.7474	0.0000	AICc = 9.51968
	SAR1 (ψ_1)	0.9997	0.0032	-304.3712	0.0001	BIC = 8.550913
	SMA1 (Θ_1)	-0.9764	0.0048	-205.0689	0.0000	
	Constant	-0.0285	0.0176	-1.6138	0.0107	
$SARIMA(2,1,1)(1,1,1)_{12}$	AR1 (ϕ_1)	0.0951	0.0404	2.3561	0.0188	AIC = 9.512175
	AR2 (ϕ_2)	-0.0190	0.0405	-0.4703	0.0386	AICc = 9.515464
	MA1 (θ_1)	-0.9932	0.0124	-80.0806	0.0000	BIC = 8.546696
	SAR1 (ψ_1)	-0.0500	0.0456	-1.0980	0.0276	
	SMA1 (Θ_1)	-0.9845	0.0765	-12.8646	0.0000	

At least all parameters of the models are significant because the P-value are less than 0.05 (level of significance) and they should be retained in the model, except ϕ (non-seasonal AR) and θ (non-seasonal MA) with p-value greater than 0.05 (i.e. 0.9752 and 0.0986 respectively). The next step was to select the suitable model for monthly rainfall data. Based on information criterion (AIC, BIC and AICc), the lower the information criterion the better the model. So after performing AIC, AICc and BIC, $SARIMA\ (2,1,1) \times (1,1,1)_{12}$ was obtained as a convenient model for monthly rainfall data. So the step which follows is to check diagnostic of the fitted $SARIMA\ (2,1,1) \times (1,1,1)12$ model. Normally, if the model agrees with the data, then the standardized residuals approximated from the $SARIMA$ model should have the characteristics of being independent and identically distributed (IID), with mean zero and constant variance δ^2 (white noise process). Figure 7, unveils the three plots for standardized residuals, ACF of the residuals and Q-statistic from lag 1 to 12. From the standardized plot in the first panel, it is discovered that no residuals are observed outside the limits of -2 to +2, which means the model follows the white noise process. Also the Ljung-Box test was employed to check the independence of the residuals. The test results under 20 degrees of freedom has chi-square of 6.0042 and p-value of 0.9989. Thus, the test statistic of Ljung-Box that the residuals of the series are independently distributed cannot be rejected. Moreover, from the white noise test, we obtain the p-value of 0, which indicates that the residuals of the series is a white noise (with zero mean and constant variance δ^2).

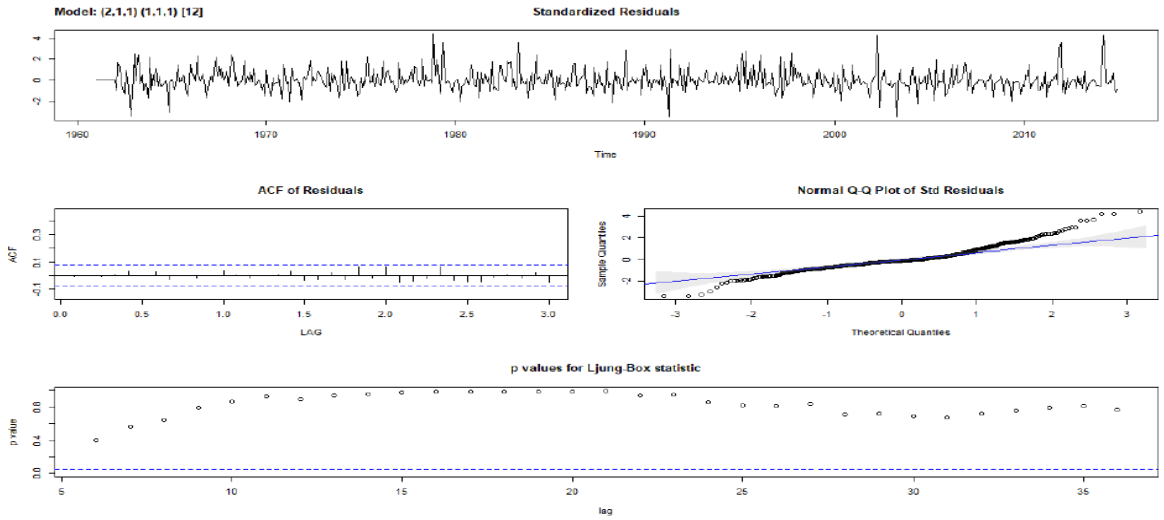

Figure6. *Residuals of Monthly Rainfall Model*

Finally, to see the clear deviation from the normality, the normal QQ-plot and histogram for the residual series are plotted. From the QQ-plot in the second panel of Figure 6, most of the points ow in a straight line, with few points are observed to be close to the line, confirming that the model residuals follows normality. The model has fulfilled all assumptions, hence $SARIMA\ (2,1,1) \times (1,1,1)_{12}$ is an appropriate model for monthly rainfall data.

Model Validation

In order to check accuracy and forecasting capability of the picked model, the actual values and the fitted ones were plotted together and presented in the Figure 7. The rainfall data from January 2005 to December 31, 2014 were designed as the test sets and were used to assess the ability of the models to fit the original data. The red and blue lines are the fitted and actual values respectively. The plot exhibited that, the fitted values are very close to the original data. This indicates that the selected model for monthly rainfall is the better one for the set of data.

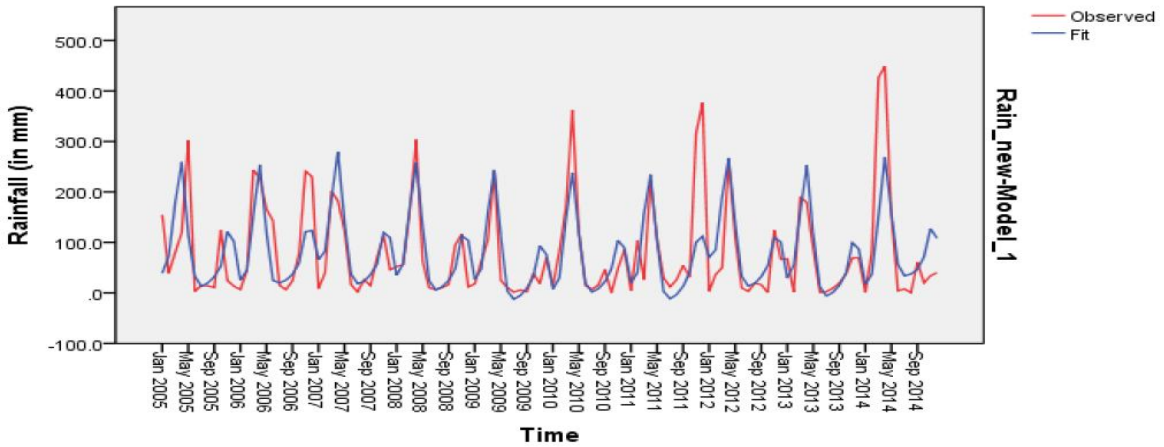

Figure7. *Observed and fitted values of Monthly Rainfall series*

Forecasting Using $SARIMA\ (2,1,1) \times (1,1,1)_{12}$

After performing the parameter estimation and conclude that all parameters were significant and then conduct diagnostic analysis and confirm that residuals followed a normal property, the next step was to do forecasting. In this case the term forecast refers to the process of predicting the future monthly rainfall of the studied time series. It should be noted that forecasting is important in decision making and planning process for all socio-economic sectors. In any time series analysis, getting the suitable model does not mean that it is a better model for prediction. Makridakis et al, (2000), asserts that the superiority of the model depends on the measure of errors. So in this study, prediction performance were judged by a number of methods, in which the measures of errors such as MAE, MASE and RMSE were used. The performance measures obtained for the monthly rainfall model respectively are shown in the Table 5 below.

Table5. *Forecasting Accuracy Statistics for Monthly Rainfall Model*

Measure of Errors	RMSE	MAE	MASE
Monthly Rainfall Model	19.06836	13.96513	0.658493

Normally, the best model must show low forecasting inaccuracy (Czerwinski et al, 2007). The performance measure of errors reported in Table 5, revealed that the prediction accuracy is high. The monthly rainfall model based on the range of monthly rainfall values show that the forecasting ability of the model is high. Thus, it is a good indication that $SARIMA\ (2,1,1)(1,1,1)_{12}$ is appropriate model for forecasting monthly rainfall values.

The forecasted values with 80% and 95% confidence limit from January 2015 to December 2024 using $SARIMA\ (2,1,1)(1,1,1)_{12}$ model is shown in Figure 10. The black color in Figure 10, indicates the actual values and the Blue color indicates the forecasted values. Most confidence limits values are observed on the negative side, however this does not make sense because the lower rainfall value recorded is zero (no rainfall). Hence, the lower negative confidence limit indicates no rainfall values are recorded. The forecasted values show that, rainfall data showed that for a Dar es salaam station, decreasing trend was observed for the rainfall seasons that is March to May (MAM) and October to December (OND), while other months rainfall values did not show any significant change.

Generally forecasted monthly rainfall was observed to have progressively decreasing trends for the upcoming ten years. In addition, it is noticed that the number of rainy seasons for Dar es Salaam will remain to be two (bimodal) that is March to May (MAM) and October to December (OND).

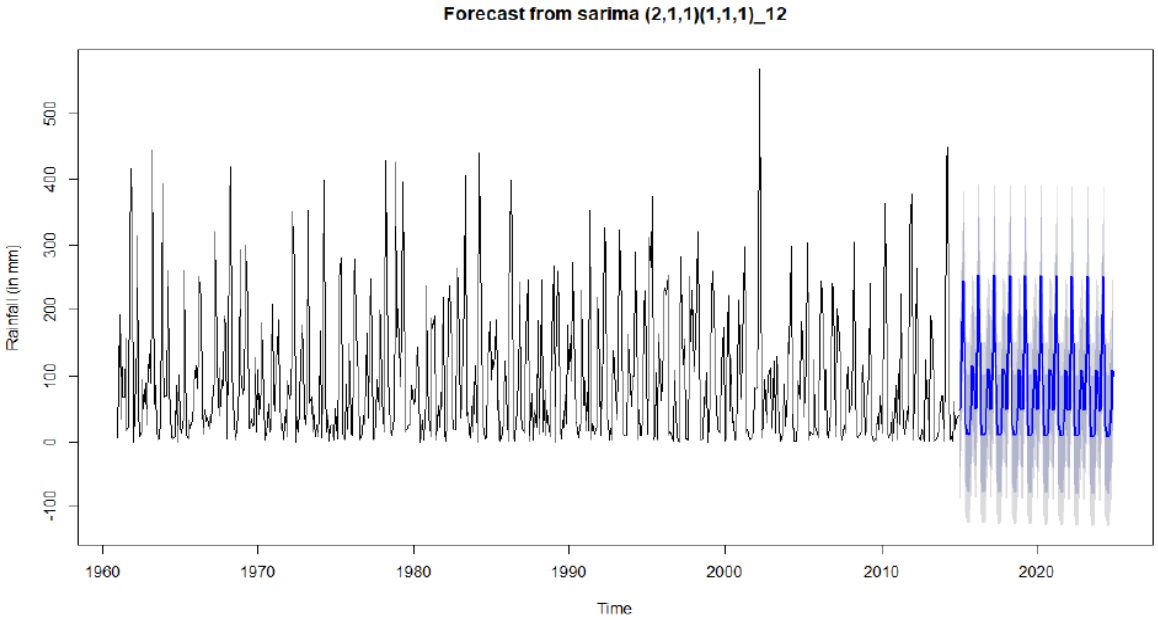

Figure8. *The forecasted values of monthly rainfall using SARIMA* $(2,1,1)(1,1,1)_{12}$ *model from January 2015 to December 2024.*

Finally, the forecasted plots from Figure 8, was observed to have minimal spread of confidence intervals from 2015 to 2020. However as time goes for example, from 2021 to 2024 the spread of confidence intervals seems to be higher implying that uncertainty of prediction becomes larger. Hence we again realized that the Box and Jenkins Seasonal ARIMA approach is the good method for short period of time forecasting of meteorological variables such as rainfall. Such result is supported by few studies like that of Erhardt, E., (2015).

ACKNOWLEDGEMENTS

I thank Tanzania Meteorological Agency (TMA) for providing their data; African Institute of Mathematical Sciences (AIMS) Tanzania for their financial supports.

REFERENCES

[1] AIDOO, E. (2010). Modelling and forecasting inflation rates in Ghana: An application of SARIMA models, Ghana

[2] Akaike, H. (1974). A new look at the statistical model identification. IEEE transactions on automatic control

[3] Box, G.E. & Jenkins, G.M. (1976). Time series analysis, control, and forecasting. San Francisco, CA: Holden Day, 3226(3228): 10.

[4] Chang'a, L.B., Kijazi, A.L., Luhunga, P.M., Ng'ongolo, H.K. & Mtongori, H.I. (2017). Spatial and Temporal Analysis of Rainfall and Temperature Extreme Indices in Tanzania.

[5] Czerwinski, I. A., Gutierrez, J. C., Estrada. & Hernando, J. A. (2007). Casal Short-term forecasting of halibut CPUE, Linear and non-linear univariate approaches. Fisheries Research

[6] Erhardt, Erik, (2015). Box-Jenkins Methodology vs Rec.Sport. Unicycling 1999-2001. Available at : http : www:statacumen:com=pub=proj=WPI=ErhardtEriktsaproj.pdf (accessed 10- August)

[7] Ghahraman, Bijan, (2007). Time trend in the mean annual temperature of Iran. Turkish journal of agriculture and forestry 30: 439-448.

[8] IPCC. (2012). Climate Change, Managing the risks of extreme events and disasters to advance climatic change adaptation. The Physical Science Basic, Summary for policymakers. Contribution of working Group 1 to the Fourth Assessment Report, Paris

[9] Kijazi, A.L. & Reason, C.J.C. (2009). Analysis of the 1998 to 2005 drought over the northeastern highlands of Tanzania. Climate Research, 38(3): 209-223.

[10] Ljung, G. M. & Box, G. E. P. (1978). On a measure of lag of _t in time series model, Biometrika, pp 65:2, 297-303

[11] McQuarrie, A.D. & Tsai, C. L. (1998). Regression and time series model selection. World Scientific

[12] Mishra, A. K. & Desai, V. R. (2005). Drought Forecasting Using Stochastic Models. Stochastic Environmental Research and Risk Assessment, pp 19(5): 326-339

[13] Mishra, U. & Jain, V. K. (2010). ARIMA forecasting of ambient air pollutants, Stochastic Environ. Res. Risk Assessment: 24:751-760. DOI: 10.1007/s00477-009-0361-8

[14] Montgomery, D. C. & Johnson, L. A. (1967). Forecasting and Time Series Analysis. McGraw-Hill Book Company, http://www.abebooks.com/Forecasting-Time Series-Analysis-Montgomery- Douglas/1323032148/bd, McGraw-Hill

[15] Nury, A.H., KochMand. & Alam, M.J.B. (2013). Time Series Analysis and Forecasting of Temperatures in the Sylhet Division of Bangladesh. In 4th International Conference on Environmental Aspects of Bangladesh (ICEAB), August: 24-26).

[16] Ojija, F., Abihudi, S., Mwendwa, B., Leweri, C.M. & Chisanga, K. (2017). The Impact of Climate Change on Agriculture and Health Sectors in Tanzania: A review. International Journal of Environment, Agriculture and Biotechnology, 2(4).

[17] Orindi, V.A. & Murray, L.A. (2005). Adapting to climate change in East Africa: a strategic approach (No. 117). International Institute for Environment and Development.

[18] Pankratz, A. (1983). Forecasting With Univariate Box-Jenkins Models Concepts and Cases. ISBN 0-471-09023- 9, pp: 414, John Wiley and Sons, Inc. New York, USA

[19] Shahin, M.A., Huq, M.M. & Ali, S. (2016). Time Series analysis, MOD Climate variable rainfall District in BANGLADESH.

[20] United Republic Of Tanzania (URT, 2007). Tanzania National Adaptation Programme of Action 2007, Vice President's Office-Dar es Salaam

[21] Urban Africa Risk Knowledge (UARK, 2017). Dar es Salaam Climatic Profile: Full Technical Version

[22] World Population Review (WPR, 2019): http://worldpopulationreview.com/world-cities/dares-Salaam-population

[23] Yurekli, K., Kurunc, A. & Ozturk, F. (2005). Testing the residuals of an ARIMA model on the Cekerek Stream Watershed in Turkey. Turkish Journal of Engineering and Environmental Sciences

[24] Yusof, F. & Kane, I. L. (2012). Modelling monthly rainfall time series using ETS state space and SARIMA models. International Journal of Current Research

[25] Zakaria, S., Al-Ansari, N., Knutsson, S. & Al-Badrany, T. (2012). ARIMA Models for weekly rainfall in the semi-arid Sinjar District at Iraq

Citation: *Paul Andrew Panga, et.al., Forecasting Rainfall in Tanzania Using Time Series Approach Case Study: Dar es Salaam, International Journal of Scientific and Innovative Mathematical Research (IJSIMR), vol. 8, no. 4, pp. 14-27, 2020. Available : DOI: https://doi.org/10.20431/2347-3142.0804002*