## *Original Research Article*

**Application seasonal autoregressive moving average models to analysis and forecasting of Time Series Monthly Rainfall Patterns in Embu County, Kenya**

UNDER PEER REVIEW

**Abstract:**

Rainfall is of critical importance for many people, particularly those whose livelihoods depend on rain-fed agriculture. Predicting the trend of rainfall is a difficult task, and statistical approaches such as time series analysis provide a means for predicting the patterns of rainfall. The models also offer the potential to improve areas such as increased food production, profitability, and improved food security policing. However, these forecasts and information systems may, in some instances, not be suitable for direct use by stakeholders in their decision-making. The objective of this study was to investigate rainfall variability and develop a Seasonal Auto-Regressive Integrated Moving Average (SARIMA) model for fitting the monthly rainfall using time series data. Secondary monthly data from 1998 to 2017 for Embu County was collected from the Kenya Meteorological Department, Embu and recorded into an excel sheet. R-software was utilized to analyse data for descriptive statistics, rainfall variability, and model fitting. The coefficient of variation for annual and seasonal rainfall was calculated. The Box Jenkin's ARIMA modelling procedure (model identification, model estimation, model validation) was used to determine the best models for the data. The main study findings indicated the existence of annual variability of 34%, March-April-May rainfall variability of 44%, and October-November-December variability of 44%. A first-order differenced SARIMA (1, 1, 1) $(0, 1, 2)_{12}$ model with an AIC score of 9.99356 was found suitable for predicting rainfall pattern in Embu, County. The study outcome revealed that Embu County experiences high seasonal and rainfall variation of rainfall, thus requires a reliable model for better prediction.

**Keywords:** Rainfall forecasting, Time Series Analysis, SARIMA, Residual Analysis.

### 1 Introduction

Seasonal rainfall forecasts are critical in rain-fed farming regions. In Africa, particularly in rural areas, the primary source of livelihood is agriculture that relies on rainfall. Empirical studies among African farmers have revealed that climate forecasts are capable of helping farmers to reduce their vulnerability to drought and adverse effects of climate change (Roudier *et al*., 2014). The predictions can also allow subsistence farmers to maximize opportunities when favorable rainfall conditions are predicted and used to make decisions.

The assessment of the potential of statistical forecasts in natural phenomena such as rainfall has ignited scientific and institutional processes for developing and disseminating climate forecasts in Africa.

Kenya's socio-economic activities to a greater extent depend on rainfall performance and distribution (Huho & Mugalavai, 2010; Wakachala *et al.,* 2015) with about 68% of these activities being weather and climate dependent. Approximately 60% of the world population is affected by low rainfall or altogether drought. About 630 million people in Africa live in Arid and semi-arid areas, which receive low or no rainfall and mainly engage in rain-fed subsistence farming for their livelihoods (Huho & Mugalavai, 2010). Arid and semi-arid areas in Kenya provide a home to about 30% of the human population and 50% of its livestock population (GOK, 2004). These areas receive low and erratic rainfall that is highly variable both in time and space, causing severe food shortages and deaths of livestock (Kisaka *et al*., 2013). Huho and Mugalavai (2010) argue that agriculture supports about 75% of the Kenyan population and generates almost all the country's food requirements (Metrine *et al*., 2015).

Wang *et al*. (2012) used a seasonal autoregressive moving average (SARIMA) model to simulate and forecast the seasonal precipitation series of Shouguang city, China. In their study they identified and fitted the data to four models namely SARIMA (2, 0, 2) $(1, 1, 1)_{12}$, SARIMA (2, 1) $(1, 1, 1)_{12}$, SARIMA (1, 1) $(1, 1, 1)_{12}$ and MA (12). After comparing the models based on available information criteria, they have argued that SARIMA (2, 0, 2) $(1, 1, 1)_{12}$ is the better one and used it for forecasting. Given an extensive time-series data set, ARIMA and SARIMA methods show high forecast accuracy. Adede (2018) used Autoregressive (AR), moving average (MA), autoregressive moving average (ARMA) and autoregressive integrated moving average (ARIMA) models to analyse annual rainfall of Debre Markos Town, Ethiopia. In his study, he identified AR (2), MA (1) and ARMA (2, 1) to be capable in describing annual rainfall time series. He further argued that ARIMA (2, 1) was the best fitting model for modelling and yearly rainfall forecasting.

Application of other linear stochastic methods has also resulted in inaccurate predictions, clearly indicating that linear statistical models do not accurately represent historical data and hence are not acceptable methods for a non-linear application such as flood forecasting (Ampaw *et al*., 2013). Mohamed and Ibrahim (2016) used linear stochastic models based on multiplicative SARIMA to simulate monthly rainfall data of Nyala station in Sudan. They carried out a first-order seasonal differencing to remove seasonality in the data and found that SARIMA $(0,0,0)x(0,1,1)_{12}$ model developed was the best fitting model to the monthly rainfall simulated data. Papalaskaris *et al.* (2016) applied stochastic time series models in forecasting rainfall patterns and trend of Kavala city, Greece. In their study, they found that among all the SARIMA models fitted SARIMA $[(0, 0, 0) \times (0, 1, 1)_{12}]$ model best fitted the total recorded monthly rainfall data of Kavala city in the period 2006 to 2014. Khan *et al*. (2014) proposed models $SARIMA(0,0,0)(1,0,3)_{12}$, $SARIMA(0,0,0)(1,0,1)_{12}$, $SARIMA(0,0,0)(1,0,2)_{12}$ and $SARIMA(0,0,0)(1,0,1)_{12}$ for maximum and minimum temperature, rainfall, and humidity on the basis of Akaike Information Criteria and Log likelihood have been captured most seasonality of the data.

Afrifa-Yamoah *et al.* (2016) used SARIMA models to fit rainfall patterns with data collected from the Department of Meteorology and Climatology in Ghana. The result showed

UNDER PEER REVIEW

that the region experienced much rainfall in September and October and the lowest amount of rainfall in January, December, and February. SARIMA (0, 0, 0), $(1, 1, 1)_{12}$, was identified as the most appropriate model for prediction of monthly average rainfall figures for the Brong Ahafo Region of Ghana. Sopipan (2014) forecasted rainfall in Thailand using SARIMA and Artificial Neural Network (ANN) models had been used to predict atmospheric variables, including precipitation in Kenya. Valipour (2015) studied the ability of the seasonal autoregressive integrated moving average (SARIMA) and autoregressive integrated moving average (ARIMA) models in investigating long-term runoff forecasting in the United States. In the first stage, the amount of runoff was predicted for 2011 in each US state using the data from 1901 to 2010. The results revealed that the accuracy of the SARIMA model is better than that of the ARIMA model.

The Box-Jenkins Seasonal ARIMA (SARIMA) model has several advantages over other models, particularly over exponential smoothing and neural network, due to its forecasting capability and richer information on time-related changes (Mahmud, Bari and Rahman, 2016). Kibunja *et al*. (2014) studied the effectiveness of SARIMA model in forecasting precipitation in Mount Kenya region and concluded that the model was good. Kibunja *et al*. (2014) studied the effectiveness of the SARIMA model in forecasting precipitation in the Mount Kenya region and concluded that the model was good. Kane and Yusof (2012) also analyzed the precipitation forecast using a SARIMA model in Golestan province and found the seasonality measure in SARIMA to be highly useful in modelling precipitation.

This paper is organized as follows: Section 2 gives the source of data and methodology, including a brief overview of SARIMA models. Section 3 provides data analysis and discussion of results. Section 4 ends the paper with some concluding remarks.

### 1.1 Objectives of the Study

The general aim of this study was to investigate rainfall variability and develop a Seasonal Auto-Regressive Integrated Moving Average (SARIMA) model for fitting the monthly rainfall using time series data in Embu County.

#### 1.1.1 Specific Objective

i. To determine rainfall variation in Embu County.
ii. To develop a Seasonal Auto-Regressive Integrated Moving Average (SARIMA) model for fitting the monthly rainfall in Embu County.

## 2 Materials and Methods

### 2.1 Data Source

The study covered Embu County, Kenya located approximately between latitude $0^o$ 8’ and $0^o$ 50’ South and longitude $37^o$ 3’ and $37^o$9’ East. The county is on the South-eastern side of Mount Kenya. Embu County borders Tharaka-Nithi to the North, Machakos to the South, Kirinyaga, and Muranga to the West, and Kitui to the East (Embu County Government, 2017). The county's location is at the foothill of Mount Kenya and altitude of between 1179 and1350 meters above the sea level. It records an approximate temperature ranging between 9°C – 28.8°C and 640mm and 1206mm of average rainfall annually (Embu County, 2016).

UNDER PEER REVIEW

Rainfall in Embu County is bimodal with short rains from mid-October to December and the long rains from March to May. This indicates that the region has two cropping seasons every year with the main crops being maize, beans, and livestock rearing. The lower parts of Embu, which includes Mbeere Sub-county, experience more moderate rainfall of between 640mm and 781 mm annually, that supports the growth of crops such as green grams, cowpeas, beekeeping, livestock rearing, and Miraa farming. Embu County produces about 20 percent of the nation's maize due to its fertile Nitosols.

The county's population is about 516,212 persons and experiences an annual growth rate of about 1.7% per year, according to the Kenya Population and Housing Census 2009 (Embu County Government, 2017). The population density is approximately 82 persons per square kilometre, with many households owning less than 5.0 hectares of land (Embu County Government, 2017). The projected population in 2017 is 519,415, indicating the need for increased agricultural production to support the people. The county is an agricultural region with the people depending on farming and livestock rearing as the main economic activities, as 70% of the residents engage in small-scale farming. The food crops include maize, beans, cassava, sweet and iris potatoes, bananas, and sorghum, among others. Cash crops include coffee, tea; macadamia, and dairy farming (Kenya Ministry of Lands & Physical Planning 2016). Thus, rainfall plays a significant role in the survival of the residents since most of them are subsistence producers. Cash crops produced in the county include tea and coffee, although some people practice daily farming.

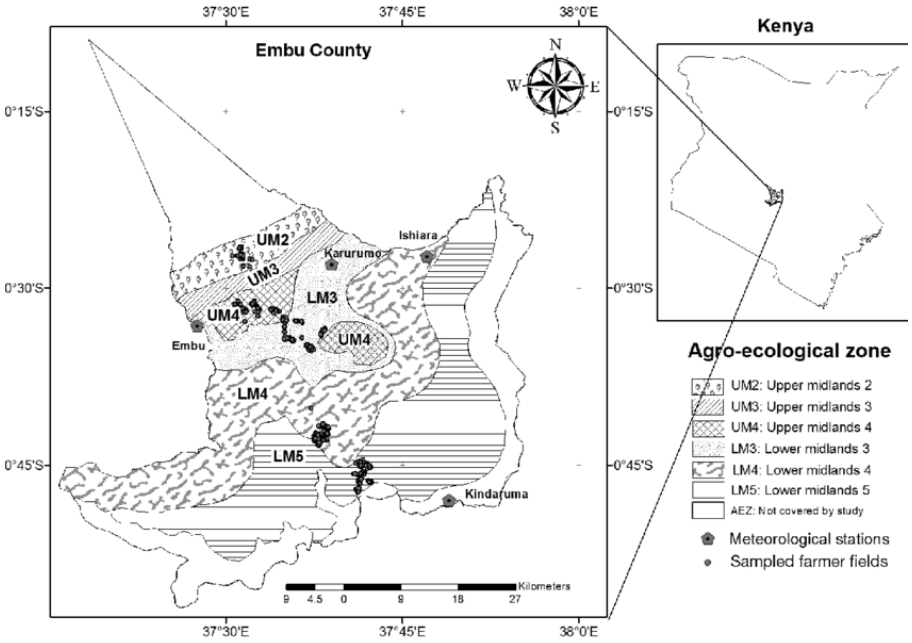


Figure 1: Embu County Map agroecology

## 2.2 Autoregressive AR Process

Let $Y_t$ be a discrete time series variable, which takes different variable over a period of time. The corresponding AR (p) model of $Y_t$ series, which is the generalizations of the autoregressive model, is expressed as;

$$Y_t = \theta_0 + \theta_1 Y_{t-1} + \theta_2 Y_{t-2} + \cdots + \theta_p Y_{t-p} + \varepsilon_t \quad (1)$$

Where $Y_t$ is the response variable at time t, $Y_t, Y_{t-1}, Y_{t-2} \ldots Y_{t-p}$ are the respective variables at varying time lags, $\theta_0, \theta_1, \theta_2, \ldots \theta_p$ are the coefficients and $\varepsilon_t$ is the error factor or white noise. Introducing a lag operator B the equation becomes

$$Y_t = (1 - \emptyset_1(B) - \emptyset_2 B^2 \ldots - \theta_p B^p) y_t = \emptyset_p(B) y_t = \epsilon_t \quad (2)$$

## 2.3 Moving Average Process

MA (q) model, which is the generalization of the moving average model, is specified as;

$$\mu_t = \varepsilon_t + \sigma_1 \varepsilon_{t-1} + \sigma_2 Y_{t-2} + \cdots + \sigma_q \varepsilon_{t-q} + \varepsilon_t \quad (3)$$

In which $\varepsilon_t \sim WN(0, {\sigma_t}^2)$ and $\varepsilon_t$ is the error term. The process uses past errors in predicting the variables in which the residuals are assumed to follow a normal distribution. Introducing a lag operator B to the equation, it becomes

$$Y_t = (1 - \theta_1(B) - \cdots - \theta_q B^q)\epsilon_t = \theta_\varphi(B)\epsilon_t \quad (4)$$

## 2.4 Autoregressive Integrated Moving average (ARIMA) Process

ARMA models may not be adequate to effectively describe the non-stationary time series, which are more frequently encountered in actual practice. The ARIMA model, which is a generalization of an ARMA model to include the case of non-Stationarity, is more appropriate. When using the ARIMA model, finite differencing is applied to the data to remove non-stationarity. When ($Y_t$) in the data is replaced with ( $\Delta Y_t = Y_t - Y_{t-1}$ ), then the ARMA models become the ARIMA (p,d,q) models, where p is the order of autocorrelation (Indicates weighted moving average over past observations), d is the order of integration (differencing) and q is the order of moving averaging. By combining the models in (1) and (2), this is referred to as ARMA to model, which have the general form of;

$$Y_t = \theta_0 + \theta_1 Y_{t-1} + \theta_2 Y_{t-2} + \cdots + \theta_p Y_{t-p} + \varepsilon_t + \sigma_0 \varepsilon_{t-1} + \cdots + \sigma_q Y \varepsilon_{t-q} \quad (5)$$

If $Y_t$ is stationary at level d(0) or at first difference d(1), then this determines the order of integration. To identify the order of p and q, the ACF and PACF are applied. For this study, the ARIMA model $Y_t = \theta_0 + \theta_1 Y_{t-1}$ where $Y_t$ is rainfall in millimeters, $\theta_0$ is the intercept and $\theta_1$ is time in months was used.

## 2.5 SARIMA Models

SARIMA models are an adaptation of autoregressive integrated moving average (ARIMA) models to fit seasonal time series specifically. That is, their construction takes into account the underlying seasonal nature of the series to be modelled. Seasonality in a time series refers to a regular pattern of changes that repeats over in time-periods, where S defines the number of time-periods until the pattern repeats. For monthly rainfall data S = 12. In a seasonal ARIMA model, seasonal AR and MA terms predict $x_t$ using data values and errors at times with lags that are multiples of *S* (the span of the seasonality). The seasonal ARIMA model incorporates non-seasonal and seasonal factors in a multiplicative model and is denoted as

ARIMA $(p, d, q) \times (P, D, Q)S$,

Where $p$ = non-seasonal AR order, $d$ = non-seasonal differencing, $q$ = non-seasonal MA order, $P$ = seasonal AR order, $D$ = seasonal differencing, $Q$ = seasonal MA order, and $S$ = time span of repeating seasonal pattern.

Without differencing operations, the model can be written as

**Comment [U1]:** Few lines may be needed buttress on the stationarity condition of AR after equation 2

**Comment [U2]:** Mathematical expression of the invertibility condition of MA model may be included after equation 4

**Comment [U3]:** Stationarity and invertibility condition of ARIMA is an important aspect of ARIMA model and I think it inclusion in this section is a plus to the study

UNDER PEER REVIEW

$$\Phi(BS)\varphi(B)(xt - \mu) = \Theta(BS)\theta(B)wt \quad (6)$$

The non-seasonal components are:

AR: $\varphi(B) = 1 - \varphi 1B - \ldots - \varphi pBp$

MA: $\theta(B) = 1 + \theta 1B + \ldots + \theta qBq$

The seasonal components are:

Seasonal AR: $\Phi(BS) = 1 - \Phi 1BS - \ldots - \Phi PBPS$

Season MA: $\Theta(BS) = 1 + \Theta 1BS + \ldots + \Theta QBQS$

### 2.5.1 Model Identification in SARIMA

The first step of applying the model is to identify the appropriate order of ARIMA (p,d,q) model. Identification of the ARIMA model involves selection of the order of AR(p), MA(q) and I(d). The order of d is estimated through I(1) or I(0) process. The model specification and selection of order p and q involves plotting of ACF and partial PACF or correlogram of variables at different lag lengths. Box-Pierce Q statistics and Ljung-Box LB statistics measures the significance level of individual coefficients. The Box-Pierce Q statistics is defined as;

$$Q = \sum_{k=1}^{m} \widetilde{\rho_k}^2 \sim {x_m}^2 \quad (7)$$

Where n=sample size and m is lag length. And Ljung Box (LB) Statistics is defined by

$$LB = n(n+2)\sum_{k=1}^{m} \frac{\widetilde{\rho_k}^2}{n-k} \sim {x_m}^2 \quad (8)$$

Where n=sample size and m is the lag length of the date. The possible SARIMA model is determined that best fit the data under consideration. SARIMA model is appropriate for stationary time series; therefore, the data under consideration must satisfy the condition of stationarity that is the mean, and variance and autocorrelation are constant over time.

**Comment [U4]:** This statement is better express as "The possible SARIMA model that best fit the data under consideration is determined by selection criteria"

### 2.5.2 Parameter Estimation SARIMA

To estimate SARIMA models, the ML method is used. Under the assumption of independent and distributed standardised $z_t$, the log-likelihood (LL) function of $\{y_t(\theta)\}$ for a T observations sample, is given by:

$$\ln L[(y_t), \theta] = \sum_{t=1}^{T} \left[ ln[D(z_t(\theta), v)] - \frac{1}{2} ln[\sigma_t^2(\theta)] \right] \quad (9)$$

where, $\theta$ is the vector of the parameters that have to be estimated for the conditional mean, conditional variance, and density function. $z_t$ is a sequence of independent and distributed random variables with mean as zero and variance as one. The approach of maximum likelihood (ML) requires the specification of a particular distribution for a sample of T observations $y_t$

$$f(Y_t, Y_{t-1}, \ldots, Y_t = yt) = f(y_{T-1}, \ldots y_t|\Psi) \quad (10)$$

denote the probability density of the sample given the unknown parameters $(n \times 1)$ parameters $\Psi$. Following the notation of Box and Jenkins, $L(\Psi|y)$ with respect to derivatives to zero and using vector notation and suppressing y the result becomes part of the unknown parameters of the vector $\Psi$ the notation is the most appropriate. Setting the $\frac{\partial(\varphi)}{\partial\varphi} = 0$

As a rule, the likelihood equations are non-linear. Therefore, the ML estimates must be found in the course of an iterative procedure.

### *2.5.3* Model diagnostic checking for SARIMA model

After estimating the parameters of our chosen model, the last step is model diagnostics. At this stage, we determine the adequacy of the selected model. One assumption of the SARIMA model is that the residuals of the model should be white noise. The ACF of the residuals is approximately zero when the residuals are white noise. Ljung-Box statistic proposed by

UNDER PEER REVIEW

Ljung and Box (1978) (Kowal, 2015) is used to check if a given observable series is linearly independent. The test examines the null hypothesis of linear independence of the series.

#### 2.5.4 Forecasting with the SARIMA model

Forecasting is the process of making a statement about events whose actual outcomes have not yet been observed. It is an important application of time series. After the model has passed the entire diagnostic test, it becomes adequate for forecasting, which the last step is in Box-Jenkins model building approach. For instance, let us consider the given Seasonal ARIMA (0, 1, 1) $(1, 0, 1)_{12}$ we can forecast the next step which is given by Cryer and Chan (2008).

$$z_t - z_{t-1} = \Phi(z_{t-12} - z_{t-13}) + \varepsilon_t - \theta\varepsilon_{t-1} - \Phi\varepsilon_{t-12} + \theta\varepsilon_{t-13}$$

$$z_t = z_{t-1} + \Phi z_{t-12} - \Phi z_{t-13} + \varepsilon_t - \theta\varepsilon_{t-1} - \Phi\varepsilon_{t-12} + \theta\varepsilon_{t-13} \quad (11)$$

The one step ahead forecast from the origin t is given by

$$\hat{z}_{t+1} = z_t + \Phi z_{t-11} - \Phi z_{t-12} - \theta\varepsilon_t - \Phi\varepsilon_{t-11} + \theta\varepsilon_{t-12} \quad (12)$$

The next step is

$$\hat{z}_{t+2} = \hat{z}_{t-1} + \Phi z_{t-10} - \Phi z_{t-11} - \Phi\varepsilon_{t-10} + \theta\varepsilon_{t-11} \quad (13)$$

and so on. The noise term $\varepsilon_{13}, \varepsilon_{12}, \varepsilon_{11}, \ldots., \varepsilon_1$ (as residuals) will enter into the forecasts for lead times$l = 1,2, \ldots,13$, but for $l > 13$ the autoregressive part of the model takes over;

$$\hat{z}_{t+l} = \hat{z}_{t-l+1} + \Phi z_{t+l-12} - \Phi z_{t+l-13}, for\ l > 13 \quad (14)$$

### 2.6 Forecasting Performance

The accuracy for each model can be checked to determine how the model performed in terms of in-sample forecast. In terms of out sample forecasting, some of the observations are left out during model building. The accuracy of the model can be compared using forecast measure or some statistic such as mean error (ME), root mean square error (RMSE), mean absolute error (MAE), mean percentage error (MPE), mean absolute percentage error (MAPE), and mean square error (MSE) among others (Cryer & Chan, 2008). The model with the minimum of MAE, MAPE, or RMSE is considered to be the best for forecasting. The mathematical expressions are defined as:

$$MAE = \frac{1}{T}\sum_{t=1}^{T}|\hat{y}_t - y_t| = \frac{1}{T}\sum_{t=1}^{T}|e_t| \quad (15)$$

$$MSE = \frac{1}{T}\sum_{t=1}^{T}(\hat{y}_t - y_t)^2 = \frac{1}{T}\sum_{t=1}^{T}(e_t)^2 \quad (16)$$

$$RMSE = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(\hat{y}_t - y_t)^2} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(e_t)^2} \quad (17)$$

Where $y_t$ is the actual observation, $\hat{y}_t$is fitted, or the forecast value and T is the sample size. If we have perfect forecast then MAE = MSE = RMSE = 0. The smaller the value, the better the prediction, and the great the value, the poorer the predictive power of the model.

**Comment [U5]:** I noticed that MAPE was used in table 6 to determine forecast accuracy however its mathematical expression is missing among other accuracy measures. Kindly address as appropriate

## 3 Results and discussion

Descriptive statistics obtained for various rainfall patterns in Embu County reveal that the highest annual rainfall sum recorded was 1824.2mm in 2002 and the lowest is 79.6mm in 2003 (Table 1). The highest monthly rainfall is 820.70mm in 2002 Table 2 and Figure 2. Some months recorded no rainfall such as January, February, June, and September 2003 (Figure 2). In addition, the months of March, April, and May recorded high rainfall totals

UNDER PEER REVIEW

during the first season while October and November received high rainfall totals in the second part of the year.

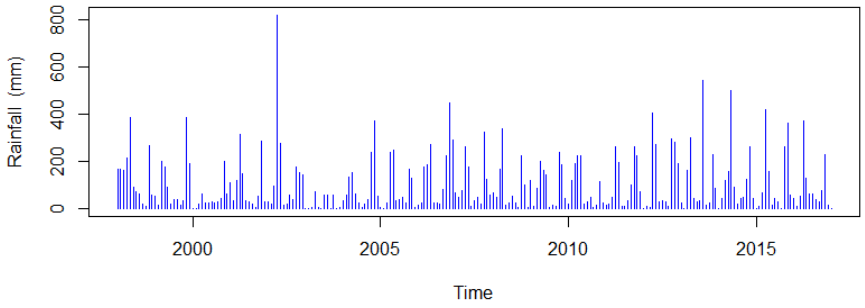


**Figure 2: Plot of Monthly-time rainfall data**

The month of April recorded the highest amount of rainfall, followed closely by March while the months of January and February recorded the least amount of rainfall. August received low amount except in 2013 when the rainfall totalled 543mm, which was an abnormal amount compared to the rest of the years (Table 1). The study reveals presence of variability of rainfall distribution over years in the area under study (Figure 3). However, no particular trends are traced on the monthly or annual rainfall.

**Table 1: Monthly Descriptive Statistics**

| Month | N | Minimum | Maximum | Mean | Std.dev |
|---|---|---|---|---|---|
| January | 20 | 0 | 166.7 | 37.605 | 46.71566 |
| February | 20 | 0 | 164.7 | 32.16 | 41.57626 |
| March | 20 | 3.2 | 197.1 | 95.79 | 64.92672 |
| April | 20 | 60.1 | 820.7 | 265.45 | 165.2701 |
| May | 20 | 1.9 | 499 | 180.775 | 127.9435 |
| June | 20 | 0 | 144 | 33.605 | 35.81813 |
| July | 20 | 3.7 | 69.8 | 31.13 | 18.01766 |
| August | 20 | 13.9 | 543 | 60.24 | 114.5205 |
| September | 20 | 0 | 100.6 | 25.755 | 26.40563 |
| October | 20 | 10.2 | 323.3 | 152.56 | 105.8849 |
| November | 20 | 0.7 | 446.1 | 229.43 | 109.9802 |
| December | 20 | 2 | 291.6 | 70.675 | 76.02191 |

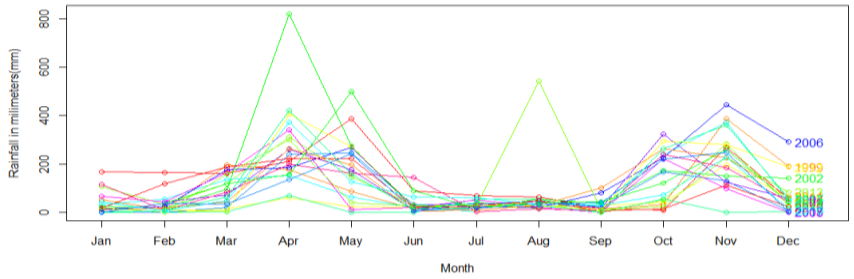


**Figure 3**: Seasonal distribution of rainfall in Embu County

UNDER PEER REVIEW

Analysis of March-April-May (MAM) and October-November-December (OND) rainfall indicated that the MAM season received the highest amount of rainfall (Table 2). Seasonal variability analysis for the MAM and OND rainfall showed irregular rainfall patterns in the study area with the coefficient of variability for MAM and OND classified as high each with a value of 44% percent. Similarly the coefficient of variation for annual rainfall is also high with a value of 34% (Table 2 & 3).

**Table 2**: Coefficient of Variation for MAM and OND (STD/Mean) 100%

| | N | Mean | STD Deviation | CV | Sum |
|---|---|---|---|---|---|
| MAM Rainfall | 20 | 536.0050 | 235.33624 | 44% | 10720.1 |
| OND Rainfall | 20 | 416.1650 | 182.80410 | 44% | 8323.3 |

The coefficients of variation are 0.44 for MAM and OND amount, which is classified as high.

**Table 3**: Variability test for mean annual rainfall totals from 1988 to 2017

| | N | Mean | Std. Deviation | CV | Sum |
|---|---|---|---|---|---|
| Rainfall | 20 | 1199.4750 | 404.85692 | 34% | 23989.5 |

The coefficients of variation were concluded based on the Hare (2003) provisions of rainfall variability coefficients. The irregular patterns in the two seasons make it difficult for farmers to make decisions on the type of agricultural practices to engage in. Annual rainfall is also highly variable with a coefficient of variation of 34 percent making it a challenge for stakeholders such as county planners, businesspersons, and other agricultural officials to make reliable decisions. The results are in agreement with Kisaka *et al*. (2014) study that examined the extent of seasonal rainfall variability using rainfall anomaly index, coefficient of variance, and probability analysis. According to this research Embu showed a 90 percent chance of below cropping threshold rainfall. The research also showed a high seasonal variability of 0.56, 0.47, 0.59, and 0.36 in regions such as Machang'a, Kiritiri, and kindaruma, and Embu.

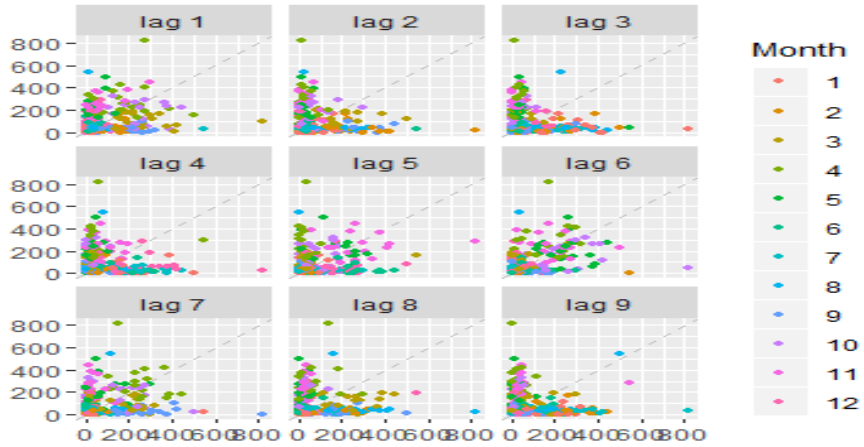


**Figure 4: Scatter plot at various lags.**

Figure 4, shows that the rainfall data is random as it gives rise to lag plots with no pattern. The points in the lag plot appear scattered from left to right and top to bottom thus there is no significant autocorrelation.

## Model Identification

The best model is the one with the lowest value of AIC. The best model for rainfall is SARIMA (1, 1, 1) (0, 1, 2)$_{12}$ with an AIC of 9.99356. The (1, 1, 1) (0, 1, 2)$_{12}$ describes a model that includes 1 non-seasonal autoregressive parameters 1 non-seasonal moving average parameter and 1 non-seasonal difference. It also indicates 1 seasonal moving average parameter and one seasonal difference. These parameters were computed for the series after it was differenced once with lag 1. The seasonal lag used for the seasonal parameters is usually determined during the identification phase and must be explicitly specified (Box, 2015).

**Table 4**: AIC values of selected models

| Model | DF | AIC | AICc | BIC |
|---|---|---|---|---|
| SARIMA (0,0,1),(0,0,1)$_{12}$ | 237 | 10.3919 | 10.4010 | 9.435496 |
| SARIMA (1,0,0),(0,0,1)$_{12}$ | 237 | 10.4163 | 10.4253 | 9.459755 |
| SARIMA(0,0,0),(0,0,1) $_{12}$ | 238 | 10.4579 | 10.46667 | 9.486917 |
| SARIMA (1,0,0),(0,0,2)$_{12}$ | 236 | 10.4007 | 10.41012 | 9.458732 |
| SARIMA (2,1,1),(1,0,1)$_{12}$ | 233 | 10.0306 | 10.04093 | 9.117599 |
| SARIMA (1,1,1),(0,1,1)$_{12}$ | 224 | 9.99594 | 10.00498 | 9.039445 |
| SARIMA (1,1,1),(0,1,1)$_{12}$ | 222 | 10.0284 | 10.0382 | 9.100876 |
| SARIMA (1,1,1),(1,1,2)$_{12}$ | 223 | 9.99670 | 10.0061 | 9.054712 |
| SARIMA (1,1,1),(1,1,2)$_{12}$ | 224 | 10.4947 | 10.50381 | 9.538272 |
| SARIMA (1,1,1),(0,1,2)$_{12}$ | 223 | 9.99356 | 10.00296 | 9.051574 |

**Comment [U6]:** What is the justification for using AIC for selecting adequate models among other criteria? Many authors indicated that when AIC and BIC select different models, BIC decision must be superior. See Okunlola and Folorunso (2015) page 104 titled Modelling Rainfall Series in Geopolitical Zones of Nigeria

## Parameter Estimation: Maximum likelihood estimates of the SARIMA (1, 1, 1) (0, 1, 2)$_{12}$.

**Table 5**: MLE Estimates of SARIMA (1, 1, 1), (0, 1, 2)$_{12}$

| | Estimate | STD Error | T-value | P-value |
|---|---|---|---|---|
| Intercept | 0.0049 | 0.0026 | 1.9094 | 0.0574 |
| AR1 | 0.1452 | 0.0673 | 2.1580 | 0.0320 |
| MA1 | -1.0000 | 0.0033 | -30.0729 | 0.0000 |
| SMA1 | -1.0671 | 0.1364 | -7.8231 | 0.0000 |
| SMA2 | 0.0671 | 0.0824 | 0.8150 | 0.4159 |

sigma^2 estimated as 7787, log likelihood = -1362.28, AIC = 2734.56

SARIMA (1, 1, 1) (0, 1, 2)$_{12}$ has an estimated variance of 7787 with a log likelihood of -1362.28 and AIC is 2734.56. The parameters were estimated according to the maximum likelihood technique; thus, results are in agreement with (Box, 2015) literature on time series modelling. Parameters are estimated through methods such as method of moments and maximum likelihood (Box, 2015). The MLE was used in this case to find the order of p, d, and q. In SARIMA (1, 1, 1) (0, 1, 2) $_{12}$, AR=1, MA=1 in a first differenced series. SAR=0, SMA=2, and Seasonal difference =1. The rule of parsimony requires a researcher to select the

UNDER PEER REVIEW

simplest model, which adequately explains the behaviour of the values, as explained by Chen *et al*. (2017). The SARIMA parameter estimates indicate that the ARIMA models with seasonal components are best fit among different models to forecast rainfall. Seasonality usually causes the series to be non-stationary because the average values at a particular time within the seasonal span are different from average values at other times.

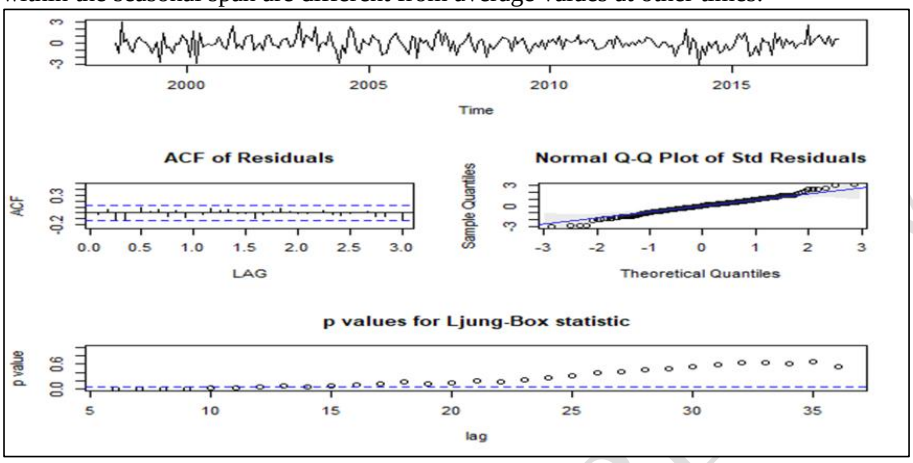


Figure 5: Diagnostic Analysis: Residuals for SARIMA (1, 1, 1), $(0, 1, 2)_{12}$ model for the Embu county Rainfall data.

## Model Validation

To test the adequacy and predictive ability of the chosen models, the actual data sets, predicted values, lower and upper limits are plotted and displayed. The predictive power of SARIMA $(1, 1, 1) \times (0, 1, 2)_{12}$ is very appreciable since it fits well to the test data since all points lie within the confidence interval. The forecasted figures tend to be very close to the actual points. The model predicts well. The prediction is indicated in red line and dots compared to the test dataset of January 2018 to December 2020 in the region. The 95% confidence interval is overlaid in the grey area.

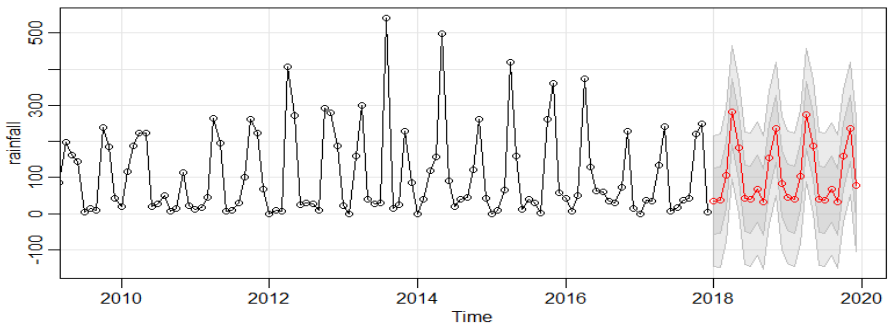


Figure 6: Prediction of SARIMA $(1, 1, 1) \times (0, 1, 2)_{12}$.

## Forecasting

Once the better model was selected, two-year a head prediction was conducted. For this purpose, SARIMA for $(1, 1, 1) \times (0, 1, 2)_{12}$ function of *astsa* package was used. This function produced predicted values based on the chosen ARMA model. Figure 6 shows the resulting prediction plot along with one and two standard error prediction bounds. The mean absolute percentage error (MAPE) of the forecast was 9.0% (Table 6). Hence the model can be considered as a better predictor.

Table 6: Forecasting Accuracy Statistic (Mean absolute percentage error(MAPE))

| Model | MAPE |
|---|---|
| SARIMA $(1,1,1),(0,1,2)_{12}$ | 9.0% |

The MAPE gives a very low value of 9.0%, indicating that the SARIMA $(1, 1, 1)$ $(0, 1, 2)_{12}$ adequately fits the monthly data for Embu County. Thus, the model can be used for rainfall prediction in the region with reliable accuracy. This is in agreement with the study by Chatfield (2016) that a lower value of MAPE provides a good measure of the accuracy of the model and its predictability of the response. For prediction models, MAPE is an important criterion for determining the fit. Forecasting helps in planning and decision-making process since it gives an insight into the future uncertainty using the past and current behaviour of given observations. From most research studies, the selected model is not always the best for forecasting. Further accuracy test by the MAPE must, therefore, be carried out on the model. Lower values of MAPE indicate better fit (Chatfield, 2016). MAPE is a good measure of how accurately the model predicts the response, and it is the most important criterion for fit if the main purpose of the model is the prediction.

Table 7: A sample of Rainfall Forecasts for the SARIMA (1, 1, 1) (0, 1, 2)12

| Year | Point Forecast | Low 99.5% CI | High 99.5% CI | Month | Year | Point Forecast | Low 99.5% CI | High 99.5% CI |
|---|---|---|---|---|---|---|---|---|
| 2018 | 34.66828 | -220.894 | 290.2309 | Jan | 2020 | 44.43287 | -214.687 | 303.5529 |
| 2018 | 36.60022 | -222.093 | 295.2931 | Feb | 2020 | 39.30592 | -219.846 | 298.4578 |
| 2018 | 106.7261 | -152.099 | 365.5508 | Mar | 2020 | 103.4251 | -155.731 | 362.5809 |
| 2018 | 281.1887 | 22.35154 | 540.0258 | Apr | 2020 | 273.743 | 14.58656 | 532.8994 |
| 2018 | 182.8405 | -75.9983 | 441.6793 | May | 2020 | 187.5799 | -71.5766 | 446.7364 |
| 2018 | 42.10109 | -216.738 | 300.9401 | Jun | 2020 | 41.18722 | -217.969 | 300.3437 |
| 2018 | 38.64529 | -220.194 | 297.4843 | Jul | 2020 | 38.71256 | -220.444 | 297.8691 |
| 2018 | 68.91899 | -189.92 | 327.7581 | Aug | 2020 | 68.00157 | -191.155 | 327.1581 |
| 2018 | 31.51366 | -227.326 | 290.3528 | Sep | 2020 | 33.41306 | -225.744 | 292.5697 |
| 2018 | 155.5851 | -103.255 | 414.4248 | Oct | 2020 | 160.5562 | -98.601 | 419.7134 |
| 2018 | 234.8791 | -23.9642 | 493.7224 | Nov | 2020 | 237.0093 | -22.1516 | 496.1703 |
| 2018 | 83.59515 | -175.268 | 342.4581 | Dec | 2020 | 79.55008 | -179.632 | 338.7325 |
| 2019 | 44.5844 | -214.152 | 303.3205 | Jan | 2021 | 45.04701 | -214.538 | 304.6316 |
| 2019 | 38.80295 | -219.924 | 297.5295 | Feb | 2021 | 39.92006 | -219.698 | 299.5385 |
| 2019 | 102.8271 | -155.898 | 361.552 | Mar | 2021 | 104.0392 | -155.584 | 363.662 |
| 2019 | 273.1312 | 14.40642 | 531.8559 | Apr | 2021 | 274.3571 | 14.73375 | 533.9805 |
| 2019 | 186.9661 | -71.7586 | 445.6909 | May | 2021 | 188.1941 | -71.4294 | 447.8175 |
| 2019 | 40.57312 | -218.152 | 299.2979 | Jun | 2021 | 41.80136 | -217.822 | 301.4248 |

UNDER PEER REVIEW

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 2019 | 38.09843 | -220.626 | 296.8232 | Jul | 2021 | 39.3267 | -220.297 | 298.9502 |
| 2019 | 67.38743 | -191.337 | 326.1122 | Aug | 2021 | 68.61571 | -191.008 | 328.2392 |
| 2019 | 32.79892 | -225.926 | 291.5237 | Sep | 2021 | 34.0272 | -225.596 | 293.6508 |
| 2019 | 159.9421 | -98.7833 | 418.6674 | Oct | 2021 | 161.1704 | -98.4538 | 420.7945 |
| 2019 | 236.3952 | -22.3336 | 495.124 | Nov | 2021 | 237.6235 | -22.0048 | 497.2517 |
| 2019 | 78.93594 | -179.812 | 337.6842 | Dec | 2021 | 80.16422 | -179.488 | 339.8161 |

## 4 Conclusion

In this study, the monthly time series rainfall data of Embu, County in Kenya was investigated. A logical procedure was followed in the search for a better stochastic model that could better explain the interesting features contained in the annual series. Among ten statistically competent SARIMA models, a first order seasonal differenced SARIMA (1, 1, 1) $(0, 1, 2)_{12}$.model was found suitable for fitting rainfall data in Embu, County. Furthermore, the model can be used as a potential alternative for the prediction of annual rainfall values. Finally, as a recommendation, other stochastic models should be investigated to see if other models can also preserve long term statistical behaviour of annual rainfall in Embu, County. Besides, seasonal behaviour of the town's monthly rainfall should also be explored.

## REFERENCES


[1]. Adede, S. A. (2018). Application of Time Series Analysis to Annual Rainfall Values in Debre Markos Town, Ethiopia. *Computational Water, Energy, and Environmental Engineering*, *7*(03), 81.

[2]. Afrifa-Yamoah, E., Saeed, B.I., & Karim, A. (2016). Sarima Modelling and forecasting of monthly rainfall in the Brong Ahafo Region of Ghana. *World Environment*, *6*(1), 1-9.

[3]. Ampaw, E. M., Akuffo, B., Larbi, S. O., & Lartey, S. (2013). Time Series Modelling of Rainfall in New Juaben Municipality of the Eastern Region of Ghana. *International Journal of Business and Social Science*, *4*(8).

[4]. Burns, P. (2002). Robustness of the Ljung-Box test and its rank equivalent. Available at SSRN 443560.

[5]. Embu County (2016). The County Platform. From http://www.thecountyplatform. or.ke/embu-county/#ffs-tabbed-12

[6]. Embu County Government (2017). The land of opportunities. From https://www.embu.go.ke/.

[7]. Chatfield, C. (2016). *The Analysis of Time Series: An Introduction*. CRC Press.

[8]. Chen, J.M., Hawkes, A.G., Scalas, E. & Trinh, M., (2017). Performance of Information Criteria used for Model Selection of Hawkes Process Models of Financial Data. *ArXiv Preprint arXiv*:1702.06055.

[9]. Cryer, J. D., & Chan, K. S. (2008). Time series regression models. *Time series analysis: with applications in R*, 249-276.

[10]. GoK (2004). Draft national policy for the sustainable development of arid and semi-arid lands of Kenya. Government printers, Nairobi

[11]. Huho, J. M., & Mugalavai, E. M. (2010). The effects of droughts on food security in Kenya. International *Journal of Climate Change: Impacts and Responses*, 2(2), 61-72.

[12]. Kane, Ibrahim Lawal, and Fadhilah Yusof (2012), Modelling monthly rainfall time series using ETS and sarima, *International Journal of Current Research* 4(1), 195–200.


UNDER PEER REVIEW

[13]. Kenya Ministry of Lands & Physical Planning (2017). http://www.ardhi.go.ke/
[14]. Khan, M. Z. K., Sharma, A., Mehrotra, R., Schepen, A., & Wang, Q. J. (2015). Does improved SSTA prediction ensure better seasonal rainfall forecasts? *Water Resources Research*, 51(5), 3370-3383.
[15]. Kibunja, H. W., Kihoro, J. M., Orwa, G. O., & Yodah, W. O. (2014). Forecasting precipitation using SARIMA Model: A Case study of Mt. Kenya Region. *Mathematical Theory and Modeling*. 11, (4), 50-58.
[16]. Kisaka, M. O., Mucheru-Muna, M., Ngetich, F. K., Mugwe, J. N., Mugendi, D., & Mairura, F. (2015). Rainfall variability, drought characterization, and efficacy of rainfall data reconstruction: case of Eastern Kenya. Advances in Meteorology, 2015.
[17]. Kisaka, M., Mucheru-Muna, M., Ngetich, F., Mugwe, J., Mugendi, D., & Mairura, F. (2014). Research Article: rainfall variability, drought characterization, and efficacy of rainfall data reconstruction: Case of Eastern Kenya.
[18]. Kowal, D. (2015). A Modified Ljung-Box Test for the Functional Linear Model.
[19]. Mahmud, Ishtiak; Sheikh Hefzul Bari, and M. Rahman (2016). "Monthly rainfall forecast of Bangladesh using autoregressive integrated moving average method." *Environmental Engineering Research* (22)2, 162-168.
[20]. Metrine, C., Kennedy, N., Omukoba, M., Lucy, M., &Frankline, T. (2015). A Time series model of rainfall pattern of Uasin Gishu County.
[21]. Mohamed, T. M. & Ibrahim, A. (2016). Time Series Analysis of Nyala Rainfall Using ARIMA Method.
[22]. Padhan, P.C. (2012). Application of ARIMA Model for Forecasting Agricultural Productivity in India. *Journal of Agriculture &Social Sciences*, 8: 50-56.
[23]. Papalaskaris, Thomas, (2016). Theologos Panagiotidis, and Athanasios Pantrakis. "Stochastic monthly rainfall time series analysis, modeling and forecasting in Kavala City, Greece, North-Eastern Mediterranean Basin." *Procedia engineering*. *162* 254-263.
[24]. Roudier, P., Ducharne, A., & Feyen, L. (2014). Climate change impacts on runoff in West Africa: a review. *Hydrology and Earth System Sciences*, *18*, 2789-2801.
[25]. Salauddin Md. Khan, Masudul Islam, Sajal Adhikary, Md. Murad Hossain and Sohani Afroja. (2014). Analysis and Predictions of Seasonal Affected Weather Variables of Bangladesh: SARIMA Models vs. Traditional Models. *International Journal of Business and Management*; Vol. 13, No. 12; 2018.
[26]. Sopipan, N. (2014), Forecasting rainfall in Thailand: A case study of nakhon ratchasima province. *International Journal of Environmental, Chemical, Ecological, Geological and Geographical Engineering* 8(11).
[27]. Tariq, M. M., and A. I. Abbasabd. "Time series analysis of Nyala rainfall using ARIMA method." *SUST Journal of Engineering and Computer Science* 17, no. 1 (2016): 5-11.
[28]. Valipour, M. (2015). Long-term runoff study using SARIMA and ARIMA models in the United States. *Meteorological Applications*, *22*(3), 592-598.
[29]. Wakachala, F. M., Shilenje, Z. W., Nguyo, J., Shaka, S., & Apondo, W. (2015). Statistical patterns of rainfall variability in the Great Rift Valley of Kenya. *J Environ Agric Sci*, 5, 17-26.
[30]. Wang, Q. J., Schepen, A., & Robertson, D. E. (2012). Merging seasonal rainfall forecasts from multiple statistical models through Bayesian model averaging. *Journal of Climate*, 25(16), 5524-5537.

UNDER PEER REVIEW