*IOSR Journal of Mathematics (IOSR-JM)*
*e-ISSN: 2278-5728, p-ISSN: 2319-765X. Volume 18, Issue 3 Ser. II (May. – June. 2022), PP 13-23*
*www.iosrjournals.org*

# Application of Bayesian Vector Autoregressive Model In Forecasting Rainfall In Kenya.

Mr. Harun Mwangi Gitonga [1], Prof. Joseph Koske[2], Dr. Mathew Kosgei[3]

1. *Mr. Harun Mwangi Gitonga: email, gitongamwangih@gmail.com, Affiliation: Moi University (Department of Mathematics, Physics and Computing in the School of Sciences and Aerospace Studies)*
2. *Prof. Joseph Koske: email, koske4@yahoo.co.uk, Affiliation: Moi University (Department of Mathematics, Physics and Computing in the School of Sciences and Aerospace Studies)*
3. *Dr. Mathew Kosgei: email, mkosgei@mu.ac.ke, Affiliation: Moi University (Department of Mathematics, Physics and Computing in the School of Sciences and Aerospace Studies)*

***Abstract***
*Forecasting for a future has turned to be of practical importance in the world in many areas. Time series modeling has grown fundamental importance in forecasting to evaluate the said future events. Many important models have been proposed to improve the accuracy of the future prediction. Weather pattern have been changing due to global warming, which has brought a big challenge to the world in affecting economic and noneconomic activities. Global warming causes severe weather changes, which are characterized by precipitation and temperature. Rainfall prediction is one of the most important and challenging tasks in today's world. The objective of this study was to develop a Bayesian vector autoregressive model to forecast the rainfall pattern in Kenya. The Bayesian vector autoregressive model was developed after a diagnostic analysis of the rainfall data. The developed model was used for forecasting ten future points were obtained for each zone. The Ljung-Box test of residuals shows that the graphs of the naïve method produced forecasts that appear to account for all available information. The mean of the residuals was close to zero and there was no significant correlation in the residual series. The time plot of the residuals shows that the variation of the residuals stays much the same across the historical data. The histogram shows that the residuals were normally distributed, which represented Gaussian behavior. The ACF graph shows that the spikes were within the required limits, so the conclusion was that the residuals had no autocorrelation among them. The Ljung-Box test shows that the developed model was good for forecasting. The developed model was used to give the prediction error for each value. The Bayesian Vector Autoregressive (BVAR) model gave an accuracy level of 88.73%. Finally, the researcher recommends the application of other techniques like random forest and bootstrapping technique to check whether the accuracy may further be improved by other models.*

***Keywords***
***Global Warming, Bayesian Vector Autoregressive. Forecasting***

---

Date of Submission: 07-06-2022 Date of Acceptance: 22-06-2022

---

## I. Introduction

The world is currently generating large datasets in virtually all fields. The amount of data produced and recorded has grown enormously in many fields which include; weather recording, biomedical, social network, mobile network data, digital archives, and electronic trading, among others. This unanticipated amount of data provides unprecedented opportunities for data-driven decision making and knowledge discovery. However, the massive sample size and high dimensionality of Big Data introduces unique computational and statistical challenges, including scalability and storage bottle neck, noise accumulation, spurious correlation, incidental endogeneity and measurement errors. The task of analyzing such large-scale data sets poses significant challenges and calls for innovative statistical methods specifically designed for faster speed and higher efficiency and accuracy. These challenges are eminent and require a new computational and statistical paradigm shift. In spite of the explosion of this big data, specific tools are required for modelling, mining, visualizing and predicting to understanding these large data sets. In many situations, it is easy to predict the outcome given the cause. However, in science, more often than not, we are faced with the question; when given the outcome of an experiment, what are the causes or the probability of the causes compared to other outcomes? Bayesian theory provides a framework for plausible reasoning and a concept which is a more powerful and general tool for handling this problem. To apply Bayesian, it is required to partition the data into the training and the testing sets, where training set is used to develop a model and testing set is for testing the developed model. This idea of Bayesian theory was championed by Jaynes (2003). There has been a growing interest in applying big data to

many analytical areas, particularly in time series prediction. The primary model in Multivariate time series analysis is the Vector Autoregressive (VAR). This study helps to integrate interdependent variables to develop a computational efficient model for VAR prediction using Bayesian model. One of the actively researched areas is the weather distribution pattern, about which the understanding is still in its early stages of inference. Numerous studies have been conducted to further the knowledge; but Bayesian methodology finds its place to aid in obtaining scientific inferences about certain facts from available data. This study provides an account of VAR and Bayesian model data analysis applied to weather distribution with a particular focus on rainfall distribution patterns in Kenya. The GVAR model is obtained by integrating the regional model through inter-linkages using different weather variables that allows for interdependencies. Weather forecasting is the application of science and technology in predicting the state of the atmosphere for a future time at a given location. It is carried out by collecting quantitative data about the current and past state of the weather conditions. This study used VAR model which was a tool for forecasting. The amount of rainfall in a given region is affected by several factors which include; temperature, atmospheric pressure, wind speed, relative humidity, radiation, and altitude, among others.

Much of the discussion around climate change focuses on how much the earth would warm up over the coming century. Climate change is not only limited to temperature, but also how precipitation (both rain and snow) changes would also have a great impact on the global population. This study considered a number of variables they included; Rainfall, which was the response variable, and the explanatory variables which were Temperature, Humidity, Atmospheric Pressure, Wind Speed, Radiation, and Wind Gust. The main purpose of this study was to get more insight about the rainfall patterns in Kenya. Several predictor variables were used in this study which were noted to influence rainfall patterns in Kenya. The effects of global warming have greatly affected rainfall patterns in Kenya, which have caused adverse economic and social effects.

Bayesian Vector Autoregressive (BVAR) is used to conduct both classical unconditional as well as conditional forecasts. Unconditional forecasts rival those obtained from factor models in accuracy Giannone *et al. (*2015) and are used for a variety of analyses. Conditional forecasts allow for elaborate scenario analyses, where the future path of one or more variables is assumed to be known. They are handy tools for analyzing possible realizations of policy-relevant variables.

**Purpose of the study**

Rainfall is one of the most affected weather components which is virtually influenced by other weather components. However, how do we deal with such a problem of variable interaction? The interaction and interdependence of variables that constitute response variables are found in many areas. VAR method are used to handle the variable inter-linkage which results from the big data. It represents the correlations among a set of variables, which are used to analyze certain aspects of the relationship between the variables of interest. The study also used Global VAR which played a key role. The main idea behind the global VAR framework is to incorporate inter-linkages between cross-sectional weather zones in a viable way. The integration of regional VAR models was used to formulate Global VAR. The next part was to ensure that improved accurate prediction was achieved. This was done through the introduction of Bayesian approaches. Bayesian theory provided a framework for plausible reasoning, a concept which was more powerful and general, an idea championed by Jayes(2003). Bayesian analysis was a useful technique which used to detect the rainfall patterns and changes using the past data to give the present information about what would happen in future. Thus, a need to develop a more accurate prediction model is necessary to overcome this global challenge. Most of the methods employed were probabilistic models, they were having a challenge of clearly identifying the part of the weather signal that was due to change, making it complex to unravel. Consequently, the probabilistic models had a weakness that they could not predict accurately. One way to overcome this weakness is through the use of Bayesian Models. The study tackled the problem of predicting adverse rainfall patterns by applying BVAR to the big data. Since an accurate forecast of rainfall patterns would save lives, support emergency management teams, mitigate the impacts of damages, and prevent economic losses, hence the important of this study. The central idea of the Bayesian method is the use of study data to update the state of knowledge about the quantity of interest has been studied. This idea in the Bayesian approach is a very intuitive one, namely, that of updating knowledge. The state of knowledge about the quantities of interest before or prior to a study is updated by the current study data, which yielded the state of knowledge after or posterior to the study. The transformation from prior to posterior is achieved by Bayes Theorem, an explicit mathematical expression for the updating process. Since from the review of the previous literatures no conclusion has been made on a good weather predictor, this motivated this study to explore the idea of the prior-to-posterior transformation by considering the rainfall data set in Kenya and using BVAR model. Models for accurate prediction of weather changes in Kenya are identified as a major area of concern that this study sought to address. This paper aims at conducting data variable analyzes to develop a predictive model of rainfall patterns using Bayesian Vector Autoregressive.

## II. Literature Review

Kenya has experienced prolonged droughts and intense flooding every year *Mary Kilavi* at el. (2018). With an increase in such extreme weather events, the glaciers around Mount Kenya have disappeared, leading to the drying up of rivers and streams. Weather changes have also led to harvest losses and food shortages, as well as landslides, soil degradation, and a loss of biodiversity *Otiende and Brian (2009).* The diminishing water sources and erratic rainfall have reduced the availability of water. However, meteorological phenomena like rainfall, normally vary more on local scales. Linacre and *Geerts* (1997) state that Numerical Weather Prediction (NWP) is a simplified set of equations called the primitive equation used to calculate the changes of conditions. According to *Lutgens and TarBuck* (1989), the word "numerical" is misleading since all types of weather forecasting are based on some quantitative data and therefore could fit under this area. The large number of variables that are included when considering the dynamic atmosphere makes this task extremely difficult. Manipulating the large data sets and performing the complex calculations necessary to predict weather and make a resolution conclusive enough to make the result useful require the use of some of the most powerful computers. In the last forty years, facilitated by advances in observing systems and improvements in the understanding and modelling of the various components of the Earth system and supported by enhancements in computing capabilities, steady advances in weather and climate prediction have occurred at major operational centers across the world, *Bauer et al.,* (2015). Complementing these advances in weather and climate prediction, there have been important milestones in advancing the science and operational infrastructure for prediction at longer timescales. The first generations of dynamic seasonal forecast systems were implemented at operational centers in the mid-1990s, *Stockdale et al*., (1998). Routine weather and climate forecasts at the global and regional levels now provide information critical for the economic welfare of society and for mitigating losses of life and property. According to the State of the Climate in 2017, *Blunden, Ji. et al*., (2018), since 1901, the mean annual global (land + ocean) surface air temperature had warmed by 0.7–0.9° Celsius per century, and the rate of warming had nearly doubled since 1975 to 1.5–1.8° Celsius per century. A steady rise in temperature has triggered important changes in the frequency and intensity of extreme weather and climate events such as heat and cold waves, droughts, floods, hurricanes, and so forth over various parts of the globe Intergovernmental Panel on Climate Change, (2013). These unprecedented long-term climatic changes have influenced sub seasonal and seasonal-to-interannual variability and had a profound impact on the natural environment as well as on the life, health and well-being of human society, *Coumou and Rahmstorf* (2012).

Bayesian Vector Autoregressive (BVAR) is used to conduct both classical unconditional as well as conditional forecasts. Unconditional forecasts rival those obtained from factor models in accuracy *Giannone et al. (*2015) and are used for a variety of analyses. Conditional forecasts allow for elaborate scenario analyses, where the future path of one or more variables is assumed to be known. They are a handy tool for analyzing possible realizations of policy-relevant variables. Impulse Response Functions (IRF) are a central tool for structural analysis. They provided insights into the behavior of weather systems and are another cornerstone of inference in VAR models. IRFS served as a representation of shocks hitting the system and are used to analyze the reactions of individual variables. The exact propagation of these shocks is of great interest, but meaningful interpretation relies on proper identification. BVAR features a framework for identification schemes, with two of the most popular schemes currently available; namely short-term zero restriction and sign restriction. The former is also known as recursive identification and is achieved via Cholesky decomposition of the Variance Covariance Vector (VCOV) matrix by *Kilian and Lutkepohl* (2017). Additionally, identification via sign restrictions comes at the cost of increased uncertainty and a loss of precision for the resulting IRF. Another related tool for structural analysis is Forecast Error Variance Decomposition (FEVD).

When BVAR models are conducted, Granger Causality tests are required to check if there is a significant association between variables. *Lütkepohl* (2005) states that there is Granger Causality, if information from one endogenous time series gives the most accurate prediction of another endogenous time series even though all other possible information is taken into account. Subsequently, *Lütkepohl* (2005) meant that the idea behind the Granger Causality test is that the effect is generated by the cause, and not the reverse. However, it is important to note that the test also identifies the direction of the association between variables and not only causality.

## III. Methodology

The study used secondary data, which was sourced from Trans- African Hydro-Meteorological Observatory (TAHMO) and Kenya Meteorological stations. The data captured over a period of four years from 2014 June to June 2017. The data was collected in Kenya a cross five regions, namely; Coastal, Arid, Semi-arid, Highlands, and Lake regions. The data was in the form of daily recordings for at least five evenly distributed weather stations in each of their respective regions. The study considered the data for seven variables, which included; Rainfall, Temperature, Atmospheric pressure, Wind speed, Wind gust, Radiation and Relative humidity. The data was converted into CSV files to import it into R statistical software for analysis. To

remove scaling, normalization was done through linear scaling technique. It was essential because all variables used different units of measurement. Moreover, a variable may have a large impact on the predictor variable only because of its numerical scale. The technique of linear scaling, which is also referred to as min-max normalization estimation has a formula stated as;

$$x = \frac{\text{x} - \text{Min(x)}}{\text{Max(x)} - \text{Min(x)}}$$

Normalization transformed the data into a common range of between 0 and 1. Thus, removing the scaling effects from all variables.

Let $x_t$ be an n x 1 random vector that takes values in the domain of real numbers. The evolution of $x_t$ the endogenous variable is described by a system of p-th order difference equations in the VAR(p):

$$x_t = \alpha + B_1 x_{t-1} + \cdots + B_p x_{t-p} + e_t$$

The vector of stochastic innovation, $e_t$, an independent and identically distributed random variable for each t. the distribution from which $e_t$ is drawn which determined the distribution of $x_t$, conditional on its past

$x_{1-p:t} = \{x_{t-p}, \dots\dots, x_0, \dots\dots\dots x_{t-2}, x_t\}$.

The standard assumption is that the errors are Gaussian.

$e_t \sim iid.\, N(0, \Sigma)$.

This implies that the conditional distribution of $x_t$ is also Normal. Bayesian inference on the $x_t$ model amount to updating prior beliefs about the VAR parameters, that are seen as stochastic variables, after having observed a sample

$x_{1-p:t} = \{x_{t-p}, \dots\dots, x_0, \dots\dots\dots x_{t-2}, x_t\}$.

Prior beliefs about the VAR coefficients are summarized by a probability density function, and updated using Bayes' Law.

$$\text{p}\big(\text{A}, \Sigma/\text{x}_{1-\text{p:t}}\big) = \frac{\text{p(A,}\Sigma\text{)p}\big(\text{x}_{1-\text{p:t}}/\text{A,}\Sigma\big)}{\text{p}\big(\text{x}_{1-\text{p:t}}\big)} \propto \text{p}\big(\text{x}_{1-\text{p:t}}/\text{A}, \Sigma\big)$$

Define A = $[A_1 \dots\dots\dots A_{P,}]$' as a k x n matrix, with k = np +1. The joint posterior distribution of the VAR(p) coefficients p(A, Σ) summarizes the initial information about the model parameters, and the sample information is the likelihood function. The posterior distribution summarizes the entire information available and is used to conduct inference on the VAR parameters. Under the assumption of Gaussian errors, the conditional likelihood of VAR is

$$p\big(x_{1-T}/A, \Sigma, x_{1-p:0}\big) = \prod_{t=1}^{T} \frac{1}{(2\pi)^{1/2}} |\Sigma|^{-1} exp\,\{-\frac{1}{2}(x_t - A'x'_t)\,\Sigma^{-1}(x_t - A'x_t)\}$$

Where

$$x'_t = [x'_{t-1} \dots\dots\dots\dots\dots\dots\ x'_{t-p}]$$

The likelihood in this equation is written in compact form, by using the apparently unrelated regression representation of the VAR.

$$x_t = Ax + e_t$$

Using this notation and standard properties of the trace operator, the conditional likelihood function is equivalently expressed as

$p\big(x_{1-T}/A, \Sigma, x_{1-p:0}\big) = \frac{1}{(2\pi)^{1/2}} |\Sigma|^{-1} \exp\{-\frac{1}{2} tr\big[\Sigma^{-1}\widehat{S}\big]\}\, X\, exp\{-\frac{1}{2} tr[\Sigma^{-1}(A - \hat{A})'x'x(A - \hat{A})\}$

Where $\hat{A}$ is the maximum likelihood estimator (MLE) of A, and $\hat{S}$ the matrix of sums of squared residuals that is $\hat{A} = (x'x^{-1})\, x'x_t, \quad \hat{S} = \big(x_t - x\hat{A}\big)'(x_t - x\hat{A})$

The likelihood is written in terms of the vectorized representation of the VAR

$x_t = (I_n \otimes x) \propto + e,\ e \sim (0, \Sigma \otimes I_T)$

Where $x_t = vec(x_t)$ and e = $vec$(e) are Tn x 1 vectors, and $\propto = vec(A)$ is nk x1. In this vectorized notation, the likelihood function is written as

$$p\big(x_{1:T}/A, \Sigma, x_{1-p:0}\big) = \frac{1}{(2\pi)^{Tn/2}} |\Sigma|^{-T/1} \exp\{-\frac{1}{2} tr\big[\Sigma^{-1}\widehat{S}\big]\} X\, exp\{-\frac{1}{2}(\propto - \hat{\propto})'\Sigma^{-1} \otimes (x'x)(\propto - \hat{\propto})\}$$

Where, consistently, $\hat{\propto} = vec(\hat{A})$ is nk x 1. The likelihood function is used to update the prior information regarding the VAR parameters.

This ability in Bayesian assists in the predictability of the future or the current situation. If time series observations are available for a variable of interest and the data from the past contains information about the future development of the available, it is plausible to use a forecast of some function of the data collected in the past. As forecasting is one of the main objectives of multiple time series analysis. Forecast for horizon h ≥ 0 of an empirical VAR(p) process are generated recursively according to Box and Jenkins (2008).

$Y_{T+h/T} = A_1 Y_{1+h-1/T} + \dots\dots\dots + A_P Y_{1+h-P/T}$

$Y_{T+j/T} = Y_{T+j}$ for j <0

where,

$$\mathrm{Cov}\left(\begin{bmatrix} Y_{T+1} - Y_{T+1/T} \\ . \\ . \\ . \\ Y_{T+h} - Y_{T+h/T} \end{bmatrix}\right) = \begin{bmatrix} I\ 0 \dots\dots\dots\dots\dots .0 \\ \Phi_1\ \ I \dots\dots\dots\dots\dots .0 \\ \dots\dots\dots\dots\dots\dots\dots \\ \dots\dots\dots\dots\dots\dots\dots . \\ \dots\dots\dots\dots\dots\dots\dots \\ \Phi_{P-1}\ \Phi_{P-2}\ \dots\dots . I \end{bmatrix} (\Sigma_U \otimes I_h) \begin{bmatrix} I\ 0 \dots\dots\dots\dots\dots .0 \\ \Phi_1\ \ I \dots\dots\dots\dots\dots .0 \\ \dots\dots\dots\dots . . \\ \dots\dots\dots . . \dots \\ \dots\dots\dots\dots\dots \\ \Phi_{P-1}\ \Phi_{P-2}\ \dots\dots . I \end{bmatrix}$$

The matrices $\Phi_i$ are the empirical coefficient matrices of the Wald moving average representation of a stable VAR(p) - process and the operator $\otimes$ is the Kronecker product.

## IV. Findings

To ensure that the time series data contains no flaws, is stable, and it is not affected by serial correlation, diagnostic analysis was put into use. To achieve these, two tests were conducted to ascertain the applicability of the data in this study. The test included Stationarity test and Granger causality test. The first step was to obtain the time plot graph. When these tests were determined, the BVAR models were developed and forecasting was done and the results presented below.

**Time plot graph for the data**

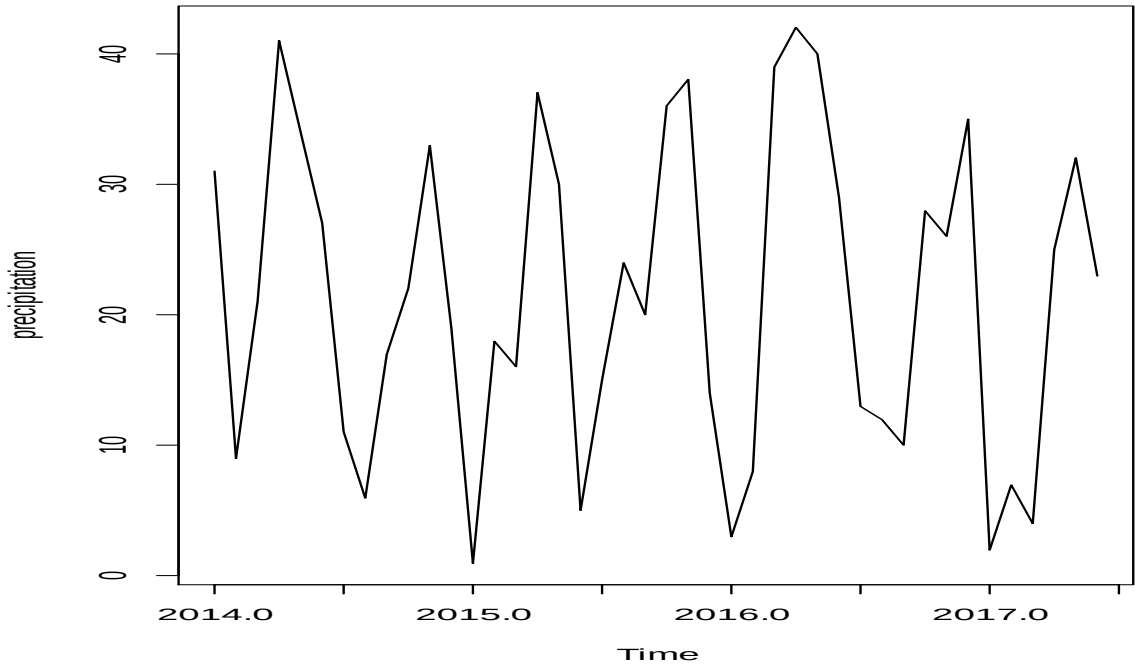


**Figure 1: Time series graph**

**Time series graph for zone one**

The plots exhibit a time series in nature. This graph is a sample representation of all other zones as they exhibited the same behavior. The time plot shows the seasonality behavior and needs to be differenced and tested for stability.

**Stationarity Test**

The test was conducted to establish the stability of the data

**Zone One**

| Variables | ADF Test Statistics | Phillips-Perron | Truncation lag parameter | P-Value ADF | P-Value P.P | Remarks |
|---|---|---|---|---|---|---|
| $X$ | -2.285 | -22.53 | 3 | 0.04631 | 0.0127 | Stationary |
| $X_1$ | -2.6144 | -12.11 | 3 | 0.03372 | 0.0343 | Stationary |
| $X_2$ | -2.129 | -14.12 | 3 | 0.0523 | 0.0218 | Stationary |

| | | | | | | |
|---|---|---|---|---|---|---|
| $X_3$ | -3.832 | -19.58 | 3 | 0.020318 | 0.033 | Stationary |
| $X_4$ | -3.893 | -10.819 | 3 | 0.0274 | 0.04245 | Stationary |
| $X_5$ | -2.312 | -17.09 | 3 | 0.0453 | 0.01643 | Stationary |
| $X_6$ | -2.206 | -9.378 | 3 | 0.04934 | 0.039382 | Stationary |

**Table 1: Zone One Stationarity Test**

From Table 1, it shows that after differencing once all variables were stationary, and they had a unit root. Thus, we make a conclusion that zone one data was stationary.

**Granger Causality Test**

The test was to determine if there was any serial correlation among different variables. It was also used to test if one time series could be used to forecast another time series.

**Zone One**

The hypothesis was that rainfall is not a granger caused by the independent variables.

Hypothesis Testing for Granger causality Zone one

| **X*$x_i$** | **F** | **P** | **Null hypothesis rejected** |
|---|---|---|---|
| $x_1$ | *-9 49254* | *2.03e-05* *** | √ |
| $X_2$ | *-9 21369* | *4.68e-05* *** | √ |
| $X_3$ | *-9 95712* | *1.045e-05* *** | √ |
| $X_4$ | *-9 28929* | *3.457e-05* *** | √ |
| $X_5$ | *-9 164550* | *6.077e-06* *** | √ |
| $X_6$ | *-9 1417* | *0.0007054* *** | √ |

**Table 2: Granger Causality test zone, one**

Table 2 shows that Granger causality existed since the P- value was statistical significant.

**Residual test**

The test was done by use of Ljung-box to determine if the developed model was fit for forecasting. The test used the residuals that were obtained from the model.

**Ljung-box test**

It shows three items; the graph of the residuals, which displays the deviations from the actual values, it also displays the ACF graph, which helps to check for uncorrelation in the residuals. It is the standard residual diagnostic to check if they behave as white noise and therefore the model can be used for forecasting. In this case, the developed model can be used for the intended purposes of forecasting. The last part is the histogram, which is used to check for the Gaussian behavior. The bell shape is well displayed in the histogram, and since a good forecast method should have normally distributed residuals, then the model would give a good forecast.

**Zone One: Residual analysis**

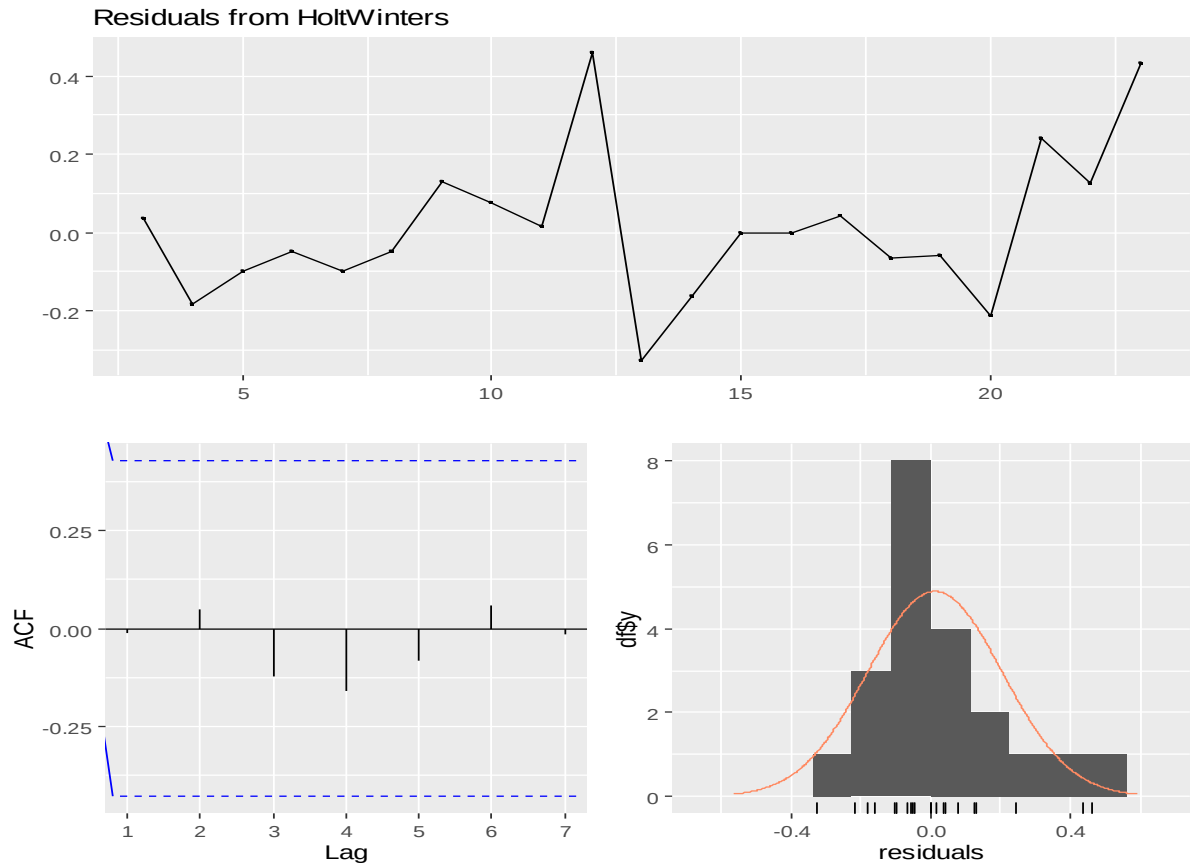


**Figure 2: Ljung-Box test**

These graphs show that the naïve method produces forecasts that appear to account for all available information. The mean of the residuals is close to zero and there is no significant correlation in the residual series. The time plot of the residuals shows that the variation of the residuals stays much the same across the historical data, apart from the two values that are beyond 0.2 or -0.2, and therefore the residual variance can be treated as constant. The histogram shows that the residuals are normally distributed, which represents Gaussian behavior. The ACF graph shows that the spikes are within the required limits, so the conclusion is that the residuals have no autocorrelation with the residuals.

**Ten points future forecasting.**

**Zone one forecasting analysis**

| | fcst | lower | upper | CI |
|---|---|---|---|---|
| 1 | 0.2435105 | -0.3638516944 | 0.8508728 | 0.6073622 |
| 2 | 0.5907031 | -0.0723471584 | 1.2537533 | 0.6630502 |
| 3 | 0.3862335 | -0.3390821988 | 1.1115492 | 0.7253157 |
| 4 | 0.7652912 | 0.0008073628 | 1.5297750 | 0.7644838 |
| 5 | 0.5235901 | -0.2554446203 | 1.3026249 | 0.7790347 |
| 6 | 0.4256933 | -0.3559468631 | 1.2073335 | 0.7816402 |
| 7 | 0.3165178 | -0.4687087960 | 1.1017444 | 0.7852266 |
| 8 | 0.3283320 | -0.4594947612 | 1.1161588 | 0.7878268 |
| 9 | 0.2991094 | -0.4991364994 | 1.0973553 | 0.7982459 |
| 10 | 0.3431256 | -0.4608876550 | 1.1471390 | 0.8040133 |

**Table3: Zone one forecasting analysis**

The table shows the forecasted values and the intervals between the maximum and minimum values where the forecasting values will lie. It also shows the confident interval from the mean forecasted value.

**Zone one model comparison results of actual and predictive**

| | Actual value | Predicted value | Error |
|---|---|---|---|
| 1 | 0.09292385 | 0.09307039 | -0.00014653 |
| 2 | 0.60660266 | 0.55920463 | 0.04739802 |
| 3 | 0.25305800 | 0.18979422 | 0.06326378 |
| 4 | 0.36735499 | 0.31946700 | 0.04788799 |
| 5 | 0.64869131 | 0.52741294 | 0.12127836 |
| 6 | 0.08976720 | 0.06899376 | 0.02077343 |
| 7 | 0.18768907 | 0.50749100 | -0.3198019 |
| 8 | 0.1282763 | 0.1236670 | 0.0046093 |
| 9 | 0.1582632 | 0.1406648 | 0.0175984 |
| 10 | 0.7836330 | 0.7575771 | 0.0260559 |

**Table 4: The Prediction Accuracy Test**

This table above is used to compare the forecasted value from BVAR model and the actual values. It also gives the prediction error for each value. The accuracy level displayed of 88.73% shows that the forecasting is more accurate compared to forecasting using VAR model which gave an accuracy of 78.68%.

**Holtwinters Forecasting Analysis**

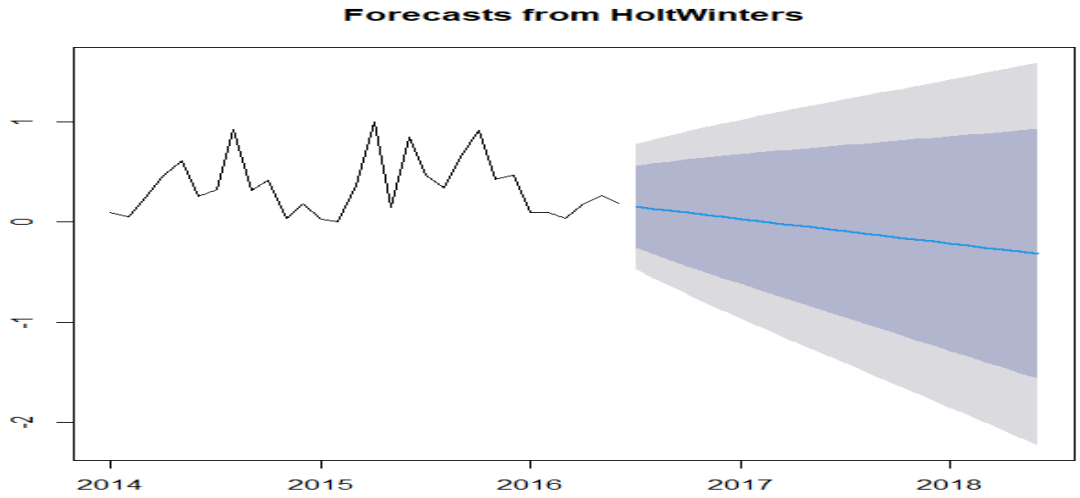


**Figure 3: HoltWinters forecast for zone one**

The figure shows the 80% and 90%, ten months forward forecasting, where the thick blue represents 80% while the light blue is 90%.

**Zone Two**

**Stationarity Test for Zone Two**

| Variables | ADF Test Statistics | Phillips-Perron | Truncation lag parameter | P-Value ADF | P-Value P. P | Remarks |
|---|---|---|---|---|---|---|
| $X$ | -4.347 | -28.68 | 3 | 0.0357 | 0.01 | Stationary |
| $X_1$ | -1.792 | -7.131 | 3 | 0.655 | 0.0268 | Stationary |
| $X_2$ | -3.285 | -26.15 | 3 | 0.04825 | 0.0119 | Stationary |
| $X_3$ | -3.148 | -15.06 | 3 | 0.01212 | 0.0183 | Stationary |
| $X_4$ | -3.531 | -14.94 | 3 | 0.0507 | 0.0190 | Stationary |
| $X_5$ | -3.914 | -27.23 | 3 | 0.0235 | 0.0437 | Stationary |
| $X_6$ | -2.686 | -13.16 | 3 | 0.0303 | 0.0138 | Stationary |

**Table 5:** Zone Two shows that the variables are Stationarity.

**Zone Two**

The hypothesis was that rainfall is not a granger causal by temperature, humidity, wind, wind gusts, atmospheric pressure, and radiation.

**Hypothesis Testing for Granger causality Zone two**

| **X*$x_i$** | **F** | **P** | **Null hypothesis rejected** |
|---|---|---|---|
| $x_1$ | *-6 4.8389* | *0.007123* ** | √ |
| $X_2$ | *-6 5.4861* | *0.004149* ** | √ |
| $X_3$ | *-6 1.0541* | *0.04334* * | √ |
| $X_4$ | *-6 4.9635* | *0.006399* ** | √ |

| | | | |
|---|---|---|---|
| $X_5$ | -6 3.1795 | 0.03492 * | √ |
| $X_6$ | -6 3.1194 | 0.03723 * | √ |

**Table 6: Granger test zone two**

Therefore, the Granger causality test for Zone two shows that all variables were having a strong significant influence on the causes of the dependent variable X. Their level of significant was less than 0.05.

**Ljung-Box test**

The Ljung-box test shows three items; the graph of the residuals, which displays the deviations from the actual values, it also displays the ACF graph, which helps to check for uncorrelation in the residuals. It is the standard residual diagnostic to check if they behave as white noise and therefore the model can be used for forecasting. In this case, the developed model can be used for the intended purposes of forecasting. The last part is the histogram, which is used to check for the gaussian behavior. The bell shape is well displayed in the histogram, and since a good forecast method should have normally distributed residuals, then the model would give a good forecast.

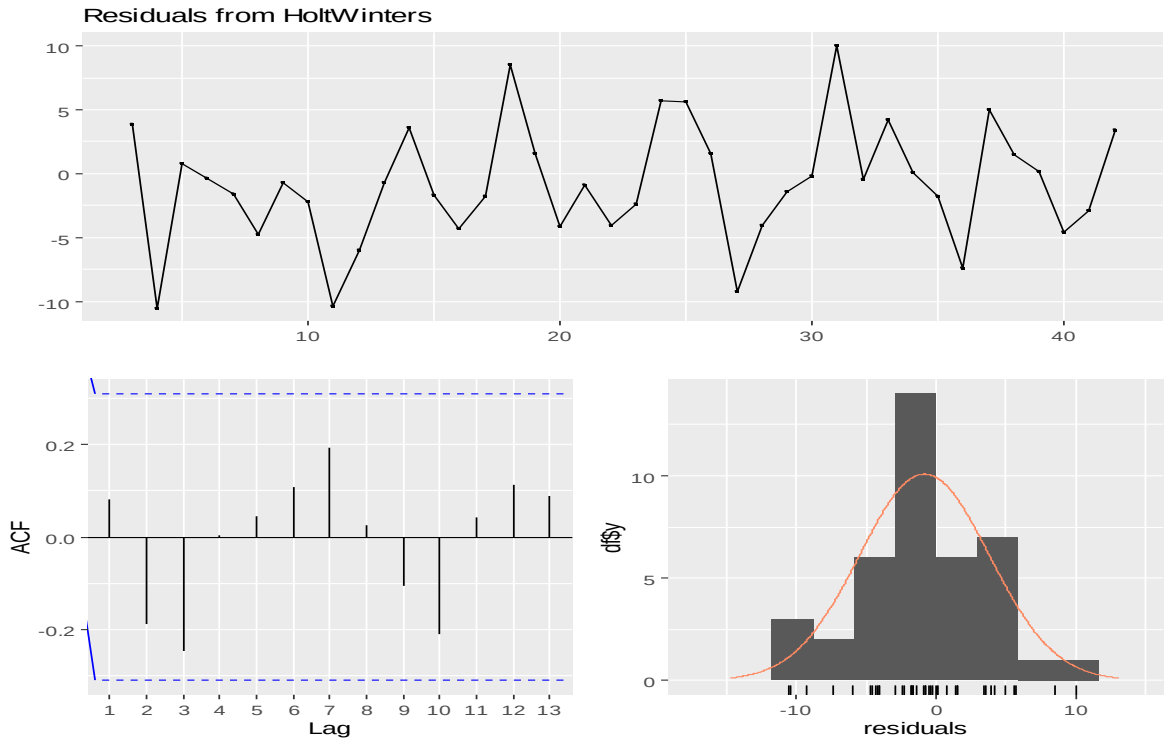


**Figure 4: forecast for zone two**

These graphs show that the naïve method produces forecasts that appear to account for all available information. The mean of the residuals is close to zero and there is no significant correlation in the residual series. The time plot of the residuals shows that the variation of the residuals stays much the same across the historical data, and therefore the residual variance can be treated as constant. This can also be seen on the histogram of the residuals. The histogram suggests that the residuals have a bell shape, which means that they are normally distributed. Consequently, the forecasts from this developed model means that it will be quite good.

**Ten points future forecasting.**

**Zone Two Forecasting Analysis**

| | fcst | lower | upper | CI |
|---|---|---|---|---|
| 1 | -0.16261840 | -0.4962751 | 0.1710383 | 0.3336567 |
| 2 | -0.06079294 | -0.4543378 | 0.3327519 | 0.3935448 |
| 3 | 0.23503271 | -0.2104697 | 0.6805351 | 0.4455024 |
| 4 | 0.27284283 | -0.1855569 | 0.7312426 | 0.4583997 |
| 5 | 0.12222782 | -0.3651445 | 0.6096001 | 0.4873723 |
| 6 | 0.03352417 | -0.4732197 | 0.5402680 | 0.5067438 |
| 7 | 0.07436820 | -0.4399424 | 0.5886788 | 0.5143106 |
| 8 | 0.15271189 | -0.3665874 | 0.6720112 | 0.5192993 |
| 9 | 0.16461172 | -0.3558062 | 0.6850296 | 0.5204179 |

| | | | | |
|---|---|---|---|---|
| 10 | 0.11846059 | -0.4034188 | 0.6403400 | 0.5218794 |

**Table 7: Zone Two Forecasting Analysis**

The table shows the forecasted values and the intervals between the maximum and minimum values where the forecasting values will lie. It also shows the confident interval from the mean forecasted value.

**Zone two model comparison results of actual and predictive**

| | Actual value | Predicted value | Error |
|---|---|---|---|
| 1 | 0.2099334357 | 0.397611215 | -0.187677780 |
| 2 | 0.0087685612 | 0.187554743 | -0.178786182 |
| 3 | 0.0009600614 | 0.136364679 | -0.135404617 |
| 4 | 0.0641001024 | 0.008314018 | 0.055786085 |
| 5 | 0.0698604711 | 0.004570093 | 0.065290378 |
| 6 | 0.0239375320 | 0.032621564 | -0.008684032 |
| 7 | 0.0501472094 | 0.435995180 | -0.385847971 |
| 8 | 0.3279249872 | 0.304371484 | 0.023553503 |
| 9 | 0.0451228879 | 0.071784835 | -0.026661948 |
| 10 | 0.3874017 | 0.3775781 | 0.0098236 |

**Table 8: Zone Two Prediction Error Analysis**

This table above is used to compare the forecasted value from BVAR model and the actual values. It also gives the prediction error for each value. The accuracy level displayed of 84.404% shows that the forecasting is more accurate compared to forecasting using VAR model which gives an accuracy of 81.68%.

**Holtwinters Forecasting Analysis**

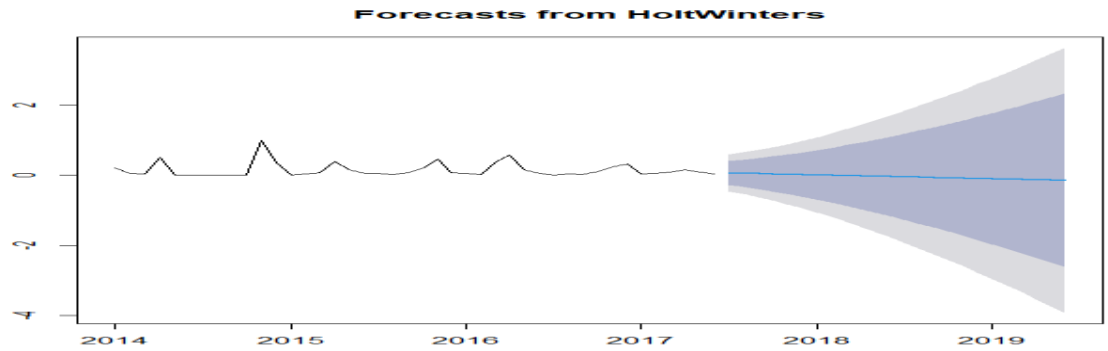


**Figure 5: HoltWinters forecast for zone two**

The figure shows the ten months forward forecasting, where the thick blue represents 80% while the light blue is 90%. The diagram shows a thin range of both sides in the two cases.

## V. Conclusions

The graphs of the initial time series data had seasonality, which prompted a need to difference, where after testing the data become stable. The stationarity test was evaluated using two tests, ADF and PP tests. The two zones were found to be stationary from the ADF and PP tests which gave a strong statistical significance of the p – values obtained. The Ganger Causality test, which was to test if there was any serial correlation and if the lag of the predictor variables influenced that of the response variable, was conducted. It was concluded that the temperature, relative humidity, atmospheric pressure, wind speed, radiation, and wind gust granger caused rainfall. This was clearly given by the statistically significant p-values in the two zones. The Ljung-Box test shows that the developed model was good for forecasting. For these purposes, the researchers established and estimated a forecasting model from the monthly meteorological variables. Applying Vector Autoregressive (VAR) method and the Bayesian method of multivariate time series analysis, it was found that the rainfall variable and other variables were interrelated. BVAR was customized for forecasting purposes, using the training data, it was possible to test and come up with the predicted values. In zone one, radiation and wind speed had very minimal influences on the dependent variable. In zone two, the wind gusts had no effect and thus it was dropped.

## VI. Recommendations

Since the predictive model has a more accurate prediction, it is recommended that it be adopted in the area of Artificial Intelligence. When comparing the weather variables for different regions, the rainfall pattern was been influenced by a number of other weather variables, therefore the study recommends that any other weather forecast influencing factors should be put into consideration. For further research, the researcher recommends the use of more weather variables like topography, cloud cover, sun shine duration, among others to improve the accuracy of the predictability. Finally, the researcher recommends the application of other techniques like Random Forest and Bootstrapping technique to check whether the accuracy may further be improved by other models.

## References


[1]. Bauer P., A. Thopre and G. Brunet, 2015: The quiet revolution of numerical weather prediction. Nature, 525:47–55.
[2]. Coumou, D. and S. Rahmstorf, 2012: A decade of weather extremes. Nature Climate Change, 2:491–49
[3]. Giannone D, Lenza M, Primiceri GE (2015). "Prior Selection for Vector Autoregressions." Review of Economics and Statistics, 97(2), 436–451
[4]. Jaynes, E. T. (2003). *Probability theory: The logic of science*. Cambridge university press.
[5]. Ji, M., A. Kumar and A. Leetmaa (2018): A multiseason climate forecast system at the National Meteorological Center. Bull. Amer. Meteo. Soc., 75:569–577
[6]. Kilian L, Lütkepohl H (2017). Structural Vector Autoregressive Analysis. Cambridge University Press. doi:10.1017/9781108164818.
[7]. Linacre, E and Geerts, B. (1997). Climates and Weather Explained Routledge London, pp. 321 – 345
[8]. Lutgens, F. K. and TarBuck, E. J. (1989). The Atmosphere: An Introduction to Meteorology, Fourth edition. Prentice Hall, New Jersey, pp. 299 – 331.
[9]. Lütkepohl, H. (2005). *New introduction to multiple time series analysis*. Springer Science & Business Media.
[10]. Mary K, Dave M, Maurine A and Joanne R. (2018). Extreme Rainfall and Flooding over Central Kenya Including Nairobi City during the Long-Rains Season.
[11]. Otiende P, & Brian M. (2009). The economic impacts of climate change in Kenya: Riparian flood impacts and cost of adaptation
[12]. Stockdale, T.N., D.L.T. Anderson, J.O.S. Alves, and M.A. Balmaseda, 1998: Global seasonal rainfall forecasts using a coupled ocean-atmosphere model. Nature, 392:371–373.


Mr. Harun Mwangi Gitonga, et. al. "Application of Bayesian Vector Autoregressive Model In Forecasting Rainfall In Kenya." *IOSR Journal of Mathematics (IOSR-JM)*, 18(3), (2022): pp. 13-23.