American Journal of Theoretical and Applied Statistics
2021; 10(6): 249-256
http://www.sciencepublishinggroup.com/j/ajtas
doi: 10.11648/j.ajtas.20211006.14
ISSN: 2326-8999 (Print); ISSN: 2326-9006 (Online)

Methodology Article

Diagnostic Analysis of Weather Variables for Forecasting Rainfall Patterns in Kenya Using Bayesian Vector Autoregressive Model

Gitonga Harun Mwangi[1], Joseph Koske[2], Mathew Kosgei[2]

[1]Department of Statistics and Actuarial Sciences, School of Pure and Applied Sciences, Kirinyaga University, Kerugoya, Kenya
[2]Department of Mathematics, Physics and Computing, School of Biological and Physical Sciences, Moi University, Mombasa, Kenya

Email address:
hgitonga@kyu.ac.ke (G. H. Mwangi), koske4@yahoo.co.uk (J. Koske), mkosgei@mu.ac.ke (M. Kosgei)

To cite this article:
Gitonga Harun Mwangi, Joseph Koske, Mathew Kosgei. Diagnostic Analysis of Weather Variables for Forecasting Rainfall Patterns in Kenya Using Bayesian Vector Autoregressive Model. *American Journal of Theoretical and Applied Statistics.* Vol. 10, No. 6, 2021, pp. 249-256.
doi: 10.11648/j.ajtas.20211006.14

Received: October 15, 2021; **Accepted**: November 5, 2021; **Published**: December 9, 2021

Abstract: Time series has fundamental importance in various practical domains in the world, more so in modeling and forecasting. Many important models have been proposed to improve the accuracy of their prediction. Global warming has been a big challenge to the world in affecting economic and Agricultural activities. It causes drastic 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 conducted a diagnostic analyzes of weather variables which were used to model the rainfall patterns by use of Bayesian Vector Autoregressive (BVAR). The diagnostic analyzes was done after the normalization of the data. The data was found to be stable after first differencing and it was tested using Augmented Dicker Fuller (ADF) and Phillips-Perron (PP) test. The tests were found to have the P-values that were statistically significant. The Granger Causality test was also conducted and found to be statistically significant. The Ljung-Box test of residuals, shows that the graphs of these residuals produced, appeared to explain all the available information in the forecasted model. The mean of the residuals was near to zero and therefore no significant correlation was witnessed. The time plot shows that the variation of the residuals remains much the same across the historical data, apart from the two values that were beyond 0.2 or -0.2 in Zone Two, and therefore the residual variance was treated as constant. 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 of the residuals. The Ljung-Box test shows that the developed model was good for forecasting. Finally, the researcher recommends application of other techniques like Random Forest and Bootstrapping technique to check whether the accuracy may further be improved from other models.

Keywords: Global Warming, Bayesian, Diagnostic Analyzes, Vector Autoregressive

1. Introduction

Kenya has experienced protracted droughts and intense flooding every year [1]. An increase in such extreme weather events, the glaciers around Mount Kenya have disappeared, leading to the drying up of rivers and streams. The weather changes have also led to harvest losses and food shortages, as well as landslides, soil degradation and a loss of biodiversity [2]. The waning water sources and unreliable rainfalls have reduced the availability of water. Climate variability and changes have adversely affected Agricultural sector and the situation is expected to worsen in the future. In the present days, weather forecast issue is resolved through the support of numerical Atmospheric Circulation Models (ACMs). These are integrated by different weather amenities on daily basis normally on coarse-grained resolution grids which covers a wide geographical coverage. The ACMs describe several meteorological variables such as humidity, temperature, wind component, geopotential among others. All these define the

predicted patterns of atmosphere for a given forecast duration. However, meteorological phenomena like rainfall, normally vary more on local scales. Numerical Weather Prediction (NWP) is a simplified set of equations so-called the primitive equation used to calculate changes of conditions [3]. The word "numerical" is deceptive since all kinds of weather forecasting are built on some quantitative data and thus could fit under this area [4]. The big number of variables that is involved when considering the dynamic atmosphere makes this task extremely difficult. Manipulating the huge 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 past about forty years, enabled by developments in observing systems and improvements in 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 taken place at major operational centers across the world [5]. Perfecting these advances in weather and climate prediction, there have been important milestones in advancing the science and operational infrastructure for predictions at extended timescales. The first generations of dynamic seasonal forecast systems were implemented at operational centers in the mid-1990s [6]. Routine weather and climate forecasts at the global and regional levels now provided 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, [7], [8], 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 almost doubled since 1975 to 1.5–1.8° Celsius per century. A steady upsurge in temperature had 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 unique long-term climatic changes had influenced sub-seasonal and seasonal-to-interannual unpredictability and had a reflective impact on the natural environment as well as on the life, health and well-being of human society, [9].

Much of the discussion around climate changes 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 together classic unconditional as well as conditional forecasts. Unconditional forecasts challenging those obtained from factor models in accuracy [10] and are used for a variety of analyses. Conditional forecasts permit for elaborate scenario analyses, where the future path of one or more variables is assumed to be known. They are handy tools for analyzing conceivable realizations of policy-relevant variables.

2. Purpose of the Study

Global warming has become a major challenge in the world. This has brought about unpredictable weather patterns that have affected the normal seasons. Extreme weather changes are identified as major global challenges of the recent times. In Kenya, unstable weather patterns which are associated with global warming have been experienced over a period of time. Despite the availability of models that are used by meteorology department to make predictions, the same devastating scenarios of unpredictable weather changes are still been experienced. Therefore, home grown models reliable for accurate predictions are needed on short and long-term time scales to reduce potential risks and damages that may occur due to unexpected weather changes. To achieve the times series models for prediction, a diagnostic analyzes is an important tool for determining the state of the data. 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 in order to develop a predictive model of rainfall patterns using Bayesian Vector Autoregressive.

3. Literature Review

To test for stationarity of the variables in the model, the Augmented Dickey-Fuller-test (ADF-test) are used. When a time series variables are independent of time and the auto covariance and the variances are not infinite, then the time series variables are said to be stationary [11]. Further, [12] when the probability distribution has no fluctuation over time, the time series variables are stationary, thus, the time series follows a random walk. Stationarity is an important criterion when using a BVAR-model [13]. If the time series is not stationary, the results from the test would not be trustworthy. In vector autoregressive (VAR) processes, if the process is stationary, the multivariate least squares (LS) estimator of the coefficients has a non-singular asymptotic distribution whereas the distribution becomes singular if some variables are integrated or cointegrated [14]. So the Wald test has a nonstandard asymptotic distribution. Differentiation is a good working method to overcome the problem of non-stationarity according to, [15]. The ADF test is based on the hypothesis testing where the null hypothesis states that the time-series variable is non-stationary [16]. In such circumstances, the variables are differenced a number of times until the ADF test shows a 5% level of significant.

When BVAR-Models are conducted, Granger Causality tests

are required to check if there is a significant association between variables. 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 [17]. Subsequently, [18] 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. Bayesian vector autoregressions are usually used for forecasting and structural analysis. Until recently, though, most empirical work had considered only small systems with a few variables due to parameter propagation concern and computational restrictions. [19].

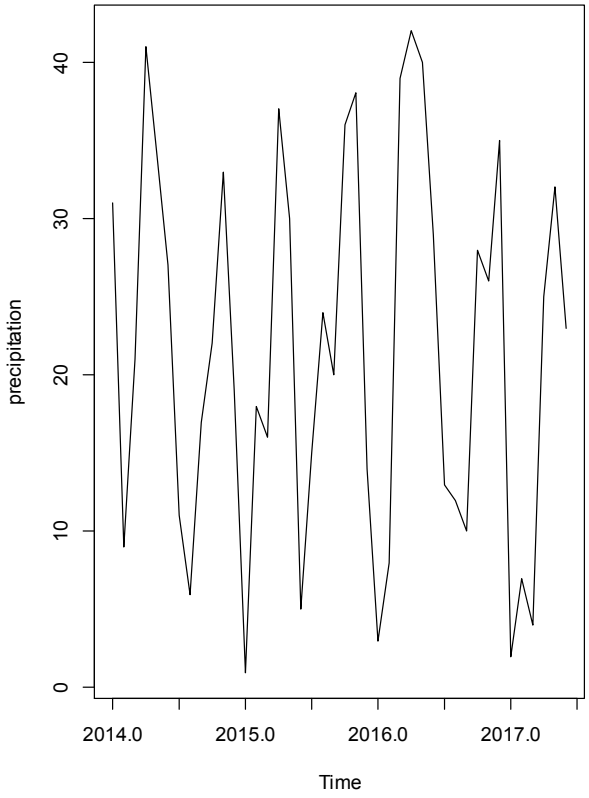

Figure 1. *Time plot graph for the initial data.*

4. Methodology

4.1. Diagnostic Procedures

The source of the data was secondary data, which was sourced from Trans- African Hydro-Meteorological Observatory (TAHMO) and Kenya Meteorological Stations. The data was stored in the form of excel format, which was captured on daily basis for a period of four years, starting from June 2014 to June 2017. The data was converted into CSV files in order to import it into R statistical software for analysis. To remove scaling, normalization was done through liner scaling technique. It was essential because all the variables used different units of measurements. Also, 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 estimations, has a formula stated as;

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

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

Before the data in time series is analyzed, the data should attain some level of stability. The study adopted two methods of stationarity test. Augmented Dickey fuller (ADF) test and Phillips- Perron (PP) test. When these two methods are used, different results are expected. Thus, the derivation of stationarity is considered at the level where both results reject the null hypothesis and therefore stationarity exists. If this assumption is violated and non-stationary data is used, then the outcome would result to an unpredictable model outcome. Many time series data in reality are not stationary and they require to be stationary in order to be analyzed. Non-stationarity can be detected by visual examining of the time series graph and by looking at the series correlogram, or by conducting a Unit Roots Statistical test. To remove non-stationarity, a time series is transformed by differencing once or several times until it becomes stationary. In this study, Unit Roots Statistical test was employed where, Dickey-Fuller test and Phillips – Perron test were used to test for stationarity.

Consider x_t a time-series which is in the form

$$x_t = \alpha + \beta x_{t-1} + u_t, \text{ where}$$

$$u_t = \rho u_{t-1} + \varepsilon_t,$$

The Unit Root tests are based on testing the null hypothesis that

H_0: ρ =1

against the alternative hypothesis that

H_1: $\rho < 1$.

The characteristic polynomial had a root equal to unity under the null hypothesis, hence the name Unit Root tests.

The Augmented Dickey-Fuller test allows for higher-order autoregressive processes by including Δx_{t-p} in the model. Taking ρ =1 then VAR(1) process is stable if all eigenvalues of β_1 have modulus less than one, this stability condition is equivalent to

$$det(\mathrm{IK} - \mathrm{B1z}) \neq 0 \text{ for } |\mathrm{z}| \leq 1$$

this can be generalized as

$$det(\mathrm{IKp} - \mathrm{Bz}) = \det(\mathrm{IK} - \mathrm{B1z} - \cdots - \mathrm{Bpzp})$$

This gives the definition of the characteristic polynomial of a matrix. The polynomial is called the reverse characteristic polynomial of the BVAR (p) process. Hence, the process is stable if its reverse characteristic polynomial has no roots in and on the complex unit circle. Since stability implies stationarity, the process is stationary when proved to be stable. The test follows AR (1) process

$$x_t = \rho x_{t-1} + u_t,$$

where, u_t is an IID series of random variables. x_t is non-stationary under the null hypothesis, and is stationary under the alternative hypothesis. The standard t-statistics would not follow t-distribution because of the non-stationarity of x_t under the null hypothesis. To test the null hypothesis, the following test statistics equation was used

$$\text{ADF} = \frac{\rho - 1}{S.e\ (\rho)}$$

The ADF test in the equation follows the assumption that the error terms are independent and identically distributed, and that the order of the underlying autoregressive process is finite and known. The procedure for the ADF test is similar to the Dickey–Fuller test procedure, the only difference is the model where it is applied. The model where ADF is applied are as shown below.

$$\Delta X_t = \alpha + \beta_t + \phi x_{t-1} + \delta 1 \Delta x_{t-1} + ... + \delta p - 1 \Delta x_{t-1} + \varepsilon_t$$

where, α denotes a constant, β is the coefficient on a time trend, and p represented the lag order of the autoregressive process. Putting the constraints $\alpha = 0$ and $\beta = 0$, this resembled a model with a random walk, and using the constraint $\beta = 0$ resembled a model of random walk with a drift. The ADF test was performed under the hypothesis

H0: $\phi = 1$ against H1: $\phi < 1$

The test statistic was computed as

$$\text{ADF} = \frac{\hat{\phi}}{SE(\hat{\phi})}$$

If the ADF τ test statistic is less than the critical value, then the null hypothesis of $\phi = 0$ is rejected and no unit root is present. When null hypothesis was not rejected, it meant that the time series was not stationary and required at least differencing once.

The Phillips-Perron (PP) test is an alternative technique for correcting the serial correlation in Unit Root testing. The PP test uses the standard DF or ADF test, but modifies the t-ratio so as to prevent serial correlation to affect the asymptotic distribution of the test statistic. The difference between the PP and ADF tests is in terms of how the tests deal with the issue of serial correlation and heteroskedasticity in the error terms. The test model for the PP test is given as

$$\Delta x_t = \beta' D_t + \pi x_{t-1} + \mu_t$$

Where μ_t denoted I(0) which may be heteroskedastic. The PP tests correct the serial correlation and heteroskedasticity in the error terms μ_t of the tested model, by directly modifying the Dickey-Fuller test statistics $t\pi = 0$ and $T\hat{\pi}$. The test statistics denoted by Z_t and $Z\pi$ are given as:

$$Z_t = \left\{ \frac{\delta^2}{\lambda^2} \right\} 2.\ t\pi = 0 - \frac{\hat{1}}{2} \left\{ \frac{\lambda^2 - \delta^2}{\hat{\lambda^2}} \right\}. \left\{ \frac{T..SE(\pi)}{\hat{\delta^2}} \right\}$$

$$Z\pi = T\hat{\pi} - \frac{\hat{1}}{2}. \left\{ \frac{T..SE(\pi)}{\hat{\delta^2}} \right\} (\hat{\lambda}^2 - \hat{\delta}^2)$$

The estimated variance parameters of:

$$\delta^2 = \lim_{T \to \infty} T^{-1} \sum_{i=1}^{T} E[\mu_t^2] \text{ and}$$

$$\lambda^2 = \lim_{T \to \infty} \sum_{i=1}^{T} E[T^{-1} S_t^2]$$

are $\sigma^2$and λ^2 Where $S_T = \sum_{i=1}^{n} \mu_t$. The sample variance of the least square residual $\hat{\mu_t}$ is a consistent estimator of σ^2, and the Newey - West long-run variance estimates of μ_t using $\hat{\mu_t}$ is a consistent estimator of λ^2. Under the null hypothesis that $\pi = 0$, the PP, Z_t and Z_π statistics have the same asymptotic distributions as the ADF t-statistics and normalized bias statistics. The advantage of the PP tests over the ADF tests is that the PP tests are robust to general forms of heteroskedasticity in the error term μ_t and the user need not to specify a lag length for the test regression.

The basis behind Bayesian Vector Autoregression is that each of the time series in the system influences each other. This relationship needs to be tested first using Granger's Causality test before the model building. So, what does Granger's Causality really test? Granger Causality tests the null hypothesis that the coefficients of past values in the regression equation is zero. In simple terms, the past values of time series (x) do not cause the other series (y). Therefore, if the p-value gotten from the test is smaller than the significance level of 0.05, then, the null hypothesis is rejected. Generally, believing that a present or future event could have been caused by a past event, this would be identified by a Granger Causality. This was the impetuous for the Granger Causality test on time series data which gave evidence that one variable caused the others. The test is based on ordinary least square regression model and on the null hypothesis test. Based on, does x Granger cause y? If it does not, the null hypothesis is rejected. The test is based on the following Ordinary Least Square Regression model

$$y_i = a_0 = + \sum_{j=1}^{m} \alpha_j y_{i-j} + \sum_{j=1}^{m} \beta_j x_{i-j} + \varepsilon_i$$

Here, the α_j and β_j are the regression coefficient and ε_i is the error term. The test is based on the null hypothesis:

$$H_0: \beta_1 = \beta_2 = \ldots\ldots\ldots\ldots = \beta_M = 0$$

If the p – value for the test is less than the designed value of alpha, then the null hypothesis is rejected and Granger Causality exists. It is assumed that the data is stationary, but if it were not the case, then differencing is employed before using the Granger Causality test.

4.2. Residual Test Using Ljung-Box Test

The "residuals" in a time series model are the errors that occurs after fitting a model. For several time series models, the residuals are equal to the difference between the observations and the corresponding fitted values.

$$et = yt - y'_t$$

To checking whether a model has adequately captured the information in the data, residuals are beneficial. The following

properties will be yielded in the residuals to depicted a good forecasting method. If there are correlations between residuals, then there is information left in the residuals which should be not be used in computing forecasts. The residuals are uncorrelated if;

The residuals have zero mean.

The residuals have constant variance.

The residuals are normally distributed.

5. Findings

To ensure that the time series data contains no flaws, is stable and it is not affected by serial correlation, diagnostic analysis is 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.

Time plot graph for the data.

Time series graph for zone one.

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

5.1. Stationarity Test

Augmented Dickey fuller (ADF) and Phillips-Perron (PP) test. The Augmented Dickey-Fuller (ADF) test was implemented to check whether the variables were stationary or not. The results for two zones and global vector are used as the representation of this study.

5.1.1. Zone One

The table 1 below shows the results of dependent and independent variables under ADF and PP tests.

Table 1. *Zone One Stationarity Test.*

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

It shows that under ADF and PP test for Zone one for all the variables are stationary.

5.1.2. Zone Two

The table 2 below, considered the variable *X* which showed ADF and Phillips-Perron Test Statistics.

Table 2. *Zone Two Stationarity Test.*

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

The result shows that the variables were stable and they can be used for time series analysis.

5.1.3. Global Vector

Table 3. *Global Vector Stationarity Test.*

Variables	ADF Test Statistics	Phillips-Perron	Truncation lag parameter	P-Value ADF	P-Value P. P	Remarks
X	-4.813	-26.706	3	0.01	0.01	Stationary
X_1	-2.4932	-29.47	3	0.03792	0.01	Stationary
X_2	-2.8393	-22.79	3	0.02428	0.01865	Stationary
X_3	-2.7592	-16.337	3	0.02744	0.01019	Stationary
X_4	-3.2486	-14.789	3	0.0298	0.01997	Stationary
X_5	-2.0473	-19.717	3	0.05155	0.0437	Stationary
X_6	-2.8747	-16.708	3	0.02288	0.0382	Stationary

Table 3 Represent stationarity of the Global Vector.

5.2. Granger Causality Test

5.2.1. Zone One

The Granger causality test was done, from the findings, model 1: x ~ Lags(x, 1:6) + Lags(x1, 1:6) and Model 2: x ~ Lags(x, 1:6) while the p value was 0.02134 *. For x2 against x, Granger causality test shown that Model 1: x ~ Lags(x, 1:6) + Lags(x2, 1:6). Model 2: x ~ Lags(x, 1:6) while the p value was 0.02139 *. For x3 against x, Granger causality test shown that Model 1: x ~ Lags(x, 1:6) + Lags(x3, 1:6). Model 2: x ~ Lags(x, 1:6) while the p value was 0.005139 **. For x4 against x, Granger causality test shown that Model 1: x ~ Lags(x, 1:6) + Lags(x3, 1:6). Model 2: x ~ Lags(x, 1:6) while the p value was 0.0171 *. For x5 against x, Granger causality test shown that Model 1: x ~ Lags(x, 1:6) + Lags(x3, 1:6). Model 2: x ~ Lags(x, 1:6) while the p value was 0.02694 *. For x6 against x, Granger causality test shown that Model 1: x ~ Lags(x, 1:6) + Lags(x6, 1:6). Model 2: x ~ Lags(x, 1:6) while the p value was 0.05171.

Therefore, the granger causality test for Zone one shows that x1 to x5 had a significant influence on the Cause of dependent variable x but x6 had minimal influence on the x variable. However, the variable was still used in the analysis since its significance value was closer to 0.05.

5.2.2. Zone Two

Table 3 showed that the Granger causality test Model 1: x ~ Lags(x, 1:6) + Lags(x1, 1:6).

Model 2: x ~ Lags(x, 1:6) while the p value was 0.007123 **. For x2 against x, Granger causality test showed that Model 1: x ~ Lags(x, 1:6) + Lags(x2, 1:6) Model 2: x ~ Lags(x, 1:6) while the p value was 0.004149 **. For x3 against x, Granger causality test showed that Model 1: x ~ Lags(x, 1:6) + Lags(x2, 1:6), Model 2: x ~ Lags(x, 1:6) while the p value was 0.006399 **. For x4 against x, Granger causality test showed that Model 1: x ~ Lags(x, 1:6) + Lags(x2, 1:6, Model 2: x ~ Lags(x, 1:6) while the p value was 0.006399 **. For x5 against x, Granger causality test showed that Model 1: x ~ Lags(x, 1:6) + Lags(x5, 1:6, Model 2: x ~ Lags(x, 1:6) while the p value was 0.03492 * For x6 against x, Granger causality test showed that Model 1: x ~ Lags(x, 1:6) + Lags(x5, 1:6) Model 2: x ~ Lags(x, 1:6) while the p value was 0.03723 *

Therefore, the granger causality test for Zone two showed that all the variables were having a strong significant influence on the Causes of dependent variable x. Their level of significant was less than 0.05.

5.2.3. Global Vector

The hypothesis was that rainfall is not granger caused by temperature, humidity, wind, wind gust, Atmospheric pressure and radiation.

For x1 against x, Granger causality test showed that Model 1: x ~ Lags(x, 1:8) + Lags(x1, 1:8).

Model 2: x ~ Lags(x, 1:8) while the p value was 0.05 **. For x2 against x, Granger causality test showed that Model 1: x ~ Lags(x, 1:8) + Lags(x2, 1:8), Model 2: x ~ Lags(x, 1:8) while the p value was 0.0002901 ***. For x3 against x, Granger causality test showed that Model 1: x ~ Lags(x, 1:8) + Lags(x2, 1:8). Model 2: x ~ Lags(x, 1:8) while the p value was 0.06 ***. For x4 against x, Granger causality test showed that Model 1: x ~ Lags(x, 1:8) + Lags(x4, 1:8) Model 2: x ~ Lags(x, 1:8) while the p value was 0.06 ***. For x5 against x, Granger causality test showed that Model 1: x ~ Lags(x, 1:8) + Lags(x4, 1:8). Model 2: x ~ Lags(x, 1:8) while the p value was 0.05 ***. For x6 against x, Granger causality test showed that Model 1: x ~ Lags(x, 1:8) + Lags(x6, 1:8), Model 2: x ~ Lags(x, 1:8) while the p value was 0.0274 *.

Therefore, the global vector had a strong granger causality influence on the dependent variable x. These have been depicted by strong level of significant in each case.

5.3. 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 displayed 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.

5.3.1. Zone One

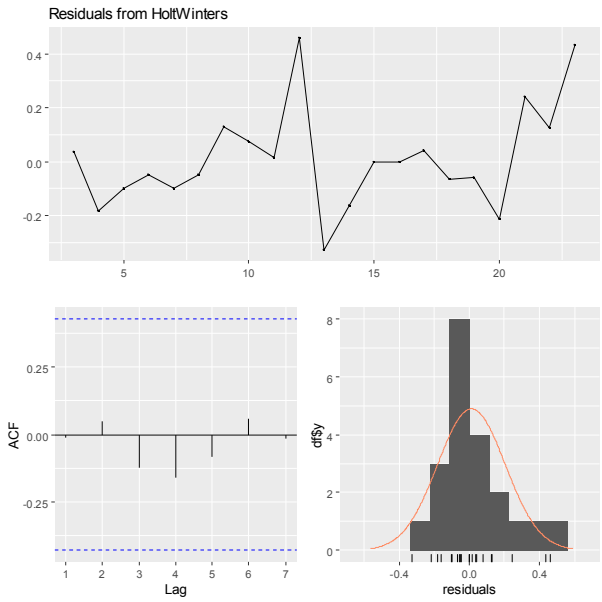

Figure 2. *Zone One checking residuals behavior.*

These graphs show that the residual 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 figure 2 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 a normal distribution of the residual, 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 of the residuals.

5.3.2. Zone Two

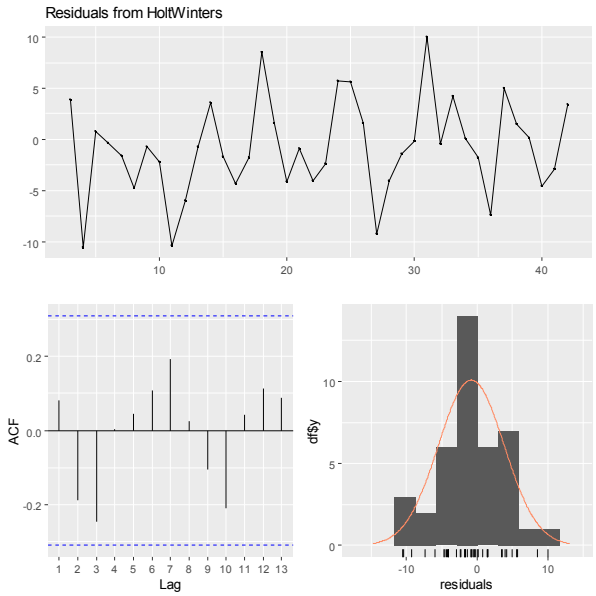

Figure 3. *Zone Two checking residuals behavior.*

These graphs show that the residual method gives 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 figure 3 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, forecasts from this developed model means that it will be quite good.

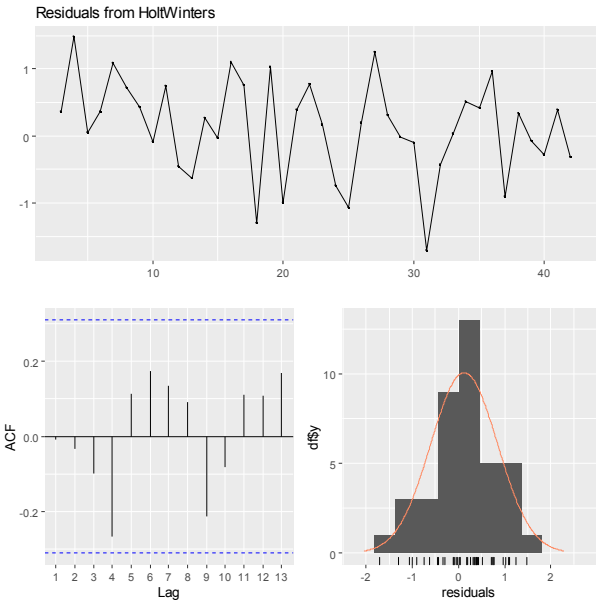

Figure 4. *Global Vector checking residuals behavior.*

5.3.3. Global Vector

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 are normally distributed. The mean of the residuals is close to zero and there is no significant correlation in the residual series. This shows that the model developed can make a good forecast.

6. Conclusions

The graphs of the initial time series data had seasonality, which prompted a need to difference once. The stationarity test was evaluated using two tests, ADF and PP tests. The conclusion was arrived at when both ADF and PP tests rejected the null hypothesis, thus the data was treated to be stationary. All 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 lags of the predictor variables influenced that of 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 all the zones. The Ljung-Box test shows that the developed model is good for forecasting.

7. Recommendations

For further research, the researcher recommends use of more weather variables like topography, cloud cover, sun shine duration among others to improve the accuracy of the predictability.

The study used secondary data for 2014 to 2017, therefore the researcher recommends that current data for 2020 and 2021 may be used to make current future predictions.

Climate models, only predict a range of possible future scenarios, the extent of how far the future would be should be studied.

Finally, the researcher recommends application of other techniques like Random Forest and Bootstrapping technique to check whether the accuracy may further be improved from other models.

References

[1] Mary Kilavi, Dave M, Maurine A and Joanne R. (2018). Extreme Rainfall and Flooding over Central Kenya Including Nairobi City during the Long-Rains Season.

[2] Otiende P, & Brian M. (2009). The economic impacts of climate change in Kenya: Riparian flood impacts and cost of adaptation.

[3] Linacre, E and Geerts, B. (2016). Climates and Weather Explained Routledge London, pp. 321 – 345.

[4] Lutgens, F. K. and TarBuck, E. J. (2016). The Atmosphere: An Introduction to Meteorology, Fourth edition. Prentice Hall, New Jersey, pp. 299 – 331.

[5] Bauer P., A. Thopre and G. Brunet, 2015: The quiet revolution of numerical weather prediction. Nature, 525: 47–55.

[6] Stockdale, T. N., D. L. T. Anderson, J. O. S. Alves, and M. A. Balmaseda, 2010: Global seasonal rainfall forecasts using a coupled ocean-atmosphere model. Nature, 392: 371–373.

[7] Blunden, J., D. S. Arndt, and G. Hartfield (eds.), 2018: State of the Climate in 2017. Bull. Amer. Meteor. Soc., 99: 8, Si–S310, doi: 10.1175/2018BAMSStateoftheClimate.1.

[8] 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.

[9] Coumou, D. and S. Rahmstorf, 2012: A decade of weather extremes. Nature Climate Change, 2: 491–49.

[10] Giannone D, Lenza M, Primiceri GE (2015). "Prior Selection for Vector Autoregressions." Review of Economics and Statistics, 97 (2), 436–451.

[11] Verbeek, Marno. (2008) A guide to modern econometrics. 3rd ed. Chichester, England ; Hoboken, NJ: John Wiley & Sons.

[12] Stock, J. H., & Watson, M. W. (2015). Generalized shrinkage methods for forecasting using many predictors. *Journal of Business & Economic Statistics*, *30* (4), 481-493.

[13] Asteriou, Dimitrios, och Stephen G. Hall. Applied Econometrics: A Modern Approach Using EViews and Microfit. Rev. ed. Basingstoke: Palgrave Macmillan, 2021.

[14] Ratsimalahelo, Z. (2017) Generalised Wald Type Test of Nonlinear Restrictions. *Open Access Library Journal*, 4, 1-8. doi: 10.4236/oalib.1103923.

[15] Subba Rao and J. Yang. Reconciling the Gaussian and Whittle likelihood with an application to estimation in the frequency domain. arXiv preprint arXiv: 2001.06966, 2020.

[16] Sumitra Iyer, Alka Mahajan. "Predicting total electron content in ionosphere using vector autoregression model during geomagnetic storm", Journal of Applied Geodesy, 2021.

[17] Korobilis, D and Pettenuzzo, D (2016) Adaptive Minnesota Prior for High= D imensional Vector Autoregressions. Working paper Essex Finance Centre working paper, Colchest.

[18] Lütkepohl, H. (2005). *New introduction to multiple time series analysis*. Springer Science & Business Media.

[19] Chan, Joshua C. C and Chan, Joshua. C. Large Bayesian Vector Autoregressions (February 14, 2019) CAMA working paper No 19/2019, Available at SSRN.