JULIUS-MAXIMILIANS-UNIVERSITÄT WÜRZBURG

INSTITUT FÜR GEOGRAPHY UND GEOLOGIE

PHSISCHE GEOGRAPHY

PROF.DR. HEIKO PAETH

East African Seasonal Rainfall prediction using multiple linear regression and regression with ARIMA errors models

---

Dissertation zur Erlangung des naturwissenschftlichen Doktorgrade der Bayerischen Julius-Maximilians-Universität Wurzburg

Vorgerlegt von

Alphonse KARAMA

Aus Ruhango (Rwanda)

Würzburg, October 2021

Eingereicht am: 20. October 2021

1. Gutachter: Prof.Dr.Heiko Paeth
2. Gutachter: Prof.Dr. Barbara Sponholz

der dissertation

1. Prüfer: Prof.Dr.Heiko Paeth
2. Gutachter: Prof.Dr. Barbara Sponholz

der mündlichen Prüfung

Tag der mündlich Prüfung: 17.December 2021

Doktorurkunde ausgehändigt am:

"*Viewed from the distance of the moon, the astonishing thing about the earth, catching the breath, is that it is alive. The photographs show the dry, pounded surface of the moon in the foreground, dead as an old bone. Aloft, floating free beneath the moist, gleaming membrane of bright blue sky, is the rising earth, the only exuberant thing in this part of the cosmos. If you could look long enough, you would see the swirling of the great drifts of white cloud, covering and uncovering the half-hidden masses of land. If you had been looking for a very long, geologic time, you could have seen the continents themselves in motion, drifting apart on their crustal plates, held afloat by the fire beneath. It has the organized, self-contained look of a live creature, full of information, marvelously skilled in handling the sun*."

Lewis Thomas (1974), *The lives of a cell*

Acknowledgments

I am extremely grateful to my supervisor, Prof. Dr. Heiko Paeth for his support and guidance during the running of this project.

My sincere thanks go to all my colleagues at the Physical geography at University of Würzburg. I would like to say a big thank you to all of you for the friendship you extended to me and academic assistance some of you provided me. Worth mentioning here are Dorothee Schill, Dr.Felix Pollinger, Rai,Dr.Praveen Kumar, Mengije Warmuth, Katrin Ziegler, Christian Hartmann, Daniel Abel, Daniel Schönbein, Luzia Keupp and Freddy Bangelesa.

I am grateful to the Government of Rwanda and German Academic Exchange Service (DAAD) for providing me financial assistance for 45 months to this study. Special thanks go to the University of Würzburg for providing an extension of PhD and financial assistance for 6 months to complete this study. I also express my profound gratitude to the Institut d'Enseignement Superieur de Ruhengeri (INES-Ruhengeri) which has provided me a study leave in order to conduct this study.

The Special thanks go to the Climate Research Unit (CRU) of the University of East Anglia (UK) who released the CRU TS 4.1 observational datasets used in this study. I also acknowledge the supply of reanalysis data of NCEP/NCAR provided by the National Oceanic and Atmospheric Administration (NOAA/OAR/ESRL PSD, Colorado, USA) from their different web sites.

Many thanks to my father Didace RUHAKANA and mother Alvera KANTAMAGE who taught me the work ethic and effort. Finally, I am also grateful to all the members of my family and friends who supported me in prayers and by way of encouragement.

I would like to express my gratitude to my family, in particular my two sons Jospin RUGAGI and Chrispin RUKUNDO and my daughter Olga UMUHIRE and I appreciate your patience and acceptance to walk away from you for a while. I would also like to express my great thanks to my wife Charlotte MUKESHIMANA, without her cooperation I could even think of doing the work.

Last but not least, big thanks go to all members of community of Saint Egidio in Würzburg for their understanding, encouragement and support.

Abstract

Nowadays, there are a lot of methods and technic available to analyse and forecast time series. One of the most used is the methodology based on autoregressive integrated moving Average (ARIMA) model by Box and Jenkins. This method uses historical data of univariate time series to analyse its own trend and forecast future cycle. Rainfall time series are often affected by different climate phenomenon which are summarized in form of indices. Therefore, one can incorporate one or more time-series in a model to predict the value of another series, by using regression with ARIMA errors (RARIMAE).

The detrimental impacts of climate variability on water, agriculture, and food resources in East Africa underscore the importance of reliable seasonal climate prediction. To overcome this difficulty RARIMAE method were evolved. Applications RARIMAE in the literature shows that amalgamating different methods can be an efficient and effective way to improve the forecasts of time series under consideration. With these motivations, attempt have been made to develop a multiple linear regression model (MLR) and a RARIMAE models for forecasting seasonal rainfall in east Africa under the following objectives:

1. To develop MLR model for seasonal rainfall prediction in East Africa.
2. To develop a RARIMAE model for seasonal rainfall prediction in East Africa.
3. Comparison of model's efficiency under consideration

In order to achieve the above objectives, the monthly precipitation data covering the period from 1949 to 2000 was obtained from Climate Research Unit (CRU). Next to that, the first differenced climate indices such as Southern Oscillation Index (SOI), Multivariate ENSO Index (MEI), Nino3.4, Dipole Mode Index (DMI), Indian Summer Monsoon Index (ISMI), Indian Monsoon Rainfall (IMR), The South Atlantic Ocean Dipole (SAOD), North Atlantic Oscillation (NAO), the Southern Annular Mode (SAM), Quasi-biennial Oscillation (QBO), Indian ocean Sea Surface Temperature (SST), Indian ocean Sea Level Pressure (SLP) and their respective lead times are considered as potential predictors. Because of a large number of potential predictors that influence the dependent variable (precipitation), the variable selection for the model has been carefully made with some theoretical background. Firstly, the variables which are significantly correlated with precipitation time series at 5% level of significance are retained. Secondly, a threshold of 5 which indicates the highest acceptable degree of multicollinearity in this research was applied on the Variance Inflation Factor (VIF) stepwise

selection. Finally, the forward stepwise regression has been used to determine variables which should enter in model construction.

In the first part of this study, the analyses of the rainfall fluctuation in whole Central-East Africa region which span over a longitude of 15˚E to 55˚E and a latitude of 15˚S to 15˚N was done by the help of maps. For models' comparison, the R-squared (R^2) values for the MLR model are subtracted from the R^2 values of RARIMAE model. The results show positive values which indicates that R^2 is improved by RARIMAE model. On the other side, the root mean square errors (RMSE) values of the RARIMAE model are subtracted from the RMSE values of the MLR model and the results show negative value which indicates that RMSE is reduced by RARIMAE model for training and testing datasets.

For the second part of this study, the area which is considered covers a longitude of 31.5˚E to 41˚E and a latitude of 3.5˚S to 0.5˚S. This region covers Central-East of the Democratic Republic of Congo (DRC), north of Burundi, south of Uganda, Rwanda, north of Tanzania and south of Kenya. Considering a model constructed based on the average rainfall time series in this region, the long rainfall season counts the nine months lead of the first principal component of Indian sea level pressure (SLP_PC19) and the nine months lead of Dipole Mode Index (DMI_LR9) as selected predictors for both statistical and predictive model. On the other side, the short rainfall season counts the three months lead of the first principal component of Indian sea surface temperature (SST_PC13) and the three months lead of Southern Oscillation Index (SOI_SR3) as predictors for predictive model. For short rainfall season statistical model SAOD current time series (SAOD_SR0) was added on the two predictors in predictive model. By applying a MLR model it is shown that the forecast can explain 27.4% of the total variation and has a RMSE of 74.2mm/season for long rainfall season while for the RARIMAE the forecast explains 53.6% of the total variation and has a RMSE of 59.4mm/season. By applying a MLR model it is shown that the forecast can explain 22.8% of the total variation and has a RMSE of 106.1 mm/season for short rainfall season predictive model while for the RARIMAE the forecast explains 55.1% of the total variation and has a RMSE of 81.1 mm/season.

From such comparison, a significant rise in R^2, a decrease of RMSE values were observed in RARIMAE models for both short rainfall and long rainfall season averaged time series. In terms of reliability, RARIMAE outperformed its MLR counterparts with better efficiency and accuracy. Therefore, whenever the data suffer from autocorrelation, we can go

for MLR with ARIMA error, the ARIMA error part is more to correct the autocorrelation thereby improving the variance and productiveness of the model.

Zusammenfassung

Heutzutage stehen viele Methoden und Techniken zur Verfügung, um Zeitreihen zu analysieren und zu prognostizieren. Eine der am häufigsten verwendeten ist die Methode, die auf dem autoregressiven integrierten gleitenden Durchschnitt (ARIMA)-Modell von Box und Jenkins basiert. Diese Methode verwendet historische Daten von univariaten Zeitreihen, um ihren eigenen Trend zu analysieren und zukünftige vorherzusagen. Niederschlagszeitreihen werden oft von verschiedenen Klimaphänomenen beeinflusst, die in Form von Indizes zusammengefasst werden. Daher kann man eine oder mehrere Zeitreihen in ein Modell integrieren, um den Wert einer anderen Reihe vorherzusagen, indem man die Regression mit ARIMA-Fehlern (RARIMAE) verwendet.

Die nachteiligen Auswirkungen der Klimavariabilität auf Wasser, Landwirtschaft und Nahrungsressourcen in Ostafrika unterstreichen die Bedeutung einer zuverlässigen saisonalen Klimavorhersage. Um diese Schwierigkeit zu überwinden, wurden die RARIMAE-Methoden entwickelt. Anwendungen RARIMAE in der Literatur zeigt, dass die Zusammenführung verschiedener Methoden ein effizienter und effektiver Weg sein kann, um die Vorhersagen der betrachteten Zeitreihen zu verbessern. Aus dieser Motivation heraus wurde versucht, ein multiples lineares Regressionsmodell (MLR) und ein RARIMAE-Modell zur Vorhersage von saisonalen Niederschlägen in Ostafrika unter folgenden Zielsetzungen zu entwickeln:

1. Entwicklung eines MLR-Modells für die saisonale Niederschlagsvorhersage in Ostafrika.
2. Entwicklung eines RARIMAE-Modells für die saisonale Niederschlagsvorhersage in Ostafrika.
3. Vergleich der betrachteten Modelleffizienz.

Um die oben genannten Ziele zu erreichen, wurden die monatlichen Niederschlagsdaten für den Zeitraum von 1949 bis 2000 von der Climate Research Unit (CRU) bezogen. Daneben die ersten differenzierten Klimaindizes wie Southern Oscillation Index (SOI), Multivariate ENSO Index (MEI), Nino3.4, Dipole Mode Index (DMI), Indian Summer Monsoon Index (ISMI), Indian Monsoon Rainfall (IMR) , Der Südatlantik-Dipol (SAOD), Nordatlantische Oszillation (NAO), Südlicher Ringmodus (SAM), Quasi-zweijährige Oszillation (QBO), Meeresoberflächentemperatur im Indischen Ozean (SST), Meeresspiegeldruck im Indischen Ozean (SLP) und ihre jeweiligen Vorlaufzeiten gelten als potenzielle Prädiktoren. Aufgrund

einer großen Anzahl potenzieller Prädiktoren, die die abhängige Variable (Niederschlag) beeinflussen, wurde die Variablenauswahl für das Modell mit einigem theoretischem Hintergrund sorgfältig getroffen. Zunächst werden die Variablen beibehalten, die signifikant mit den Niederschlagszeitreihen auf einem Signifikanzniveau von 5 % korrelieren. Zweitens wurde ein Schwellenwert von 5, der den höchsten akzeptablen Grad an Multikollinearität in dieser Untersuchung anzeigt, auf die schrittweise Auswahl des Varianz-Inflationsfaktors (VIF) angewendet. Schließlich wurde die schrittweise Vorwärtsregression verwendet, um Variablen zu bestimmen, die in die Modellkonstruktion eingehen sollten.

Im ersten Teil dieser Studie wurden die Analysen der Niederschlagsfluktuation in der gesamten Region Zentral-Ostafrika, die sich über einen Längengrad von 15˚O bis 55˚O und einen Breitengrad von 15˚S bis 15˚N erstreckt, von der von Karten. Für den Modellvergleich werden die R-Quadrat-(R^2)-Werte für den MLR-Modell von den R^2-Werten des RARIMAE-Modells abgezogen. Die Ergebnisse zeigen positive Werte, was darauf hinweist, dass R2 durch das RARIMAE-Modell verbessert wird. Auf der anderen Seite werden die Root-Mean-Square-Error (RMSE)-Werte des RARIMAE-Modells von den RMSE-Werten des MLR-Modell subtrahiert und die Ergebnisse zeigen einen negativen Wert, was darauf hinweist, dass der RMSE durch das RARIMAE-Modell für Trainings- und Testdatensätze reduziert wird.

Für den zweiten Teil dieser Studie umfasst das betrachtete Gebiet einen Längengrad von 31,5˚O bis 41˚O und einen Breitengrad von 3,5˚S bis 0,5˚S. Diese Region umfasst den Zentral-Osten der Demokratischen Republik Kongo (DRC), nördlich von Burundi, südlich von Uganda, Ruanda, nördlich von Tansania und südlich von Kenia. Betrachtet man ein Modell, das auf der durchschnittlichen Niederschlagszeitreihe in dieser Region basiert, zählt die lange Regensaison den neunmonatigen Vorsprung der ersten Hauptkomponente des indischen Meeresspiegeldrucks (SLP_PC19) und den neunmonatigen Vorsprung des Dipolmodus-Index (DMI_LR9) als ausgewählte Prädiktoren für statistische und prädiktive Modelle. Auf der anderen Seite zählt die kurze Regenzeit den dreimonatigen Vorsprung der ersten Hauptkomponente der indischen Meeresoberflächentemperatur (SST_PC13) und den dreimonatigen Vorsprung des Southern Oscillation Index (SOI_SR3) als Prädiktoren für das Vorhersagemodell. Für das statistische Modell der kurzen Regenzeit wurde die aktuelle SAOD-Zeitreihe (SAOD_SR0) zu den beiden Prädiktoren im Vorhersagemodell hinzugefügt. Durch die Anwendung eines MLR-Modell wird gezeigt, dass die Vorhersage 27,4 % der Gesamtvariation erklären kann und einen RMSE von 74,2 mm/Saison für eine lange Regenzeit hat, während für RARIMAE die Vorhersage 53,6% der Gesamtvariation erklärt und einen

RMSE von 59,4 . hat mm/Jahreszeit. Durch die Anwendung eines MLR-Modell wird gezeigt, dass die Vorhersage 22,8% der Gesamtvariation erklären kann und einen RMSE von 106,1 mm/Saison für das Vorhersagemodell für kurze Regenzeiten hat, während die Vorhersage für das RARIMAE 55,1% der Gesamtvariation erklärt und einen RMSE hat von 81,1 mm/Saison.

Aus einem solchen Vergleich wurde ein signifikanter Anstieg von R^2 und eine Abnahme der RMSE-Werte in RARIMAE-Modellen für gemittelte Zeitreihen sowohl für kurze Regenfälle als auch für lange Regenzeiten beobachtet. In Bezug auf die Zuverlässigkeit übertraf RARIMAE seine MLR-Pendants mit besserer Effizienz und Genauigkeit. Wenn die Daten unter Autokorrelation leiden, können wir uns daher für MLR mit ARIMA-Fehler entscheiden. Das ARIMA-Fehler Modell dient mehr dazu, die Autokorrelation zu korrigieren, wodurch die Varianz und Produktivität des Modells verbessert werden.

Résumé

De nos jours, il existe de nombreuses méthodes et techniques disponibles pour analyser et prévoir des séries chronologiques, dont le model de moyenne mobile intégrée autorégressive (ARIMA)de Box et Jenkins, qui s'avère le plus utilisé. Le model ARIMA utilise des données historiques de séries chronologiques univariées pour analyser sa propre tendance et prévoir le cycle futur. Les séries temporelles de précipitations sont souvent affectées par différents phénomènes climatiques appelées communément indices climatiques. Par conséquent, une ou plusieurs séries chronologiques peuvent être incorporer dans un modèle pour prédire la valeur d'une autre série, en utilisant la régression avec les erreurs ARIMA (RARIMAE).

Les impacts néfastes de la variabilité climatique sur l'eau, l'agriculture et les ressources alimentaires en Afrique de l'Est soulignent l'importance d'une prévision climatique saisonnière fiable. Pour surmonter cette difficulté, la méthode RARIMAE a été développée. Applications RARIMAE dans la littérature montre que la fusion de différentes méthodes peut être un moyen efficace et efficient d'améliorer les prévisions des séries temporelles considérées. Avec ces motivations, des tentatives ont été faites pour développer un modèle de régression linéaire multiple (MLR) et un modèle RARIMAE pour la prévision des précipitations saisonnières en Afrique de l'Est sous les objectifs suivants :

1. Développer un modèle MLR pour la prévision des précipitations saisonnières en Afrique de l'Est.
2. Développer un modèle RARIMAE pour la prévision des précipitations saisonnières en Afrique de l'Est.
3. Comparer l'efficacité du modèle considéré

Afin d'atteindre les objectifs ci-dessus, les données mensuelles sur les précipitations couvrant la période de 1949 à 2000 ont été obtenues auprès de l'Unité de Recherche sur le Climat (CRU). À côté de cela, les premiers indices climatiques différenciés tels que l'indice d'oscillation australe (SOI), l'indice ENSO multivarié (MEI), Nino3.4, l'indice de mode dipolaire (DMI), l'indice de mousson d'été indien (ISMI), les précipitations de mousson indienne (IMR) , Le dipôle de l'océan Atlantique Sud (SAOD), l'oscillation nord-atlantique (NAO), le mode annulaire austral (SAM), l'oscillation quasi-biennale (QBO), la température de surface de la mer de l'océan Indien (SST), la pression au niveau de la mer de l'océan Indien (SLP) et leurs délais respectifs sont considérés comme des prédicteurs potentiels. En raison d'un grand nombre de prédicteurs potentiels qui influencent la variable dépendante

(précipitations), la sélection des variables pour le modèle a été soigneusement effectuée avec un certain contexte théorique. Premièrement, les variables qui sont significativement corrélées avec les séries chronologiques des précipitations à un niveau de signification de 5% sont retenues. Deuxièmement, un seuil de 5 qui indique le degré de multicolinéarité acceptable le plus élevé dans cette recherche a été appliqué à la sélection par étapes du facteur d'inflation de la variance (VIF). Enfin, la régression pas à pas vers l'avant a été utilisée pour déterminer les variables qui devraient entrer dans la construction du modèle.

Dans la première partie de cette étude, les analyses de la fluctuation des précipitations dans toute la région de l'Afrique centrale et orientale qui s'étendent sur une longitude de 15˚E à 55˚E et une latitude de 15˚S à 15˚N ont été effectuées à l'aide de cartes. Pour la comparaison des modèles, les valeurs R au carré (R^2) pour le model MLR sont soustraites des valeurs R^2 du modèle RARIMAE. Les résultats montrent des valeurs positives qui indiquent que R^2 est amélioré par le modèle RARIMAE. D'un autre côté, les valeurs d'erreur quadratique moyenne (RMSE) du modèle RARIMAE sont soustraites des valeurs RMSE du model MLR et les résultats montrent une valeur négative qui indique que le RMSE est réduit par le modèle RARIMAE pour l'entraînement et le test des ensembles de données.

Pour la deuxième partie de cette étude, la zone considérée couvre une longitude de 31,5˚E à 41˚E et une latitude de 3,5˚S à 0,5˚S. Cette région couvre le Centre-Est de la République Démocratique du Congo (RDC), le nord du Burundi, le sud de l'Ouganda, le Rwanda, le nord de la Tanzanie et le sud du Kenya. Considérant un modèle construit sur la base de la série chronologique des précipitations moyennes dans cette région, la longue saison des pluies compte les neuf mois d'avance de la première composante principale de la pression au niveau de la mer indienne (SLP_PC19) et les neuf mois d'avance de l'indice de mode dipolaire (DMI_LR9) comme prédicteurs sélectionnés pour le modèle statistique et prédictif. D'un autre côté, la courte saison des pluies compte les trois mois d'avance de la première composante principale de la température de surface de la mer indienne (SST_PC13) et les trois mois d'avance de l'indice d'oscillation australe (SOI_SR3) comme prédicteurs du modèle prédictif. Pour le modèle statistique de courte saison des pluies, la série temporelle actuelle de la SAOD (SAOD_SR0) a été ajoutée sur les deux prédicteurs du modèle prédictif. En appliquant un modèle MLR, il est montré que la prévision peut expliquer 27,4% de la variation totale et a un RMSE de 74,2 mm/saison pour la longue saison des pluies tandis que pour le RARIMAE la prévision explique 53,6% de la variation totale et a un RMSE de 59,4 mm/saison. En appliquant un modèle MLR, il est montré que la prévision peut expliquer 22,8% de la

variation totale et a un RMSE de 106,1 mm/saison pour le modèle prédictif de courte saison des pluies tandis que pour le RARIMAE la prévision explique 55,1% de la variation totale et a un RMSE de 81,1 mm/saison.

À partir de cette comparaison, une augmentation significative de R^2 et une diminution des valeurs RMSE ont été observées dans les modèles RARIMAE pour les séries temporelles moyennes à la fois pour des courtes et des longues saisons de pluie. En termes de fiabilité, RARIMAE a surpassé ses homologues MLR avec une meilleure efficacité et précision. Par conséquent, chaque fois que les données souffrent d'autocorrélation, nous pouvons opter pour le MLR avec erreur ARIMA, la partie erreur ARIMA est davantage destinée à corriger l'autocorrélation, améliorant ainsi la variance et la productivité du modèle.

Table of Contents

List of Tables

List of Figures

List of acronyms

ACF	Auto-Correlation Function
ADF	Augmented Dickey-Fuller
ADW	Angular Distance Weighting
AIC	Akaike Information Criterion
ANFIS	Adaptive Neuro-Fuzzy Inference System
ANN	Artificial Neural network
AR	Autoregressive
ARIMA	Autoregression Integrated Moving Average
ARIMAX	Autoregression Integrated Moving Average with exogenous variables
BIC	Bayesian Information Criterion
BSMDDS	Basic Structural Model with Dummy Seasonality
BVAR	Bayesian Vector Autoregressive
CAM	Climate Anomaly Method
CAR	Central African Republic
CART	Classification and Regression Trees
CDD	Correlation decay distance
CRU	Climate Research Unit
CTM	Cyclical Trend Model
Df	Degree of freedom
DMI	Dipole Mode Index
DRC	Democratic Republic of Congo
ENSO	EL-Nino Southern Oscillation
EOF	Empirical Orthogonal Function
FAO	Food and Agriculture Organization
GARCH	Generalized Autoregressive Conditional Heteroscedastic
GDP	Gross Domestic Product
GMT	Generic Mapping Tools
GPCC	Global Precipitation Climatology Centre
GPCP	Global Precipitation Climatology Project

HadGHCND	Hadley- Global Historical Climatology network -Daily
I/P-O/P	InPut-OutPut
ICOADS	international Comprehensive Oceanic – Atmospheric Dataset
IM	Indian Monsoon Index
IOD	Indian Ocean Dipole
ISMR	Indian Summer Monsoon Rainfall
ITCZ	Inter-Tropical Convergence Zone
JAS	July-August-September
LSSVM	Least Squares Support Vectors Machine
M5P	Multivariate Regression prediction model M5P
MAM	March-Avril-May
MARS	Multivariate Adaptive Regression splines
MJO	Madden-Julian-Oscillation
MLR	Multiple Linear Regression
MPL-ANN	Multilayer Perceptron Artificial Neural Network
MPR	Multi Parameter Regression
NAO	North Atlantic Oscillation
NCAR	National Center for Atmospheric Research
NCEP	National Center for Environmental Prediction
NOAA	National Oceanic and Atmospheric Administration
OLS	Ordinary Least Squares
ON	October-November
OND	October-November-December
PACF	Partial Auto-Correlation Function
PAR	Periodic Autoregressive
PC	Principal Component
PCA	Principal Component Analysis
QBO	Quasi-Biennial-Oscillation
RFI	Radiative Forcing Index
RMSE	Root Mean Square Error

RARIMAE	Regression with ARIMA errors
SAM	Southern Annular Mode
SARIMA	Seasonal Autoregression Integrated Moving Average
SLP	Sea level Pressure
SOND	September-October-November-December
SST	Sea Surface Temperature
STSM	Structural Time Series Models
SVM	Support Vector Machine
TAMSAT	Tropical Applications of Meteorology using SATellite data and ground-based observations.
TRMM	Tropical Rainfall Measuring Mission
VAR	Vector Autoregressive
VIF	Variance Inflation Factor
WA	West Australia

1 INTRODUCTION

Agriculture takes a large share of National Economies throughout East Africa. According to Food and Agriculture Organization (FAO) and World Bank development Indicators, agriculture accounts for 43% of the total Gross Domestic Product (GDP) in the region. In Tanzania and Burundi agriculture share of GDP exceeds 50% and in Uganda and Rwanda it is about 50%. Only in Kenya, it contributes less than 30% because Kenya's structural transformation towards a less agricultural-based economy is more advanced than in other countries in the sub-region.

East Africa region agriculture is highly depending on rainfall, with irrigation agriculture accounting less than 1% of the regions total cultivated land. Thus, the amount and temporal distribution during the growing season are critical to crop yields and can induce food shortages and famine (Di Falco et al.,2012)

East Africa is highly vulnerable to climate variability, as seen by the recent devasting drought happened between 2010 and 2011(Haile et al.,2019). This drought was the worst in decades and struck a severe food crisis across many countries, including Somalia, Sudan, Kenya and Uganda. Similar widespread droughts occurred between 1984 and 1985(Broad and Agrawala,2000), 2005 and 2008(Hastenrath et al.2007,2010), all of which had harmful impacts on water, agriculture, energy, and environment (Funk et al.,2005; Verdin et al.,2005). At the other end of climate variability, floods that took place in 1994, 1997, and 2006 claimed thousands of lives and hundreds of thousands of properties (Birkett *et al.*,1999; Hastenrath et al.,2010). Unfortunately, the risks from future droughts and floods are expected to rise in view of the growing population, expended development of coastal areas and flood plains, unbated urbanization and land use changes, and climate change (Doocy et al.,2013). Together, historical experience and future projections call attention to the need for improved East African preparedness to droughts and floods, a critical component of whichis access to reliable climate forecasts.

The topic of seasonal rainfall variability and forecasting for East Africa is discussed in many studies. Large regions of East Africa exhibit two rainy seasons distinguished in 'short rains' (occurring in boreal autumn in October and November) and 'long rains' (occurring in boreal spring from March to May).

East Africa rains are known to be dominated by varied large-scale forcing, such as the migration of the Inter-Tropical Convergence Zone (ITCZ), effects of abrupt orography (e.g. Ethiopia Highlands, Mountain Kilimanjaro, and the Great Rift Val- ley), and ocean-induced wind systems (from the Atlantic and Indian Oceans). Several excellent studies of the short rains (also known as *Vuli* in Tanzania ,*Deyr* in Somalia and Umuhindo in Rwanda) addressed seasonal forecasting (e.g. Mutai et al.(1998) ; Philippon et al. (2002); Ntale et al. (2003); Hastenrath et al. (2004); Mwale and Gan (2005)), the occurrence of specific extreme events (e.g. Behera et al. (1999); Birkett et al. (1999); Latif et al. (1999); Webster et al. (1999)), and the dynamic relationship with the Indian Ocean (e.g. Hastenrath (2007); Ummenhofer et al.(2009)).

Likewise, several excellent studies of the long rains (also known as *Masika* in Tanzania, *Belg* in Ethiopia, *Itumba* in Rwanda, and *Gu* in Somalia) addressed their general mechanisms and interannual variability (e.g. Nicholson (2019); Camberlin and Philippon (2002); Camberlin and Okoola (2003); Pohl and Camberlin (2006)) and long-term trends (e.g. Williams and Funk (2011) ; Lyon and DeWitt (2012)). A study by Omondi et al. (2013) examined the decadal variability of the short, long, and summer rains and their statistical linkages with sea surface temperatures (SSTs) over the Indian, Atlantic, and Pacific Oceans. Their results indicate that while El Niño–Southern Oscillation **(**ENSO) and the Indian Ocean dipole (IOD) are the prominent modes, all three oceans contribute in explaining a significant portion of East African rainfall variance.

1.1 Research Problem

Most commonly used classical linear time series models are ARIMA and linear regression models. The major drawback of these models is presumed linear form of the model, *i.e*. a linear correlation pattern is assumed among the time series hence, no nonlinear patterns can be modelled by these models. Sometimes the time series often contain both linear and nonlinear components, rarely they are pure linear or nonlinear and under such condition neither ARIMA nor linear regression models are adequate in modelling such series.

To overcome this difficulty Regression with ARIMA errors method were evolved. Applications of Regression with ARIMA errors methods in the literature shows that amalgamating different methods can be an efficient and effective way to improve the forecasts of time series under consideration. With these motivations, attempt have been made to develop

a time series linear regression model and a Regression with ARIMA errors models for forecasting seasonal rainfall in east Africa region. The details methodology is explained in subsequent chapters.

1.2 Research objectives

With above discussed problems and research gaps the following objectives were framed to forecast seasonal rainfall in East Africa:

1. To develop MLRM for seasonal rainfall prediction in East Africa.
2. To develop a RARIMAE model for seasonal rainfall prediction in East Africa.
3. Comparison of model's efficiency under consideration.

1.3 Research Methodology

The Global Climate Research Unit (CRU) are used to investigate the seasonal rainfall situation in the whole part of the region. The mean seasonal data from 1949 up 2000 are used in model's development as predictands whereas 10 teleconnections data including their 5 different times steps i.e (current time series, three months' lead, six months lead, nine months lead and 12 months lead) and 2 teleconnections with a single time series also have been used. Two kinds of models are developed i.e a statistical model and a predictive model with 72 predictors and 58 predictors respectively.

1.3.1 Statistical models

A statistical model is a stochastic model which contains unknown parameters, and these parameters need to be estimated based on assumptions about the model and the data under considerations. The error term in the model carries appropriate assumptions *viz.*, independence and homoscedasticity and the distribution being normal.

1.3.1.1 Linear Model

A linear model is one in which all the parameters appear linearly. Some examples of linear models are multiple linear regression model and polynomial models with one predictor variable

1.3.1.1.1 Multiple Linear Regression model

$$Y = \beta_0 + \beta_1 X_1 + \cdots, + \beta_p X_p + \varepsilon \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (1.1)$$

Where, Y is the dependent (response) variables are independent (predictor or stimulus) variables, $\beta_0, \beta_1, \cdots, \beta_p$ are the regression coefficients and ε is the error term.

1.3.1.1.2 Polynomial models with one predictor variable

$Y=a+bX+\varepsilon$ (First-order model) …… (1.2)

$Y = a + bX + cX^2 + \varepsilon$ (Second order model) … (1.3)

These models are commonly used in many fields' *viz.*, agriculture, climatology, medicine, education, industry, *etc.* The Ordinary Least Square (OLS) is generally employed for parameter estimation.

1.3.1.2 Nonlinear Models

In real world, most of the existing phenomenon are not linear in nature rather they are very complex and in unidentified state. Nonlinear models play a prominent role in comprehending the complex nonlinear inter-relationships among many variables under consideration. Nonlinear models are one in which at least one of the parameters appears in nonlinear form. More precisely, in nonlinear model, at least one derivative with respect to a parameter should include that parameter. Examples of a nonlinear model are:

$$Y(t) = \exp\,(at + bt^2) + \varepsilon \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (1.4)$$

$$Y(t) = at + \exp(-bt) + \varepsilon \ldots\ldots\ldots\ldots\ldots\ldots \quad (1.5)$$

Sometimes the nonlinear models can be transferred into linear model form by using some transformations, such models are called 'intrinsically linear models' (Draper and Smith, 1998).

1.3.1.3 Time series models

In most of the phenomenon including climatology, large amounts of data pertaining to precipitation, Sea Surface Temperature (SST), Sea Level Pressure (SLP), *etc.* are being recorded sequentially over a period. Important properties of such data are the successive observations under considerations are dependent. Much efforts have been made by researchers over many years to develop the efficient forecasting models to improve the prediction accuracy of the models involving time series data.

1.3.1.3.1 Linear time series models

In the context of time series, the function of the dependent variable, where generally the time is linear, the model is called as a linear time series model. In other words, if a function relating to the observed time series phenomenon Y_t and the underlying shocks is linear, it is termed as linear time series model. Some important linear time series models are discussed below:

1.3.1.3.1.1 Autoregressive (AR) model

An observed time series Y_t can be elucidate by linear function of its previous observation Y_{t-1} and some unexplainable random error ε_t. Let us consider equally spaced time series Y_t, Y_{t-1}, Y_{t-2} …, over an equal period of time say t, t_{-1}, t_{-2}, …, then Y_t can be defined as:

$$Y_t = \emptyset_1 Y_{t-1} + \emptyset_2 Y_{t-2} + \cdots + \emptyset_p Y_{t-p} + \varepsilon_t \quad \ldots\ldots\ldots\ldots \text{(1.6)}$$

If we represent the series in Backshift operator format, then it becomes

$$\emptyset(B) = 1 - \emptyset_1(B) - \emptyset_2 B^2 - \cdots - \emptyset_p B^p \quad \ldots\ldots\ldots\ldots \text{(1.7)}$$

Where, B is the backshift $BY_t = Y_{t-1}$ then the AR model can be written as $\emptyset(B)Y_t = \varepsilon_t$.

1.3.1.3.1.2 Moving Average (MA) model

Another important model of great practical utility in the framework of time series is finite moving average model. The MA (*q*) model is defined as:

$$Y_t = \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \cdots - \theta_q \varepsilon_{t-q} \ldots\ldots\ldots\ldots\ldots\ldots \quad (1.8)$$

In terms of backshift operator, the MA model of order q is given as follows:

$$\theta(B) = 1 - \theta_1(B) - \theta_2 B^2 - \cdots - \theta_q B^q \ldots\ldots \quad (1.9)$$

Where *B* is the backshift operator, and the moving average model can be expresses as:

$$Y_t = \theta(B)\varepsilon_t \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (1.10)$$

1.3.1.3.1.3 Autoregressive Moving Average (ARMA) model

To obtain the higher efficiency and greater flexibility in modelling we combine both autoregressive and moving average processes together. These models are called as "mixed models" and are represented as ARMA *(p,q)* models:

$$Y_t = \emptyset_1 Y_{t-1} + \emptyset_2 Y_{t-2} + \cdots + \emptyset_p Y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \cdots - \theta_q \varepsilon_{t-q} \ldots\ldots \quad (1.11)$$

Generally, in Backshift operator it is expressed as follows:

$$\emptyset(B) Y_t = \theta(B)\varepsilon_t \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (1.12)$$

1.3.1.3.1.4 Autoregressive Integrated Moving Average (ARIMA) model

Often most of the time series are non-stationary in nature, to obtain the stationary time series, we need to introduce the differencing term *d.* to make the non-stationary series to stationary series we add the differencing term then the general form of ARMA model becomes ARIMA and are represented as ARIMA *(p,d,q).* The process Y_t is said to follow integrated ARMA model if $\Delta Y_t = (1 - B)^d \varepsilon_t$.

The ARIMA model is expressed as follows:

$$\emptyset(B)(1-B)^d Y_t = \theta(B)\varepsilon_t \quad \ldots\ldots\ldots\ldots\ldots\ldots (1.13)$$

Where, $\varepsilon_t \sim WN\ (0, \sigma^2)$ and WN is the white noise. The Box-Jenkins ARIMA model building consists of three steps *viz.,* identification, estimation, and diagnostic checking.

There are many linear time series models available in literature, some prominent models among them are family of exponential smoothing models, Vector Autoregressive (VAR) model, Bayesian Vector Autoregressive (BVAR) models, Periodic Autoregressive (PAR) models, Structural Time Series Models (STSM), Cyclical Trend Model (CTM), Basic Structural Model with Dummy Seasonality (BSMDS) and *etc.* However, in this study we confined only to ARIMA family models because of their popularity for linear time series and due to its well-known Box-Jenkins model building procedure.

1.3.1.3.2 Nonlinear time series models

Main drawback of both ARIMA and other linear time series models is that the underlying relationship among variables is nonlinear and highly complex and cannot be explained through a linear modelling approach. Modelling and forecasting of data sets has to be carried out by some nonlinear models. From last three decades or so, a new area of "nonlinear time series modelling" has rapidly been developing.

Many studies and findings in literature shows that parametric nonlinear time series models like bilinear time series, doubly stochastic model, Generalized Autoregressive Conditional Heteroscedastic (GARCH) model, mixture autoregressive model and Threshold Autoregressive (TAR) model yields better performance, if underlying data generating process is follow some distribution and normal form. Regression with ARIMA errors (or ARIMAX) is a nonlinear model which combines two powerful statistical models namely, Linear Regression, and ARIMA into a single super-powerful regression model for forecasting time series data.

1.3.1.3.2.1 ARIMAX model

This is when you have at least two time series and you believe that one series is causing another. The X is indicating an exogenous variable or multiple exogenous variables. ARMAX

model is a special case of ARIMAX model of order (p, 0, q). An ARIMAX model (i.e an ARIMA model with an exogenous variable) without constant takes the form

$$y_i = \beta x_i + \sum_{j=1}^{p} \emptyset_j y_{i-j} + \varepsilon_i + \sum_{j=1}^{q} \theta_j \varepsilon_{i-j} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (1.14)$$

This is simply an ARMAX model with extra independent variable (covariant) on the right side of the equation. Using the lag operator, this is equivalent to

$$\emptyset(L) y_i = \beta x_i + \theta(L)\varepsilon_i \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (1.15)$$

Or

$$y_i = \frac{\beta}{\emptyset(l)} x_i + \frac{\theta(L)}{\emptyset(L)} \varepsilon_i \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (1.16)$$

One way to deal with such a model is to reinterpret it as a linear regression plus ARIMA errors.

1.3.1.3.2.2 Regression with ARIMA errors model

Regression with ARIMA errors model is mathematically equivalent to ARIMAX model above. ARIMAX is emphasizing that this model handles exogenous variable but did not say how. Regression with ARIMA errors can be a more difficult word to pick up for beginners because it is not an intuitive name, but it very clearly describes what happens in the actual formula that you have beta coefficients just like an OLS regression and then an error term that is an ARIMA process.

$$y_i = \beta x_i + u_i \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (1.17)$$

Where

$$u_i = \sum_{j=1}^{p} \emptyset_j u_{i-j} + \varepsilon_i + \sum_{j=1}^{q} \theta_j \varepsilon_{i-j} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (1.18)$$

This model is equivalent to

$$y_i = \beta x_i + \frac{\theta(L)}{\emptyset(L)} \varepsilon_i \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (1.19)$$

1.3.2 Computer programs

The whole work of analysing the data was done using computer programs; the essential programs which have been used are R software for data analysis and GMT (The Generic Mapping Tools) for maps production.

1.3.3 Significance of Research

Linear models are not always adequate for the time series that have both linear and non-linear structures. In this context, the regression with ARIMA errors which combines both linear and nonlinear component can be effective and efficient way to improve the forecasting performance of the time series under consideration. Findings of this research is also an important result for the regression with ARIMA errors models studies in the future. The results of the research will go a long way to help the policy makers and farmers who are involved in agricultural sector, natural disaster preparedness and water resources planning in East Africa.

1.3.4 Organisation of the Thesis

This research consists of four chapters. Chapter one addresses the general introduction of the research. It also includes the statement of the problem, general and specific objectives and scope of the study, significance of the study and the organization of the study.

The second chapter reviews the key issues in the existing literature. In brief, this chapter includes the literature reviews of what other researchers have done concerning the topic of the research.

The Chapter three is the methodology part, and it describes the type of data to be used from where (source of data) to be used, sample size, and how it ought to be analysed. Finally, the results are presented and discussed in chapter four. In this chapter a linear regression and a regression with ARIMA errors for statistical model and predictive model are evaluated for seasonal rainfall in East Africa. The last part of this chapter presents the summary of the main findings, the conclusion as well as the recommendation.

2 REVIEW OF LITERATURE

A review of the available literature relevant to the proposed study has been furnished in this chapter with a perspective to overview the various methodologies and procedures employed by the researchers. The region of East Africa is characterized by bimodal rainy seasons. The longer rain falls from March to May (MAM) and shorter rains fall from October to December (OND). Some previous studies related to these two seasons are presented in the section bellow.

2.1 The short rainfall season in East Africa

The short rains, although the first season in most of eastern equatorial Africa, provide the largest contribution to interannual rainfall variability. They also have one of the strongest associations ever demonstrated to global circulation: the correlation between East African rainfall and the surface westerlies over the equatorial Indian Ocean is −0.85 (Hastenrath et al.,1993). This suggests a significant degree of predictability, assuming a fair degree of persistence of circulation parameters. Statistical forecast models for this season were developed by Philippon et al. (2002), Mutai et al. (1998), Ntale et al. (2003), Mwale and Gan (2005) and Hastenrath et *al*. (2004). Batte and Deque (2011) also examined the predictability of this season. They evaluated both deterministic (single model) predictions and probabilistic (multimodel) skill scores.

The main months of the short rains are October and November. It should be noted that few of the studies were confined to these months. While it is well known that rainfall variability is highly coherent within the ON period, it is not clear whether or not the variability is coherent within the longer seasons (September–December or October–December) used by several studies. Camberlin and Philippon (2002) found that the coherence is limited to ON, but Hastenrath et al. (2004) found that for the coastal region the correlation between the ON season and the September–December season is 0.97. However, Hastenrath et al. (1993) show much greater skill in predicting October and November rainfall in this region than December rainfall.

2.2 The long rainfall season in East Africa

The boreal spring is the main rainy season in most of Kenya, Uganda, Somalia, Rwanda and northern Tanzania. This season is termed the masika in Kenya–Uganda, gu in Somalia and

itumba in Rwanda. The northern protrusion of these rains into Ethiopia is locally termed as the belg (or small rains) season.

The most extensive study of the predictability of the boreal spring rains is that of Camberlin and Philippon (2002). They distinguished two geographical regions separately considering Ethiopia and Kenya–Uganda, but predictability was tested only for the latter region. Four February indices, involving several time scales and both atmospheric and oceanic parameters, served as predictors in linear multiple regression and linear discriminant analysis models. The predictors were SST in Niño-1.2, zonal wind over the Congo basin at 1000 mb, geopotential height of the 500-mb surface over the Near East, and the east–west moist static energy gradient between the East African highlands and the Sahel. The models were applied for the period 1951–97 and were evaluated using cross validation. For the multiple regression model, the correlation between the predicted and observed MAM rainfall for the Kenya–Uganda section was 0.66 in the cross-validation mode. The discriminant analysis model correctly classified the seasonal anomalies 70% of the time.

Considering two rainfall seasons, the review of the available literature is categorized under the following sections: Studies related to Regression Techniques and weather-based time series forecasting, Studies related to Autoregressive Integrated Moving Average (ARIMA) model and studies related to the combination of these statistical models

2.3 Regression Techniques and weather-based time series forecasting

Fisher (1925) was the first to tackle the pre-harvest forecasting problem using fifth degree polynomial regression model for modeling rainfall distribution and obtained the rainfall constants. A multiple regression equation was developed using crop yield as dependent variable and rainfall distribution constants as independent variables. It was found that wheat crop yield was significantly affected by rainfall. Davis and Harrell (1942) fitted third degree polynomials to study effect of rainfall and average maximum temperature on corn yield at various locations from the Great Plains to the Atlantic coast. It was found that a systematic change occurs in the pattern of precipitation climate indices relationships from one end of the region to the other.

A model was proposed to estimate rainfall in Esparto using data mining process. Author used monthly rainfall of Senirkent, Uluborlu and E˘girdir station. The relative error of this model was 0.7 (Terzi ,2012).

A forecasting model was proposed for prediction of gold price using linear regression. Author used factors such as inflation, money supply and concluded that MLR perform better than Naïve method of prediction (Ismail, et.al, 2009). MPR technique, an effective way to describe complex nonlinear I/P-O/P relationship for prediction of rainfall and then compared the MPR and MLR technique based on the accuracy (Zaw and Naing (2008)). This described the development of a statistical forecasting method for SMR over Thailand using multiple linear regression and local polynomial-based nonparametric approaches. SST, SLP, wind speed, ENSO, and IOD were chosen as predictors. The experiments indicated that the correlation between observed and forecast rainfall was 0.6 (Nkrintra., et al 2005).

Philippon et al. (2002) used a multiple linear regression model to predict the October–December rainfall in a large sector of East Africa that included inland Kenya, northern Tanzania, plus most of Rwanda, Burundi, and Uganda. Based on September predictors identified from correlations for the 1968–97 period, their model explained 64% of the interannual variance. The predictors included a monsoon index involving the northeast and southwest wind components at 200 and 850 mb, respectively; meridional wind at 200 mb over the south-eastern tip of Africa; and an index of circulation over the western Indian Ocean.

Mutai et al. (1998) found that the JAS global SST pattern is strongly correlated with October–December seasonal rainfall aggregated for a large sector of East Africa extending from 5°N, in Kenya, southward to Malawi at 15°S. They developed a multiple linear regression forecast model based on three rotated EOFs for SSTs in the north western Pacific, the eastern equatorial Pacific (the ENSO signal), and the South Atlantic. The model showed significant forecast skill, with a correlation between predicted and observed of 0.69 for the period 1945 to 1988 for rainfall averaged over the entire region. The strongest predictor was an SST EOF with maximum variance in the northwest Pacific.

Ntale et al. (2003) used canonical correlation analysis to predict standardized seasonal rainfall totals for September–November at 3-month lead time. Predictors included SLP and SST anomaly fields in the Indian and Atlantic Oceans. The strongest association was with SSTs off the Somali and Benguela coasts. Mwale and Gan (2005) continued the work, comparing several methods of predicting standardized seasonal precipitation at 21 stations within a homogeneous region that comprises most of East Africa. Skill was higher with a nonlinear model known as artificial neural network than with the more standard linear canonical correlation model. In the latter case, the percent variance explained at individual stations for

the 11 seasons 1987 and 1997 ranged from 49% to 81%, with root-mean-square error (RMSE) of 0.4–0.75 standardized units. With linear correlation the model explained 6% to 32%, with RMSE of 0.4–1.2.

Hastenrath et al. (2004) conducted several prediction experiments using a linear forecast model and a variable number of predictors, including two experiments with the Southern Oscillation index as the only predictor. The best predictors were zonal temperature and pressure gradients across the equatorial Indian Ocean. A cross validation for 1958–96 based on six predictors, produced a correlation between predicted and observed rainfall of 0.45. However, when the model was tested using separate training and validation periods, correlation in the validation period was much lower. It also appeared that the correlation with individual predictors changed markedly over time.

Diro et al. (2008) and Ntale et al. (2003) also used empirical methods to predict rainfall in the boreal spring over East Africa. Both studies focused on Ethiopia. The latter study applied canonical correlation analysis to predict standardized MAM rainfall totals at a 3-month lead time, using SLP and SST anomaly fields of the Indian Ocean adjacent to East Africa and in the Gulf of Guinea in the Atlantic. Camberlin and Philippon (2002) similarly found strong local influence (the Red and Arabian Seas) on MAM rainfall in Ethiopia.

2.4 Autoregressive Integrated Moving Average (ARIMA) Model

At present, several time-series analyses are used as a statistical method for modelling and developing rainfall forecast models. Among them, the ARIMA technique has become very popular due to its effective forecasting abilities over other conventional methods. Additionally, the ARIMA technique has shown effective results in terms of predicting the variability with better accuracy (Momani and Nail,2009).

The idea of stochastic time series model was generated from the deterministic models in 19th century. Yule in the year 1927 initiated the idea of stochastic time series with the assumption that the time series under consideration are the realization of a random process. This notion of concept leads to a landmark in time series analysis. Researchers such as Yule (1927), Gilbert (1931), Slutzky (1937) and Yaglom (1955) initiated the concept of autoregressive (AR) and moving average models for the first time. Since then many theories and concepts have been developed by many researchers in the area of Autoregressive Moving Average Modeling. The credit for popularization of ARMA models goes to Box and Jenkins

(1970) with their fundamental book in Time Series Analysis *i.e.* Time Series Analysis: Forecasting and Control. There are many studies available in the literature regarding theoretical development and practical utility of ARIMA models, outcome of some related studies is briefed in this section.

Box and Jenkins (1970) integrated the existing knowledge and came up with the book entitled "Time series analysis: Forecasting and Control". This book has had an enormous impact on time series analysis and forecasting. They also developed a coherent, versatile three stage iterative procedure for development of ARIMA model *viz.*, identification, estimation and validation, popularly known as Box – Jenkins approach. They developed the three-stage model building methodology based on transfer function models which is still now the robust procedure for linear time series under consideration.

Newbold and Granger (1974) compared many models and came up with the conclusion that every model has its own advantages and disadvantages. For the time series with less than 30 observations, stepwise regression was found better compared to other models. For the observations between 30 to 50, combination of Holt-Winters and step wise regression was found suitable. For the time series of more than 50 observations the Box-Jenkins approach performed better compared to other methods. Over the years, several studies have considered ARIMA for developing rainfall forecasting models.

Tularam (2010) has used the ARIMA model for rainfall forecasting in Queensland, Australia, where the relationship between rainfall and temperature was investigated. Kumar et al. (1995) investigated climate variability and predictability of Indian summer rainfall using the ARIMA technique.

Otok and Suhartono (2009) developed a rainfall forecast model for Indonesia using the ARIMA method. Weeks and Boughton (1987) have used the ARIMA model for rainfall–runoff prediction, while Han et al. (2010) applied the ARIMA model for drought forecasting. Zhang (2003) also developed a hybrid ARIMA and neural network model for forecasting.

Panga (2021) has tried to develop a Seasonal Autoregressive Integrated Moving Average (SARIMA) Model to analyze long term monthly rainfall data of Dar es Salaam region in Tanzania for the period of 53 years (1961 to 2014). Rainfall observations were discovered to have seasonality and also non-stationarity and hence differencing and seasonal differencing

was used to attain stationarity. Rainfall data were found to have two seasons namely October to December (OND) and March to May (MAM). The analysis exhibited that the seasonal ARIMA model which is satisfactory in describing the monthly rainfall data in Dar es Salaam Tanzania is SARIMA (2, 1, 1)(1, 1, 1)12. The model was then used for predictions of monthly rainfall values from January 2015 to December 2024. The forecasting results showed that monthly rainfall values have a decreasing trend, hence that may be a threat to agriculturists and water managers in the region.

2.5 Novel combination of ARIMA model and multiple linear model

However, studies which have considered ARIMA techniques never included climate indices as predictors to develop rainfall forecasting model in East Africa. Some researchers have successfully employed several different techniques such as adaptive neuro–fuzzy inference system (ANFIS), ANN, M5P Model Tree, multivariate adaptive regression splines (MARS), least squares support vector machine (LSSVM), classification and regression trees (CART) model for rainfall/streamflow forecasting in different parts of the world (Choubin, et al., 2014; Choubin, et al., 2016; Choubin, et. al., 2017; Choubin, et al., 2018; Kisi, et al., 2019)..

However, the effective independent variable(s) are unlikely to be the same for all regions, i.e., some climatic variable(s), which are effective for one part of the world are not necessarily to be effective for other parts. Additionally, a single technique may not produce the best results for the entire world. As such, it is necessary to investigate different techniques for a region, while focusing on the stakeholders' needs. To satisfy such a requirement, a simple ARIMAX model was developed to predict autumn rainfall in WA and its prediction performance was compared with previously developed multiple linear regression (MLR) models for the same region. ARIMAX model has been selected due to its superiority in terms of prediction performance over ARIMA and other models (Chadsuthi et al.,2012; Fan et al.,2009; Ling et al.,2019; Peter and Silvia, 2012).

A study conducted by Jalalkamali., et al. (2015) reported that forecasting using ARIMAX is possible with 9 months lagged period whereas the performance has been as outstanding if compared to multilayer perceptron artificial neural network (MLP–ANN), support vector machine (SVM) models, and adaptive neuro–fuzzy inference systems (ANFIS) models. Considering such facts, the ARIMAX model could produce much necessary flexibility required to meet the stakeholders' needs.

An ARIMAX model adds in the covariate on the right-hand side as follows:

$$Y_t = \beta X_t + \emptyset_1 Y_{t-1} + \cdots + \emptyset_p Y_{t-p} - \theta_1 Z_{t-1} - \cdots - \theta_q Z_{t-q} + Z_t \ldots\ldots\ldots (2.1)$$

Where X_t is a covariate at time t and β is its coefficient. While this looks straight-forward, one disadvantage is that the covariate coefficient is hard to interpret. The value of β is not the effect on Y_t when the X_t is increased by one (as it is in regression). The presence of the lagged values of the response variable on the right-hand side of the equation mean that β can only be interpreted conditional on the value of previous values of the response variable, which is hardly intuitive.

If we write the model using backshift operators, the ARIMAX model is given by

$$\emptyset(B)Y_t = \beta X_t + \theta Z_t \ldots\ldots\ldots\ldots\ldots\ldots (2.2)$$

Or

$$Y_t = \beta\emptyset(B)X_t + \theta(B)\emptyset(B)Z_t \ldots\ldots\ldots\ldots\ldots\ldots\ldots (2.3)$$

Where $\emptyset(B) = 1 - \emptyset_1 B - \cdots - \emptyset_p B_p$ and $\theta(B) = 1 - \theta_1 B - \cdots \theta_q B_q$

Notice how the AR coefficients get mixed up with both the covariates and the errors term.

Van den Bossche et al. (2004) developed models to explain and forecast the frequency and severity of accidents in Belgium. The objective of his study was to enhance the understanding of the developments in road safety by studying the impact of various explanatory variables on traffic safety. It is investigated whether the number of accidents and victims is influenced by weather conditions, economic conditions and policy regulations. The model is used to predict the frequency and severity of accidents for a 12-months out-of-sample data set. Using a regression model with ARIMA errors, the impact of variables on aggregate traffic safety is quantified and at the same time the influence of unknown factors is captured by the error term. The results show a significant effect of weather conditions and laws and regulations on traffic safety, but there seems to be negligible statistical impact of economic conditions. The model can easily be used to forecast traffic safety, as can be seen from the reasonably good fit obtained on a 95% confidence level.

3 MATERIALS AND METHODS

The present chapter consists of materials used and the methodology adopted for forecasting. The chapter is divided into the following sections:

3.1 Description of the Study Area

3.2 Sources and description of data

3.3 Statistical methodologies employed

3.4 Comparison of forecasting ability of different statistical techniques

3.1 Description of the Study Area

3.1.1 Area of study

The study was carried over the Central-East Africa. This area is divided into two parts. The first part span over a longitude of 15°E to 55°E and a latitude of 15°S to 15°N. The selected area comprises 20 countries: Angola, Burundi, Cameroon (East), Congo-Brazzaville (East), Djibouti, Eritrea (South), Ethiopia, Kenya, Madagascar (North), Malawi (North), Mozambique (North), Rwanda, D.R.C, Central African Republic (CAR), Somalia, Sudan (South), Chad, Uganda, Tanzania, Zambia (North). Rainfall in these countries has similarities in magnitude and duration because their climate is controlled in regard to the tropical climate types which relates to the position of the Inter Tropical Convergent Zone (ITCZ). This area of study counts 3569 geographical grid points in which the seasonal rainfall patterns are similar. The amount of rainfall with respect to the seasons is different considering the geographical position of each point.

The second part of this study covers a longitude of 31.5°E to 41°E and a latitude of 3.5°S to 0.5°S. This is the regions of East Africa which exhibits two rainy seasons distinguished in "short rains" (occurring in boreal autumn in October, November and December) and "long rains" (occurring in boreal spring from March to May). To perform residual analysis for the linear regression model and for Regression with ARIMA errors model, a single time series was constructed by averaging monthly precipitation in selected grid box.

3.1.2 The systems that impact the distribution of rainfall over East Africa

The two seasons, namely MAM and OND coincide with the double passage of the ITCZ, which lags behind the overhead sun by 3-4 weeks over the region. They also coincide with the transition between the northeast and southeast monsoon circulations. The OND season is a transition period from the southeast monsoon to the northeast monsoon and vice versa for the MAM season. The transition period is associated with convergence along which the ITCZ propagates. The ITCZ can be associated with a quasi-continuous belt of unsettled, often rainy weather (Folland et al. 1991). The convergence of these flows creates strong upward motion that causes rainfall if sufficient moisture is available.

Even though the OND and MAM periods are considered transition periods, Nicholson (2019) described the air streams which govern the region's climate as the Congo air with westerly and southwesterly flow, northeast monsoon and the southeast monsoon. Both monsoons are thermally stable and associated with subsiding air. The Congo air is humid, convergent, and thermally unstable and generally associated with high amounts of rainfall. These air streams are separated by two surface convergent zones, the ITCZ and the Congo Air Boundary; the former separates the two monsoons, the latter, the easterlies and westerlies. Normally, the passage of ITCZ leads the onset of the two rainy seasons by 3-4 weeks, but this may be modulated from season to season by the interactions between the ITCZ and perturbations in the global climate circulation, as well as with changes in the local circulation systems initiated by land surface heterogeneity induced by variable vegetation characteristics, large inland lakes and topography.

The inter-annual variability of the East African climate is linked to perturbations in the global SSTs, especially over the equatorial Pacific and Indian Ocean basins, and to some extent, the Atlantic Ocean (Mutai and Ward, 2000; Indeje et al., 2000; Saji et al., 1999) among others. ENSO anomaly patterns play a dominant influence on the interannual variability of the region. The zonal temperature gradient over the equatorial Indian Ocean, often referred to as IOD Mode (Saji et al., 1999) and the coupled IOD-ENSO influence have also been linked to some of the wettest periods in the region, such as 1961, 1997 and 2006 (Black et al., 2003, Bowden and Semazzi, 2007; Owiti et al., 2008). This part of the region has a classical annual cycle of regions in the vicinity of equator, with two peaks in MAM and OND coinciding with the passage of the ITCZ.

The two major seasons described above, are largely controlled by the location and intensity of anticyclones such as St.Helena, Mascarenes, Azores and Siberian (Ilunga et al.2004; Anyah and Semazzi 2007;Kizza et al.2009). Rainfall in the region generally occurs during the rain's seasons (MAM and SOND) as the ITCZ shifts to the equator from North to South, and Vice-versa (Mutai and Ward 2000).

The ITCZ is the most system controlling the rainfall season over the east Africa region. The subtropical anticyclones are regions of high pressure, which form the sources of the winds. They act as pumps of moisture into the areas of convergence. Their location and intensity influence the seasonal rainfall performance in the East African countries. The subtropical anticyclones with important effect on the climate of the country include Azores (situated Northern Atlantic Ocean), St. Helana (situated Southern Atlantic Ocean), Mascarene (Situated in the Southern Indian Ocean) and the Arabian high-pressure ridge (situated in the Arabian Sea).

The Mascarene high pressure is a major pump of moisture into the region. It is at its strongest during the Southern Winter (June-August) when it is associated with the East African high-pressure ridge, which render the wind flow over Eastern Africa mainly diffluent at low levels. The Arabia ridge is fully developed during southern summer in the period of December-February, it is mainly associated with the diffluent flow over the region creating mainly short dry period in the region and little rainfall in some areas due to its topographic features. The maritime location is favourable rainfall occurrence. The St.Helena high pressure is an important pump of humidity into the area from the Congo air basin. The Congo basin is an important of moisture for the region, which bring significant rainfall during March and May when the subtropical Anticyclones in the southern hemisphere are fully developed. The Azores high pressure is useful in the enhancement of the convergence in the region.

The tropical cyclone affecting the region form in the Arabian and southwestern India. They form in the Arabian sea region during the period March to May and in the southern India during the period December-February. The tropical cyclone days over Indian ocean contains prominent decadal cycles, higher frequencies linked to Quasi Biennal Oscillation (QBO) and had positive relationships with SSTs over the entire southwest India Ocean from September to March (Jury et al.1999) suggested that the possibility of association between the occurrence of tropical cyclones and Madden Julian Oscillation (MJO). The MJO has a strong impact in the development of the tropical cyclone activities. The effects of Tropical cyclones on weather and

climate of the region depend on time of the year, location of the cyclones and the associated large-scale flow (Anyamba 1984,1993). The cyclones that move to the Mozambique channel can have adverse effects on the weather and climate of the region in March-May season which induce low level diffluent flow in the region (Anyamba,1993). However, the cyclones in the Mozambique channel during the months of December and January tend to enhance rainfall and are often associated with floods affecting the region during the period. They are characterized with the increase in pressure gradient between North Africa and Atlantic Ocean and Southwest India Ocean resulting to moist westerlies convergence over the region. It can therefore, be concluded that the effect of the tropical cyclones on region rainfall depends on the season, track and location of the cyclones.

ENSO has important effect on precipitation over the region (Indeje et al.,2000). El Nino is linked with improved rainfall over the region especially in September to December (SOND) season. La Nina is linked with scarce rainfall over some parts of the region. It also influences or impacts the onset, cessation and the peacks of seasonal rainfall (Indeje, 2000).

The IOD is the modes that have been observed to have important impact on rainfall over the region and other areas neighbouring the Indian Ocean (Owiti ,2005).

SSTs of the global oceans are the most frequently used predictors of seasonal rainfall. Various effort has been made to determine useful relationships between SST and seasonal rainfall over the region and other parts of the tropics that could use to predict rainfall during the season (Nyakwanda,2003).

Enhanced/depressed seasonal rainfall over a region has been linked with the warming and cooling aver the western Indian Ocean (Owiti, 2005) also the cooling over the Eastern Indian Ocean observed that wet/dry seasons over the region were closely associated with distinct anomalously warm/cool SSTs over parts of the western Indian/Eastern Atlantic Oceans. The SST based models have been observed to give climate outlooks for the region with useful skills. The skills of the forecasts are, however, influenced by the statistics of the weather within the season, which are dependent internal chaotic variations (Zebiak, 2003).

3.2 Sources and description of data

The monthly precipitation data covering the period from 1949 to 2000 was obtained from Climate Research Unit (CRU). Next to that, the first differenced climate indices such as Southern Oscillation Index (SOI), Multivariate ENSO Index (MEI), Nino3.4, Dipole Mode Index (DMI), Indian Summer Monsoon Index (ISMI), Indian Monsoon Rainfall (IMR), The South Atlantic Ocean Dipole (SAOD), North Atlantic Oscillation (NAO), the Southern Annular Mode(SAM), Quasi-biennial Oscillation (QBO), Indian ocean Sea Surface Temperature (SST), Indian ocean Sea Level Pressure (SLP) and their respective lead times (i.e three months lead, six months lead, nine months lead and 12 months lead) are considered as potential predictors.

3.2.1 Observational climate data

Observational climate data are needed to describe climate patterns, asses the performance of climate models and calibrate impact models in present day. The data are obtained from either land-based network of meteorological stations (e.g HAdGHCND, GPCC, and CRU datasets) or meteorological observation satellites (e.g TRMM 3B 42 and TAMSTAT datasets). Some datasets are also derived from the combination of gauge measurements and the satellite products (e.g operational RFE 2.0, climatological RFE, and GPCP). For regional applications, the stations records are spatially interpolated to generate gridded datasets$(X(t,s)$ where X stand for variable of interest (e.g precipitation, temperature, etc.) whereas t and s stand for the temporal and spatial resolutions respectively.

The different datasets come from various international data centers. Each of these datasets covers different time periods at different spatial and temporal resolutions. The fidelity of all these datasets in representing the real African climate is questionable (Paeth at al.,2005). Indeed, uncertainty is inherent in all the observation products, especially in data sparse areas (Sylla et al.(2012); Gbobaniyi et al.(2014)). As pointed out by Pinker et al. (2006), the satellite estimates generally overestimate precipitation over semi-arid regions of the African continent. This is certainly due to the fact that algorithms translating measured radiation to effective rainfall amount are still subject to some uncertainties (Paeth et al.,2005).

Nikulin et al. (2012) have compared Tropical Rainfall Measuring Mission (TRMM) data to Global Precipitation Climatology Project (GPCP) satellites-gauge combination data

(Adler et al.2003) and found that TRMM exhibits significant dry bias up to 50% over some regions in tropical Africa. Many other scholars have also shown that combined satellite-gauge information often outperform the current satellite only products (e,g Nicholson et al.(2003a,b); Dinku et al.(2007); Paeth et al.(2011b);Parker et al.(2011)). But these combined products also do not lead to much added-value when compared to some gauge-datasets like Global Precipitation Climatology Centre (GPCC), especially in the Western Africa (Nicholson et al.(2003a); Nicholson et al.(2003b); Ali et al. (2005)).

In general, some authors have suggested that the discrepancy between all the datasets might be due to the facts that: (1) they use different gauge analysis products (Huffman et al.(2009); Nikulin et al.(2012)), (2) the number of observations used in the products varies over time and regions (Sylla et al.,2012), and (3) different retrieval, merging and interpolation techniques are applied (sylla et al.(2012); Panitz et al.(2013)). As pointed out by Zhang et al. (2012) and Panitz et al. (2013) in station errors are particularly relevant in areas where almost no gauge data are available (e,g central Africa) because of their large spatial influence. However, good agreement between GPCC, GPCP, and CRU datasets have generally been reported except in areas like Angola and the Democratic Republic of the Congo where the number of gauge stations is very limited (Zhang et al. (2012); Panitz et al. (2013); Harris et al.(2014a)). Gbobaniyi et al. (2014) have also confirmed that GPCP and CRU agree well in representing the inter-annual variability in the Sahel with a high correlation coefficient of around 0.96 but reported relatively low correlation (0.63) over the Gulf of Guinea. Nevertheless, they have all concluded that one or the other dataset as reference does not change the conclusion.

In this study, the CRU time-series was selected for impact application due its high spatial resolution, long temporal coverage, and large spatial coverage. This dataset has also advantage of being freely available and consistent over time for one of the climate variables of interest in this study (i.e total precipitation). Herein, the CRU dataset serves as reference for representing the pattern, mean and trend of the present-day climate. The CRU dataset has already been extensively analysed by Brohan et al. (2006). It has also been intensively used for different research purposes; for example, climate models assessment (Paeth et al. (2005); Trenbert et al. (2007); Paeth (2011); Jacob et al. (2012); Nikulin et al. (2012); Kim et al. (2013)) and human disease transmission (Gaardbo Kuhn et al. (2002); Ermert et al. (2012)) and is thus regarded as suitable for this study.

In this study, CRU TS 4.01 for 1901-2016 was used because it was the most recent at the start of the analysis. CRU TS 4.01 (herein referred to as CRU) is publicly on the Climatic Research Unit portal (https://crudata.uea.ac.uk/cru/data/hrg/cru_ts_4.01/) on a regular high spatial resolution (0.5°grid) and represents century-long time series. Indeed, it is a monthly time series of various climate variables including precipitation, potential evapotranspiration and air temperature developed by the Climatic Research Unit of University of East Anglia in Norwich, UK (Mitchell and Jones 2005; Harris et al 2014a). The main process change in version 4 is the move to Angular Distance Weighting (ADW) for gridding the monthly anomalies. Compared to the previous approach, which used IDL routines TRIANGULATE and TRIGRID to effect triangulated linear interpolation, ADW allows us total control over how station observations are selected for gridding, and complete traceability for every datum in the output files. For secondary variables, this means that observed and synthesised data values are used in the same way in the gridding process.

The monthly database is built from in situ meteorological stations from around the world and spans the period from 1901 to 2016 at a spatial resolution of $0.5^{\circ}X5^{\circ}$ latitude/longitude over all land masses. The CRU dataset is constructed using Climate Anomaly Method (CAM) developed by (Peterson et al 1998). Only the stations with at least 75% of non-missing values in each month through the reference period (1949-2000) were included in the gridding operation. Those stations values were used to compute monthly climatology per station provided that they have fallen within the range of 3 times (4 times for precipitation) standard deviation departure from the normal. For the stations that passed the screening, the time series were converted to monthly anomalies relative to their average on the reference period. Depending on the station locations, the anomaly values were further on interpolated to a half degree grid cell resolution through triangulated linear interpolation. First, for each variable a correlation decay distance (CDD) (New et al 2000) was defined to determine the stations to be considered to infill each land grid cell. The monthly anomalies were then passed to the gridding routines only if a least one station falls in a land grid cell within the CDD. In case no station falls in a given land grid cell, the empty cell is given 0 as anomaly value. These yields 0.5°regular gridded anomalies for all global land areas. The gridded anomalies were finally converted to absolute values (construction of the time series) by combining them the monthly gridded reference climatology (New et al 1999) used in the earlier versions of the CRU TS dataset (cf. Harris et al 2014 for detailed description of the dataset).

3.2.2 Predictors (Global scare circulation variables)

The Atlantic and Indian oceans are major sources of moisture for the East African region. The oceans do not influence the regional climate independently but in some integrated manner through the interactions associated with the oceanic and atmospheric circulations (Wolter 1987). ENSO and Walker circulation (Chervin and Druyan,1984), and the Great Ocean Conveyor (GOC) (Gross (1972); Saenko et al. (2002)) are some examples of the atmospheric and oceanic processes that may be associated with the combined influence of the global oceans on global climate.

The low-level circulation patterns associated with the above-normal rainfall over the region is dominated by easterly inflow from the Indian Ocean and westerly inflow from the Congo tropical rain forest into the positive rainfall region (Anyah and Semazzi (2006); Schreck and Semazzi (2004)). Goddard and Graham (1999) observed significant influence of the Indian Ocean on seasonal rainfall over the region. Okoola (1996) observed that the cooling over the eastern Atlantic Ocean together with the warming over the Indian Ocean are associated with enhanced rainfall over the region.

3.2.2.1 Extended Reconstructed Sea Surface Temperature (ERSST) v4

In this study the monthly National Oceanic and Atmospheric Administration Extended Reconstructed Sea Surface Temperature (NOAA ERSST) v4 dataset (Huang et al.,2014; Liu et al.2014) has been used. Itis a global monthly sea surface temperature dataset derived from the International Comprehensive Ocean–Atmosphere Dataset (ICOADS). It is produced on a $2^{\circ} \times 2^{\circ}$ grid with spatial completeness enhanced using statistical methods. This monthly analysis begins in January 1854 continuing to the present and includes anomalies computed with respect to a 1971–2000 monthly climatology. The newest version of ERSST, version 4, is based on optimally tuned parameters using the latest datasets and improved analysis methods.

SST anomalies in the western Indian Ocean exert a strong influence on the equatorial East African short rains than central and eastern Indian Ocean SST anomalies both in terms of the coverage of significantly changed precipitation and the magnitude of precipitation response.

3.2.2.2 Indian Ocean Dipole (IOD)

Intensity of the IOD is represented by anomalous SST gradient between the western equatorial Indian Ocean (50°E-70°E and 10°S-10°N) and the south eastern equatorial Indian Ocean (90°E-110°E and 10°S-0°N). This gradient is named as Dipole Mode Index (DMI). When the DMI is positive then, the phenomenon is refereed as the positive IOD and when it is negative, it is refereed as negative IOD.

Positive western Indian Ocean SST anomalies significantly increases the short rains over 95% of the equatorial East African domain ($30^\circ - 40^\circ\text{E}, 5^\circ\text{S} - 5^\circ\text{N}$), while only 30% of the region responds to central and eastern Indian Ocean SST anomalies. This was approved by Ummenhofer et al. (2009) in his study on the relationship of October-November rainfall over ($31^\circ - 45^\circ\text{E}, 1^\circ\text{S} - 10^\circ\text{N}$) and IOD using ensemble simulations with an atmospheric general circulation model (GCM). They asses the contributions of individual (and combined) poles of the IOD to above-average precipitation over East African region. They show that increased East African short rains during positive IOD are driven mainly by warming over the western Indian Ocean ($38^\circ - 70^\circ\text{E}, 12^\circ\text{S} - 12^\circ\text{N}$), leading to a reduction in sea level pressure over the western half of the Indian Ocean. Converging wind anomalies over East Africa lead to the moisture convergence and increased convective activity. IOD data used in this study is downloaded from the following link: https://psl.noaa.gov/gcos_wgsp/Timeseries/DMI/

3.2.2.3 NINO3.4

There are several indices used to monitor the tropical Pacific, all of which are based on SST anomalies averaged across a given region. Usually the anomalies are computed relative to a base period of 30 years. The Niño 3.4 index and the Oceanic Niño Index (ONI) are the most commonly used indices to define El Niño and La Niña events.

The Niño 3.4 (5N-5S, 170W-120W) anomalies may be thought of as representing the average equatorial SSTs across the Pacific from about the dateline to the South American coast. The Niño 3.4 index typically uses a 5-month running mean, and El Niño or La Niña events are defined when the Niño 3.4 SSTs exceed $\pm 0.4°\text{C}$ for a period of six months or more. The Niño 3.4 index are downloaded from the following link: https://psl.noaa.gov/gcos_wgsp/Timeseries/Data/nino34.long.anom.data

3.2.2.4 Southern Oscillation Index (SOI)

The SOI is defined as the normalized pressure difference between Tahiti and Darwin. There are several slight variations in the SOI values calculated at various centres. Here we calculate the SOI based on the method given by Ropelewski and Jones (1987). It uses a second normalization step and was the Climate Analysis Centre's standard method in 1987. The reader is also referred to Allan et al. (1991) and Können et al. (1998) for details of the early pressure sources and methods used to compile the series from 1866 onwards.
The SOI index used in this study are downloaded from the following link: https://crudata.uea.ac.uk/cru/data/soi/soi_3dp.dat

3.2.2.5 Multivariate El Niño Index (MEI)

El Niño/Southern Oscillation (ENSO) is the most important coupled ocean-atmosphere phenomenon to cause global climate variability on interannual time scales. The monitoring of ENSO by the Multivariate ENSO Index (MEI) is based on the six main observed variables over the tropical Pacific. These six variables are: sea-level pressure (P), zonal (U) and meridional (V) components of the surface wind, sea surface temperature (S), surface air temperature (A), and total cloudiness fraction of the sky (C). These observations have been collected and published in ICOADS for many years.

The MEI is computed separately for each of twelve sliding bi-monthly seasons (Dec/Jan, Jan/,..., Nov/Dec). After spatially filtering the individual fields into clusters (Wolter,1987), the MEI is calculated as the first unrotated Principal Component (PC) of all six observed fields combined. This is accomplished by normalizing the total variance of each field first, and then performing the extraction of the first PC on the co-variance matrix of the combined fields (Wolter and Timlin, 1993). In order to keep the MEI comparable, all seasonal values are standardized with respect to each season and to the 1950-93 reference period.
The MEI index are downloaded from the following link: https://www.psl.noaa.gov/enso/mei.ext/table.ext.html

3.2.2.6 The South Atlantic Ocean Dipole (SAOD)

The South Atlantic Ocean Dipole (SAOD) is the mechanism of warming of the surface waters off the coasts of West/Central Equatorial Africa associated with concurrent cooling of similar magnitude off the Argentina-Uruguay-Brazil coasts. These SST patterns are coupled to the atmospheric circulation field and regional climates.

A simple measure of the dipole, the SAOD Index (SAODI) is defined by differencing the domain-averaged normalized SST anomaly (SSTA) of the two centers of intense warming and cooling associated with the SAOD, viz:

$$SAODI=[SSTA]_{NEP}-[SSTA]_{SWP}$$

where the square brackets indicate domain averages, the subscripts show the two regions over which the SSTA averages are computed. These domains are described by their locations in the South Atlantic Ocean as the northeast pole (NEP: 10°E–20°W, 0° - 15ºS) and the southwest pole (SWP: 10º–40ºW, 25ºS - 40ºS). As shown below, this index is closely reproduced by the time series of the SAOD-mode determined by the singular value decomposition of the South Atlantic Ocean SST and mean sea level pressure. The data are downloaded from the following link: http://ljp.gcess.cn/dct/page/65592/

3.2.2.7 Quasi-biennial Oscillation (QBO)

The quasi-biennial oscillation (QBO) is a quasiperiodic oscillation of the equatorial zonal wind between easterlies and westerlies in the tropical stratosphere with a mean period of 28 to 29 months. The alternating wind regimes develop at the top of the lower stratosphere and propagate downwards at about 1 km (0.6 mi) per month until they are dissipated at the tropical tropopause. Downward motion of the easterlies is usually more irregular than that of the westerlies. The amplitude of the easterly phase is about twice as strong as that of the westerly phase. At the top of the vertical QBO domain, easterlies dominate, while at the bottom, westerlies are more likely to be found. At the 30mb level, with regards to monthly mean zonal winds, the strongest recorded easterly was 29.55 m/s in November 2005, while the strongest recorded westerly was only 15.62 m/s in June 1995. The QBO data used in this study are downloaded from the following link: https://psl.noaa.gov/data/correlation/qbo.data

3.2.2.8 North Atlantic Oscillation (NAO)

The North Atlantic Oscillation (NAO) is a weather phenomenon in the North Atlantic Ocean of fluctuations in the difference of atmospheric pressure at sea level (SLP) between the Iceland Low and the Azores High. Through fluctuations in the strength of the Icelandic low and the Azores high, it controls the strength and direction of westerly winds and location of storm tracks across the North Atlantic (Hurrell., et al.2003).

The NAO was discovered through several studies in the late 19th and early 20th centuries (Stephenson, et al.2003). Unlike the ENSO phenomenon in the Pacific Ocean, the NAO is largely an atmospheric mode. It is one of the most important manifestations of climate fluctuations in the North Atlantic and surrounding humid climates (Hurrell,1995). The North Atlantic Oscillation is closely related to the Arctic oscillation (AO) (or Northern Annular Mode (NAM)) but should not be confused with the Atlantic Multidecadal Oscillation (AMO). The NAO data are downloaded from the following link: https://psl.noaa.gov/gcos_wgsp/Timeseries/Data/nao.long.data

3.2.2.9 The Southern Annular Mode (SAM)

The Southern Annular Mode (SAM), also known as the Antarctic Oscillation (AAO), describes the north–south movement of the westerly wind belt that circles Antarctica, dominating the middle to higher latitudes of the southern hemisphere.

The changing position of the westerly wind belt influences the strength and position of cold fronts and mid-latitude storm systems and is an important driver of rainfall variability in southern Australia. In a positive SAM event, the belt of strong westerly winds contracts towards Antarctica. This results in weaker than normal westerly winds and higher pressures over southern Australia, restricting the penetration of cold fronts inland.

Conversely, a negative SAM event reflects an expansion of the belt of strong westerly winds towards the equator. This shift in the westerly winds results in more (or stronger) storms and low-pressure systems over southern Australia. During autumn and winter, a positive SAM value can mean cold fronts and storms are farther south, and hence southern Australia generally misses out on rainfall. However, in spring and summer, a strong positive SAM can mean that southern Australia is influenced by the northern half of high-pressure systems, and hence there are more easterly winds bringing moist air from the Tasman Sea. This increased moisture can turn to rain as the winds hit the coast and the Great Dividing Range. In recent years, a high positive SAM has dominated during autumn–winter and has been a significant contributor to the 'big dry' observed in southern Australia from 1997 to 2010. The SAM data are downloaded from the following link: https://psl.noaa.gov/data/20thC_Rean/timeseries/monthly/SAM/sam.20crv2c.long.data

3.2.2.10 Monthly mean Sea Level Pressure from the NCEP Reanalysis

The NCEP/NCAR Reanalysis project is using a state-of-the-art analysis/forecast system to perform data assimilation using past data from 1948 to the present. A subset of this data has been processed to create monthly means of a subset of the original data. There are also files containing data from variables derived from the reanalysis and some other statistics. The seal level pressure data are downloaded from the following link: ftp://ftp.cdc.noaa.gov/Datasets/ncep.reanalysis.derived/surface/slp.mon.mean.nc

3.2.2.11 Indian monsoon Index (IM)and Indian summer monsoon rainfall (ISMR)

From a circulation perspective, the monsoon is dominated by the lowest baroclinic mode, which is stimulated by the latent heat released in the middle troposphere. The vertical shears defined by the difference of 850- and 200-hPa zonal winds, U850−U200, provided a first-order approximation to the strength of the gravest baroclinic mode. The large zonal (westerly) vertical shears in pressure coordinates, denoted by WS (westerly shear), extend along 10°N from Africa to the western North Pacific with a maximum of 36 m s−1 (10°N, 60°E). The WYI defined by the WSs averaged in the area (0°–20°N, and 200-hPa winds display maximum intensity (termed action centers). The data are downloaded from the following link: http://apdrc.soest.hawaii.edu/projects/monsoon/ismidx/ismidx-jjas.txt/

3.3 Statistical tools employed

Time series forecasting models are very useful techniques for forecasting weather time series phenomenon. Generally, whether time series data contains both linear and nonlinear structures hence, no single model is capable to capture both linear and nonlinear pattern present in the data. Consequently, various types of linear and nonlinear parametric time series models are used for forecasting, like time series linear regression models, ARIMA models (Box and Jenkins 1970), regression with ARIMA errors models and finally the methods for comparing different models. The time series forecasting models employed in the present study are described as below:

3.3.1 Regression Analysis

In regression analysis procedure the impact of weather parameters on seasonal rainfall prediction in East Africa was assessed by relating the weather parameters on seasonal rainfall.

3.3.1.1 Multiple Linear Regression Analysis

The Multiple Linear Regression (MLR) models are applied when two or more independent variables are influencing the dependent variable. The dependent variable is also called the response variable and independent variables are named as Predictors. Three assumption check is necessary before conducting linear regression analysis: linearity, equal of variance, and normality. In a linearity check, the linear relationship between the dependent and independent variables is verified. In equal of variance, the spread of the residuals is checked and in normality check, data distribution is sought whether it is normally distributed or not. If selected variables satisfy all these assumption checks, linear regression analysis can proceed further. The equation for the MLR model is given below:

$$Y = b_{0+}b_1X_1 + b_2X_2 + \cdots, b_nX_n + e_t \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (3.1)$$

Where, b_0 is the intercept, $b_{1,}b_{2,} \ldots, b_n$ are the regression coefficients representing the contribution of explanatory variables $X_{1,}\ X_2\ \ldots, X_n$ on the dependent variable Y and error e_t at time *t* is *i.i.d.* with zero mean and finite variance (Drapper and Smith 1966).

3.3.1.2 Least square estimation

In practice, of course, we have a collection of observations, but we do not know the values of the coefficients $b_1, b_2, \ldots, b_n$. These need to be estimated from the data.

The least squares principle provides a way of choosing the coefficients effectively by minimising the sum of the squared errors. That is, we choose the values of $b_1, b_2, \ldots, b_n$ that minimise

$$\sum_{t=1}^{T} e_t^2 = \sum_{t=1}^{T}(Y_t - b_0 - b_1X_{1,t} - b_2X_{2,t} - \cdots, -b_nX_{n,t})^2 \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (2.2)$$

This is called least squares estimation because it gives the least value for the sum of squared errors. Finding the best estimates of the coefficients is often called “fitting” the model to the data, or sometimes “learning” or “training” the model.

3.3.1.3 Variable’s selection Technics

When there are many possible predictors, we need some strategy for selecting the best predictors to use in a regression model.

A common approach that is not recommended is to plot the forecast variable against a particular predictor and if there is no noticeable relationship, drop that predictor from the model. This is invalid because it is not always possible to see the relationship from a scatterplot, especially when the effects of other predictors have not been accounted for.

Another common approach which is also invalid is to do a multiple linear regression on all the predictors and disregard all variables whose p-values are greater than 0.05. To start with, statistical significance does not always indicate predictive value. Even if forecasting is not the goal, this is not a good strategy because the p-values can be misleading when two or more predictors are correlated with each other. The methods for selection are described below:

3.3.1.4 Correlation Analysis

The relationship between two variables between x and y is represented by the correlation coefficient r_{xy} of Pearson (Pearson Product-moment coefficient of linear correlation). Correlation analysis is one of the important step during variables sections for checking the influences of available predictors on the predictors. The calculation of r_{xy} is based on the ratio of the covariance of the two variables and the standards deviation (Willks,2006):

$$\mathrm{r_{xy}} = \frac{\mathrm{cov(x,y)}}{\mathrm{s_x * s_y}} = \frac{\frac{1}{\mathrm{n-1}}\sum_{\mathrm{i=1}}^{\mathrm{n}}(\mathrm{x_i}-\bar{\mathrm{x}})(\mathrm{y_i}-\bar{\mathrm{y}})}{\sqrt{\frac{1}{\mathrm{n-1}}\sum_{\mathrm{i=1}}^{\mathrm{n}}(\mathrm{x_i}-\bar{\mathrm{x}})^2} * \sqrt{\frac{1}{\mathrm{n-1}}\sum_{\mathrm{i=1}}^{\mathrm{n}}(\mathrm{y_i}-\bar{\mathrm{y}})^2}} \quad \text{..... (3.3)}$$

where $-1 \leq r_{xy} \leq 1$

If r_{xy} is positive (negative) the result is a concordant (inverse) with an increasing (decreasing) for y if x increases and vice versa. If r_{xy} is $|1|$ the linear relationship is a perfect one, at a value of 0 there is no relationship.

After calculating correlation coefficients r_{xy} between two variables, testing for significance is necessary. For this one can look up critical values which enclose the null hypothesis H_0 .If the absolute value of $|\hat{t}|$, with respect to the degree of freedom Φ and the significance level α is greater than the absolute critical value of the critical $|t_{crit}|$ then H_0 is discarded and the correlation is significant (Sachs and Hedderich, 2006):

$$H_0: \rho = 0 \text{ and } H_1: \rho \neq 0 \text{..... (3.4)}$$

$$\hat{t} = r * \sqrt{\frac{\Phi}{1-r^2}}, \text{ where } \Phi = n - 2 \ldots\ldots\ldots\ldots \text{ (3.5)}$$

After a study of relationship between rainfall and time series teleconnections in each grid box, next step in this research was to check multicollinearity in each model. For computational reasons, the explanatory variables $X_{1,t}, \ldots, X_{k,t}$ may not be (perfect) correlated. From a practical point of view, the estimated coefficients will be unstable und unreliable if explanatory variables are highly correlated. In the presence of multicollinearity, the effect of a single explanatory variable can't be isolated, as the regression coefficients are quite uninformative and their confidential intervals very wide. If the purpose of the model is only to predict the dependent variable, multicollinearity is not real a problem. However, if one is interested in the individual estimated coefficients, results should be interpreted with caution, since only imprecise information can be obtained from the regression coefficients. In the study at hand, the impact of explanatory variables on seasonal rainfall time series is important. Therefore, the models should be checked for multicollinearity.

3.3.1.5 Collinearity and stepwise VIF selection

A simple approach to identify collinearity among explanatory variables is the use of variance inflation factors (VIF). VIF calculations are straightforward and easily comprehensible, the higher the value, the higher the collinearity. Furthermore, to eliminate highly overlapping predictors, which introduce multi-collinearity issues, the predictors selection procedure relies on the variance inflation factors (VIF):

$$VIF_j = 1/(1 - R_j^2) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \text{ (3.6)}$$

Where R_j^2 (coefficient of determination) from a regression between the j^{th} candidate predictor and each selected predictor.

A VIF is calculated for each explanatory variable and those with high values are removed. Neter et al. (1960) indicate that VIF should not be larger than 10 to minimise multi-collinearity among predictors. The definition of 'high' is somewhat arbitrary but values in the range of 5-10 are commonly used**.** Chen and Georgakakos (2014) found that a VIF of 4 is more effective and does not undermine forecast accuracy and in this study the VIF threshold used is 5. In the last step for predictors selection, the backward stepwise linear regression is applied.

The method used is a forward selection method, in which all independent variables entered in the model at each step are reassessed based on their partial F-statistics.

3.3.1.6 Stepwise Regression Analysis

An important issue in regression modelling is the selection of explanatory variables which are really influencing the dependent variable. There are many methods for selection, stepwise regression analysis is frequently used variable selection algorithm in regression analysis. This is a modification of forward selection method, in which all independent variables entered in the model at each step are reassessed based on their partial F-statistics. An explanatory variable incorporated at earlier step may now be unnecessary because of the relationships between it and the latest variable entered in the model (Montgomery *et al.* 2003). The predictor variables finally selected by the stepwise algorithm were included in the final model.

If there are a large number of predictors, it is not possible to fit all possible models. For example, 40 predictors lead to 2^{40}> 1 trillion possible models! Consequently, a strategy is required to limit the number of models to be explored.

An approach that works quite well is backwards stepwise regression:

- Start with the model containing all potential predictors.
- Remove one predictor at a time. Keep the model if it improves the measure of predictive accuracy.
- Iterate until no further improvement.

If the number of potential predictors is too large, then the backwards stepwise regression will not work and forward stepwise regression can be used instead. This procedure starts with a model that includes only the intercept. Predictors are added one at a time, and the one that most improves the measure of predictive accuracy is retained in the model. The procedure is repeated until no further improvement can be achieved.

Alternatively, for either the backward or forward direction, a starting model can be one that includes a subset of potential predictors. In this case, an extra step needs to be included. For the backwards procedure we should also consider adding a predictor with each step, and for the forward procedure we should also consider dropping a predictor with each step. These are referred to as hybrid procedures. It is important to realise that any stepwise approach is not guaranteed to lead to the best possible model, but it almost always leads to a good model. For further details see James, Witten, Hastie and Tibshirani (2014).

3.3.1.7 Principal component Analysis

As it was introduced in section 3.2, the SST and SLP data used spans from 1949 to 2000 and was composed of monthly data. So, the time step $t = 1,2,3, \ldots \ldots n$ with n=624. Prior to the PCA annual cycles were calculated and subtracted from the SST/SLP data for each month per grid point. Let X (t, s) represent the new data with annual cycles removed. In practice, PCA is performed on either a correlation matrix or a covariance matrix. In this study the covariance matrix was used because it allows identifying the strongest variations in the dataset contrary to the correlation matrix in which the spatial variations in the dataset are removed (Wilks 2011). Hence, per grid point **s** the data were transformed as follows:

$$X'(t,s) = w(s) * (X(t,s) - \bar{X}(s)) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (3.7)$$

Here $\bar{X}(s)$ represents the arithmetic mean per grid point.

Basically the goal of PCA is to find the set of new variables or principal components $u_i(t)$ that summarise the information in the data X'(t,s), together with their associated variability modes or eigenvectors $e_i(s)$. Knowing those principal components time series and eigenvectors at any time the data can be reconstructed as follows (Storch and Zwiers 2004):

$$X'(t,s) = \sum_{i=1}^{m} u_i(t).\, e_i^T(s) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (3.8)$$

In principle, before obtaining the principal components a main mathematical issue has to be solved. Indeed, the mathematical issue that arises here is to find the eigenvectors $e_i(s)$ and the associated eigen values λ_i. This yield a system of equations that can be expressed as (Storch and Zwiers 2004):

$$A * E = \lambda * E \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (3.9)$$

$$(m * m)\ (m * 1) \qquad (m * 1)$$

Since the data set contains m grid points, this equation should be resolved m times. In fact, starting with the covariance matrix of the data ($A = X'^T X'$) has to be solved and the eigenvector $e_1(s)$ and its eigenvalue λ_1 obtained. This first eigenvector will have the largest variance. Subsequently the matrix A should be recomputed each time after subtracting the information explained by the i^{th} eigenvector and the next eigenvector ($[i+1]^{th}$) with the largest possible variance can be calculated together with its eigenvalue. Moreover, all the eigenvectors should be uncorrelated and thus can be denoted as empirical orthogonal functions (EOFs). The share

of information holds by each eigenvector or the variance of explained by each eigenvector $e_i(s)$ can be expressed as:

$$R_i^2 = \frac{\lambda_i}{\sum_{i=1}^{m} \lambda_i} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (3.10)$$

In general, the variance explained by each eigenvector decreases as the index of eigenvector increases.

The four tables (Appendix A, B, C and D) show the amount of information (eigen values in %) held by each of the five leading empirical orthogonal functions. The variance explained by each EOF decreases with the decreasing order of the EOFs. The season SST and SLP principal components variables considered in this study are constructed based on five data steps i.e actual data, three moths lead, six months lead, nine months lead and 12 months lead. For the long rainfall season counts 47.91%, 52.65%, 51.87%, 44.88% and 45.05% of the total SST variances respectively whereas the short rainfall season counts 50.30%, 46.16%, 47.91%, 52.65% and 51.87% of the total SST variances respectively. For the long rainfall season counts 50.43%, 59.82%, 52.09%, 55.82% and 50.78% of the total SLP variances respectively whereas the short rainfall season counts 52.15%, 56.23%, 50.43%, 59.82% and 52.09% of the total SLP variances respectively (cfr Appendices A, B, C, D). In this study only the spatial and the temporal characteristics of the first two PCA modes that accounted for most of the variance are discussed.

Once the eigenvectors were all obtained, they were further normalized ($\|e_i\| = 1$) and sorted. For each eigenvector e_i the normalisation implies that:

$$\sum_{s=1}^{m} e_i^2 (s) = 1 \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (3.11)$$

Therefore, the observed principal components were computed. Each principal component $u_i(t)$ was the result of the projection of the data $X'(t,s)$ onto the i^{th} eigenvector. Mathematically, this projection can be expressed as (Storch and Zwiers 2004):

$$u_i(t) = \sum_{s=1}^{m} e_i(s) * X'(t,s) \ldots\ldots\ldots\ldots \quad (3.12)$$

Each PC $u_i(t)$ holds the same amount of information as the EOF $e_i(s)$ it is generated from. In the following the principal components obtained from the SST and SLP data are denoted as SST PCs and SLP PCS.

3.3.2 Seasonal rainfall model development

Two kinds of models are developed i.e a statistical model which is developed based on all selected predictors (in this case are 72 predictors) and a predictive model developed based on only leading predictors (in this case are 58 predictors).

3.3.2.1 Time series linear regression model

The tslm function has been used to run the multiple linear regression model (MLR). The function tslm is largely a wrapper for lm() except that it allows variables "trend" and "season" which are created on the fly from the time series characteristics of the data. The variable "trend" is a simple time trend and "season" is a factor indicating the season (e.g., the month or the quarter depending on the frequency of the data).

3.3.2.2 ARIMA modeling

In the previous section, the multiple linear regression was described, together with possible problems that should be taken care of in order to benefit from desirable properties of the estimators. When regression is applied to time series data, the errors terms are often autocorrelated. If they are, ARIMA models can be used to model the information they contained. The resulting model is then a combination of a multiple regression and an ARIMA model in the error terms. This should enable us to obtain more reliable estimates for the effect of the explanatory variables on the dependent variable.

The ARIMA modelling approach expresses a variable as a weighted average of its own past values. The model is in the most cases a combination of an autoregressive (AR) part and moving average (MA) part. Suppose a variable N_t is modelled as an autoregressive process, AR*(P)*. Then, N_t can be expressed as a regression in terms of its own passed values:$N_t = C + \emptyset_1 N_{t-1} + \emptyset_2 N_{t-2} + \cdots + \emptyset_p N_{t-p} + a_t$, where C is a constant term, $\emptyset_i (i = 1,2, \dots, p)$ are the weights of the autoregressive terms and $a_t$is a new random term, which is assumed to be normally distributed "white noise" , containing no further information. Using the backshift operator B^i on N_t, defined as $B^i N_t = N_{t-i} (i = 1,2,3 \dots)$, this process can be written as $N_t = C + \emptyset_1 B N_t + \emptyset_2 B^2 N_t + \cdots + \emptyset_p B^p N_t + a_t$, or $\left(1 - \emptyset_1 B - \emptyset_2 B^2 - \cdots - \emptyset_p B^p\right) N_t = C + a_t$.The series N_t can also be expressed in terms of the random errors of its past values which is then a moving average MA *(q)* model : $N_t = C + a_t - \theta_1 a_{t-1} + \theta_2 a_{t-2} + \cdots + \theta_q a_{t-q}$, where

$\theta_j (j = 1,2, \dots, p)$ are the weights for the moving average terms. Using the Backshift operator, this equal $N_t = C - \theta_1 B a_t - \theta_2 B^2 a_t - \cdots - \theta_q B^q a_t + a_t$, or

$N_t = C + \left(1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_p B^q\right) a_t$. In more general settings, it is possible to include autoregressive and moving average terms in one equation, leading to ARMA*(p,q)* model: $\left(1 - \emptyset_1 B - \emptyset_2 B^2 - \cdots - \emptyset_p B^p\right) N_t = C + \left(1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_p B^q\right) a_t$, where a_t is again assumed to be "white noise".

However, An ARIMA model can't be applied in all circumstances. It is required that the series be stationary. For practical purposes, it sufficient to have weak stationary, which means that the data is in equilibrium around the mean and the variance remains constant over time. If a series is non-stationary because the variance is not constant, it often helps to log-transform the data. To have a series that is stationary in the mean, differencing is used. For example, in order to obtain a stationary, the data may be differenced once for the period by the period(monthly) fluctuations $\nabla X_t = X_t - X_{t-1}$. When an ARMA model is built on differenced data, it is called an ARIMA model, where "*I*" indicates the differencing.

3.3.2.3 Unit root and Stationarity Tests

Most time series data could be nonstationary due to the presence of random walk, drift, or trend. One best way to test these is to evaluate a regression that nests a mean, a lagged term which checks for difference stationarity, and a term for deterministic trend which also looks for trend stationary in one particular model:

$$Y_t = \alpha + y_{t-1} + \beta t + \varepsilon_t \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots \quad (3.13)$$

By taking the first difference of (3.13), we get $\nabla Y_t = \alpha + (\rho - 1) y_{t-1} + \beta t + \varepsilon_t$.
This model forms the basis of the Dickey-Fuller unit root test. The application of the Dickey-Fuller test mainly depends on the regression context in which the lagged dependent variable is tested. The three identified model contexts are those of (1) a pure random walk, (2) random walk plus drift, and (3) the combination of deterministic trend, random walk, and drift. In line with this, three different regression equations are considered:

$$Y_t = \rho_1 y_{t-1} + \varepsilon_t, \varepsilon_t \sim i.i.d(0, \sigma^2) \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots \quad (3.14)$$

$$Y_t = \alpha_0 y_{t-1} + \rho_1 y_{t-1} + \varepsilon_t, \varepsilon_t \sim i.i.d(0, \sigma^2) \dots\dots\dots\dots\dots\dots \quad (3.15)$$

$$Y_t = \alpha_0 y_{t-1} + \rho_1 y_{t-1} + bt + \varepsilon_t, \varepsilon_t \sim i.i.d(0, \sigma^2) \dots\dots\dots\dots \quad (3.16)$$

Equation (3.14) specifies a regression model without a constant term. This regression model is used to test for pure random walk process without drift. Here, the null hypothesis of nonstationary random walk is tested against a stationary series. If $\rho_1 = 1$, then the null hypothesis cannot be rejected and the data generating process is inferred to have a unit root.
The second Dickey-Fuller case in (3.15) involves a context of random walk plus drift around a nonzero mean. The null hypothesis is that the series under consideration is integrated at the first order, that is, $I(1)$. In the other words, the null hypothesis of whether $\rho_1 = 1$ is tested against a stationary series around a constant mean of $\frac{\alpha_0}{(1-\rho)}$.

The third Dickey-Fuller case is one with a context of random walk plus drift in addition to a deterministic linear trend shown in (3.16). As in the earlier cases, the null hypothesis is that $\rho_1 = 1$ $(\rho_1 - 1 = 0)$ and the alternative hypothesis is that the series is stationary. Nonetheless, not all Dickey-Fuller regression models have white noise residuals. This means, in a situation where the error term (ε_t) in (3.14), (3.15) and (3.16) are autocorrelated, the Dickey-Fuller distribution might not be applicable. However, if there is autocorrelation in the series, it has to be removed from the residuals (ε_t) of the regression before Dickey-Fuller tests are executed. Under the conditions of residual serial correlation, where the Dickey-Fuller regression models are not valid for the unit root test, a new test called the Augmented Dickey-Fuller (ADF) test in (3.5) may be applied. This new regression model addresses the issue of serial correlation.

$$Y_t = \alpha_0 + \rho_1 y_{t-1} + \sum_{1=2}^{\rho=1} \beta_i \nabla y_{t-i} + \varepsilon_t \ldots\ldots\ldots\ldots\ldots (3.17)$$

In a situation where the process is ARIMA(p, q), Said and Dickey were reported by Yaffee and McGee (2000) to have discovered that the MA(q) parameter invertibility can be represented by an AR(p) process of the kind in (3.17) when p gets large enough. The Augmented Dickey-Fuller equation is identical to the three Dickey-Fuller equations discussed earlier, except that the ADF equation contains higher order lags of the differenced dependent variable which take care of serial correlation before testing for nonstationarity. If the series has a higher order serial correlation which result to an AR unit root, higher order differencing will be required in order to transform the residuals into white noise disturbances. Moreover, utmost care should be taken since over-differenced series might also result to an MA unit root.

In the case where the series under study exhibit patterns of random walk plus drift around a stochastic trend, the Dickey-Fuller test can be reconstructed with the additional of a time trend variable as shown in (3.18)

$$Y_t = \alpha_0 + \rho_1 y_{t-1} + \sum_{1=2}^{\rho=1} \beta_i \nabla y_{t-i} + bt + \varepsilon_t \ldots\ldots\ldots\ldots\ldots (3.18)$$

Where $\rho_1 y_{t-1} + \sum_{1=2}^{\rho=1} \beta_i \nabla y_{t-i}$ is the Augmented part, y_{t-1} is the lagged term of Y_t , ∇y_{t-i} shows the lagged change, t represent the deterministic trend, α is the drift component, ε_t represents a well-behaved error term (unobserved series) and b, ρ_1, β are coefficients to be estimated.

Generally, an ADF test with hypothesis:$H_0: \rho_1 = 0$ and $H_A: \rho_1 < 0$ can be tested in the regression model (3.6).

3.3.2.4 Regression with ARIMA errors model

The ARIMA modelling approach can now be applied to the multiple regression equation to model the information that remains in the error terms. Assume a regression model with an explanatory variable, denoted as $Y_t = \beta_0 + \beta_1 X_{1,t} + N_t$. Suppose further that the error terms are autocorrelated, and that they can be appropriately described by an ARMA (1,1) process. This model can be written as $Y_t = \beta_0 + \beta_1 X_{1,t} + N_t$, where $(1 - \emptyset_1 B) N_t = (1 - \theta_1 B) a_t$, and a_t is assumed to be wite noise. Substituting the correction of error term into the regression equation gives:

$$Y_t = \beta_0 + \beta_1 X_{1,t} + \frac{(1-\theta_1 B)}{(1-\emptyset_1 B)} a_t \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3.19)$$

Because of the specific form in the error terms, the classical least squares methods are not appropriate to estimate the parameters of this equation.

An important consideration when estimating a regression with ARMA errors is that all of the variables in the model must first be stationary. Thus, we first have to check that Y_t and all of the predictors $(X_{1,t}, \ldots, X_{k,t})$ appear to be stationary. If we estimate the model when any of these are non-stationary, the estimated coefficients will not be consistent estimates (and therefore may not be meaningful). One exception to this is the case where non-stationary variables are co-integrated. If there exists a linear combination of the non-

stationary Y_t and the predictors that is stationary, then the estimated coefficients will be consistent (Harris and Sollis,2003).

We therefore first difference the non-stationary variables in the model. It is often desirable to maintain the form of the relationship between Y_t and the predictors, and consequently it is common to difference all of the variables if any of them need differencing. The resulting model is then called a "model in differences," as distinct from a "model in levels," which is what is obtained when the original data are used without differencing.

If all of the variables in the model are stationary, then we only need to consider ARMA errors for the residuals. It is easy to see that a regression model with ARIMA errors is equivalent to a regression model in differences with ARMA errors. For example, if the above regression model with ARIMA (1,1,1) errors is differenced we obtain the model

$$Y_t' = \beta_1 X_{1,t}' + \cdots + \beta_k X_{k,t}' + \eta_t' \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (3.20)$$

$$(1 - \emptyset_1 B)\eta_t' = (1 - \theta_1 B)\varepsilon_t \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (3.21)$$

Where $Y_t' = Y_{t-1}, X_t' = X_{t,i} - X_{t-1,i}$ and $\eta_t' = \eta_t - \eta_{t-1}$, which is a regression model in differences with ARMA errors.If differencing is applied to the errors in multiple regression, Pankratz (2012) shows that all corresponding series (both dependent and the explanatory variables) should be differenced. This can be seen in our small regression example. Differencing the error terms once results in the following expression, with the ARMA (1,1) model now in the differenced error terms:

$$\nabla N_t = \frac{(1-\theta_1 B)}{(1-\emptyset_1 B)} a_t \iff N_t = \frac{(1-\theta_1 B)}{\nabla(1-\emptyset_1 B)} a_t \ldots\ldots\ldots\ldots\ldots \quad (3.23)$$

Substituting back the expression into the regression equation gives:

$$Y_t = \beta_0 + \beta_1 X_{1,t} + \frac{(1-\theta_1 B)}{(1-\emptyset_1 B)} a_t \iff \nabla Y_t = \beta_0' + \beta_1 \nabla X_{1,t} + \frac{(1-\theta_1 B)}{(1-\emptyset_1 B)} a_t \ldots\ldots\ldots \quad (3.24)$$

The intercept is now possibly different, but the (theoretical) regression coefficient β_1 is not affected by the differencing operation. Its estimated value may differ slightly, since the estimation is done on different (although related) time series.

Regression with ARIMA errors consist of three steps (Han et al. (2010); Box and Jenkins (1979); Cryer and Chan (2008)):

1. Identification:

In this stage, first, the raw data series is plotted to identify whether the data is stationary or not. If the raw data series is found to be non–stationary, differencing is required. After the first order differencing, correlograms of the autocorrelation function (ACF) and partial autocorrelation function (PACF) is investigated. From these plots, the order of AR and MA gets identified.

2. Parameter Estimation and Selection:

The number of AR depends on the lag of PACF cuts and the number of MA depends on the lag of the ACF plot. However, decision making on the order of AR and MA by looking at the cuts/spikes is not straightforward. Most of the time it required experimentation with several alternative orders of different models to choose the appropriate order. The following guidelines are usually followed during the selection of the AR and MA order:

- If the ACF plot shows exponential decay and PACF spikes at lag–1, no correlation for other lags, in that case, one autoregressive parameter ($p = 1$) can be selected.
- If the ACF plot shows a sine–wave shape pattern or a set of exponential decay and PACF spikes at lag–1 and lag–2, no correlation for other lags, in that case, two autoregressive parameters ($p = 2$) can be selected.
- If the PACF plot shows exponential decay and ACF spikes at lag–1, no correlation for other lags, in that case, one moving average parameter ($q = 1$) can be selected.
- If the PACF plot shows a sine–wave shape pattern or a set of exponential decay and ACF spikes at lag–1 and lag–2, no correlation for other lags, in that case, two moving average parameters ($q = 2$) can be selected.
- One auto–regressive and one moving average parameter can be selected if both shows exponential decay starting at lag–1.
- Sometimes, using both AR and MA orders in a model can cancel each other's impact. Therefore, it is often wise to use mixed AR and MA models with a smaller number of orders.

3. Diagnostics Check:

The diagnostic check is required to verify the adequacy of the developed model. The residual of the developed model should be white noise (no autocorrelation). To check whether the residual is white noise or not, at first, an inspection of the residual ACF and PACF plot is required. If 95% of the spikes stay between the black lines, it indicates that the autocorrelation is white noise. If two or more spikes or more than 5% of spikes are located outside of the boundary line, then the series is not white noise. Another way of checking the model accuracy is to perform the Ljung–Box test. Such a test is conducted to verify the null hypothesis of being white noise of residual if the p–value is greater than 0.05 (Ljung and Box (1978)). A p–value greater than 0.05 implies that lag autocorrelation among the residuals is zero and the developed model is adequate to fit the data set.

3.3.3 Time series cross validation

When you build your model, you need to evaluate its performance. Cross-validation is a statistical method that can help you with that. For example, in K-fold-Cross-Validation, you need to split your dataset into several folds, then you train your model on all folds except one and test model on remaining fold. You need to repeat this step until you tested your model on each of the folds and your final metrics will be average of scores obtained in every fold. This allows you to prevent overfitting and evaluate model performance in a more robust way than simple train-test.

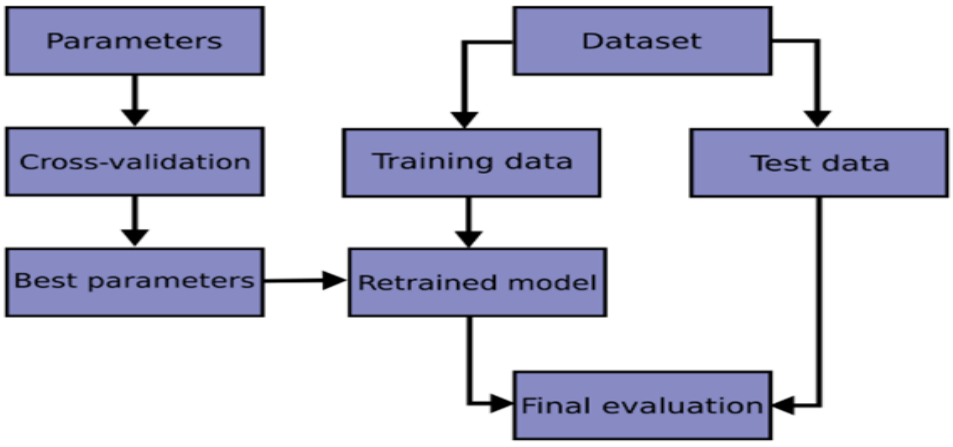


Figure 3.1:The concept at the base of Cross Validation

The most accepted technique in the ML world consists in randomly picking samples out of the available data and split it in train and test set. Well to be completely precise the steps are generally the following:

1. Split randomly data in train and test set.
2. Focus on train set and split it again randomly in chunks (called folds).
3. Let's say you got 10 folds; train on 9 of them and test on the 10th.
4. Repeat step three 10 times to get 10 accuracy measures on 10 different and separate folds.
5. Compute the average of the 10 accuracies which is the final reliable number telling us how the model is performing.

In the case of time series, the cross-validation is not trivial. We cannot choose random samples and assign them to either the test set or the train set because it makes no sense to use the values from the future to forecast values in the past. In simple word we want to avoid future-looking when we train our model. There is a temporal dependency between observations, and we must preserve that relation during testing.In this procedure, there is a series of test sets, each consisting of a single observation. The corresponding training set consists only of observations that occurred prior to the observation that forms the test set. The following diagram illustrates the series of training and test sets, where the blue observations form the training sets, and the red observations form the test sets.

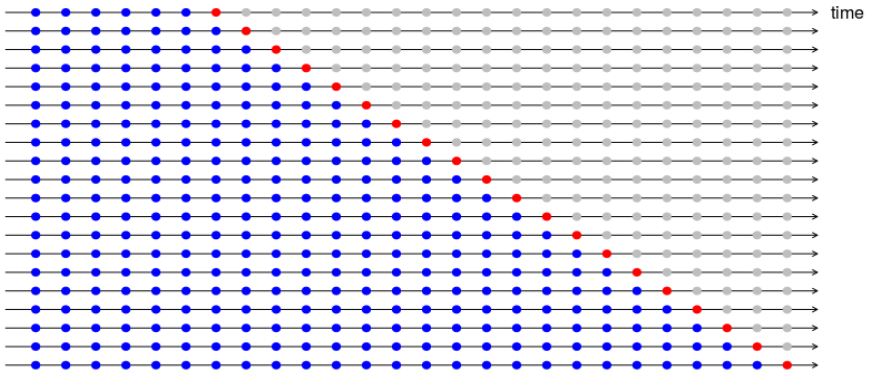


Figure 3.2:Time series Cross Validation (Bergmeir., et al. (2018))

3.4 Comparison of forecasting ability of different statistical techniques

Among all the statistical parameters that are used to evaluate time series model's performance, RMSE calculates prediction errors, the measure of how much a dependent series varies from its model–predicted level. MAE is the average of the absolute errors/residuals between observed and predicted value, R-squared explains to what extent the variance of one variable explains the variance of the second variable. For both RMSE and MAE, a value of 0 indicates a perfect predictability performance. Thus, the lower the value of RMSE, MAE and the improved value of R-squared, the better is the model's performance (Saigal and Mehrotra (2012); Singh et al. (2005)). The equation for RMSE, MAE, and R-squared is presented below:

3.4.1 Root Mean squared error (RMSE)

The square root of mean squared error which is also known as standard error of estimate in regression analysis or the estimated white noise standard deviation in ARIMA analysis. It is expressed as:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2}{N}} \quad \text{..... (3.25)}$$

Where, Y_i is the Actual value, $\hat{Y}_i$ is the predicted value and N is the number of observations.

3.4.2 Mean Absolute error (MAE)

Mean absolute error is another criterion to measure the performance of forecasting model and is written as:

$$MAE = \frac{\sum_{i=1}^{N}(\hat{Y}_i - Y_i)}{N} \quad \text{..... (3.26)}$$

Where, Y_i is the Actual value, $\hat{Y}_i$ is the predicted value and N is the number of observations.

3.4.3 R-squared (R^2)

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{N}(Y_i - \bar{Y})^2} \quad \text{..... (3.27)}$$

Where, Y_i is the Actual value, $\hat{Y}_i$ is the predicted value, $\bar{Y}$ is the mean value and N is the number of observations.

4 RESULTS AND DISCUSSION

Keeping in view of the objectives of the present research study, the data chronicled on pretipitation and selected weather teleconections of East Africa was analysed by using proposed methodologies delineated in the previous chapter. Therefore, according to the objectives of the study, the results obtained in the study are discussed in this chapter. The chapter is divided into the following sections:

4.1. Temporal and spatial rainfall fluctuation time series and models comparison in East Africa

4.2. Linear Regression models and Regression with ARIMA errors models in East Africa

In this study both linear and nonlinear time models i.e multiple linear regression analysis and Regression with ARIMA errors models are used to analyse the past behaviour of rainfall time series data, in order to make inferences about its future behaviour for seasonal rainfall of East Africa.

4.1 Temporal and spatial rainfall time series analysis

The first part of this study analyses the rainfall fluctuation in whole Central- East Africa region which span over a longitude of 15°E to 55°E and a latitude of 15°S to 15°N as it is descried in chapter 3. The CRU monthly precipitation data collected from 1949 up to 2000 are used to investigate the objectives of this research in the whole part of the region in order to increase the sample size which is very important for rainfall trend analysis.

4.1.1 Long term means time series precipitations in Central-East Africa

The long-term rainfall characteristics are presented spatially with the help of maps. The area which is considered cover a longitude of 15°E to 55°E and a latitude of 15°S to 15°N. The selected area comprises almost 20 countries: Angola, Burundi, Cameroon (East), Congo-Brazzaville (East), Djibouti, Eritrea (South), Ethiopia, Kenya, Madagascar (North), Malawi (North), Mozambique (North), Rwanda, D.R.C, Central African Republic (CAR), Somalia, Sudan (South), Chad, Uganda, Tanzania, Zambia (North). Rainfall in these countries has similarities because their climate is controlled in regard to the tropical climate types which relates to the position of ITCZ.

The seasonal rainfall patterns in this region are determined by the presence or absence of rainfall, the amount of rainfall with respect to seasons are different considering the geographical position of each grid point.

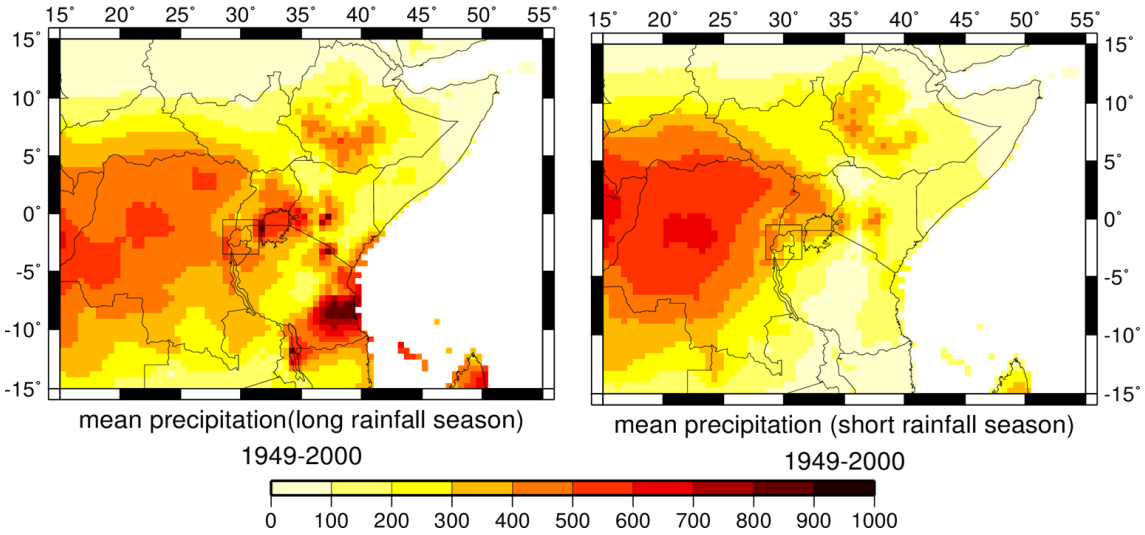


Figure 4.1: Spatial and temporal distribution of seasonal mean precipitation

The figure 4.1 shows an average of less than 1,000 millimetres of rainfall per year across most of parts of central African region. Rainfall tends to decrease with distance from the equator and is negligible in the Sahara (north of between latitude 10°N -15°N), in eastern Somalia for both long rainfall and short rainfall seasons. For Short rain fall season, the amount of rainfall tends to decrease in central part of Tanzania and continue to decrease in northern Mozambique, North Malawi, and North-Est of Zambia. Rainfall is most abundant on the North of Madagascar, portions of the highlands in eastern Africa and large areas of the Congo Basin and central Africa.

4.1.2 Seasonal variability of rainfall in the region

For present study, the rainfall variability is measured by the standard deviation. As it is done for mean precipitation in the region, the spatial and temporal season standard deviation in specified countries is presented spatially with the help of maps.

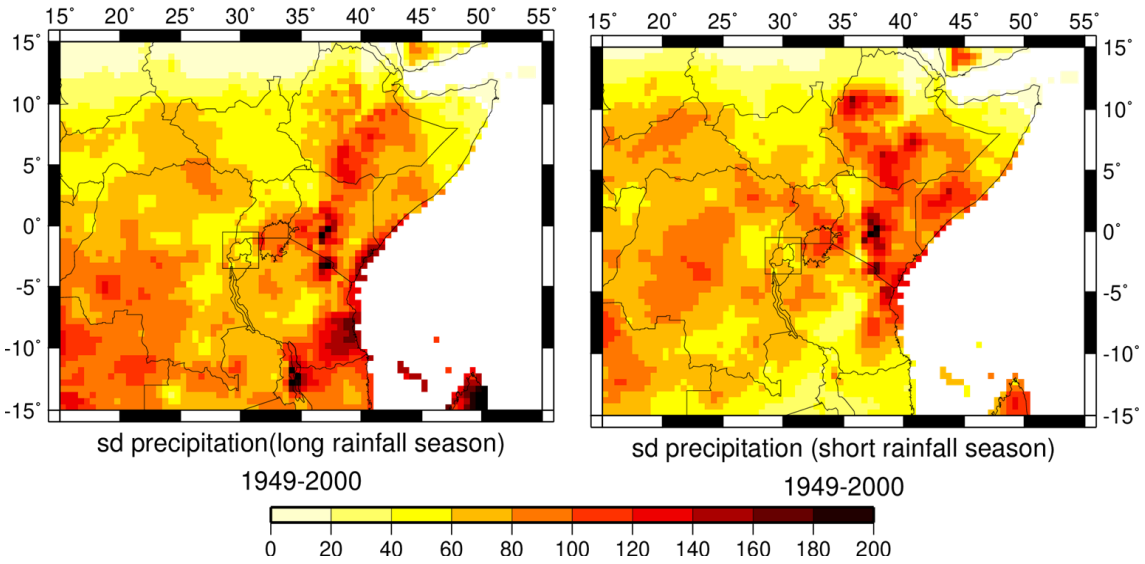


Figure 4.2: Spatial and temporal distribution of standard deviation for long and short rainfall Seasons

It is shown in figure 4.2 that some points of the region exhibit more variation within each season in terms of precipitation. Based on the two maps in Figure 4.2, less precipitation variation in the northern part of the region and eastern part of Somalia which are characterized by less precipitation during long and short rainfall season, is observed. Similar pattern is also observed in southern part of this region during the short rainfall season. The rest of the region in the maps shows variations that closely differ in space.

4.1.3 Prediction performance of seasonal rainfall models in central east Africa

The MLRM and RARIMAE model in ich each grid box was developed. The R squared and RMSE values for each model was computed.

4.1.3.1 Prediction performance of seasonal rainfall models explained by R-squared

In this study, R^2 is a statistical measure that represents the proportion of the variance for a precipitation that is explained by the selected variable or variables in a model. Whereas correlation explains the strength of the relationship between an independent and dependent variable, R^2 explains to what extent the variance of one variable explains the variance of the second variable. So, if the R^2 of a model is 0.50, then approximately half of the observed variation can be explained by the model's inputs.

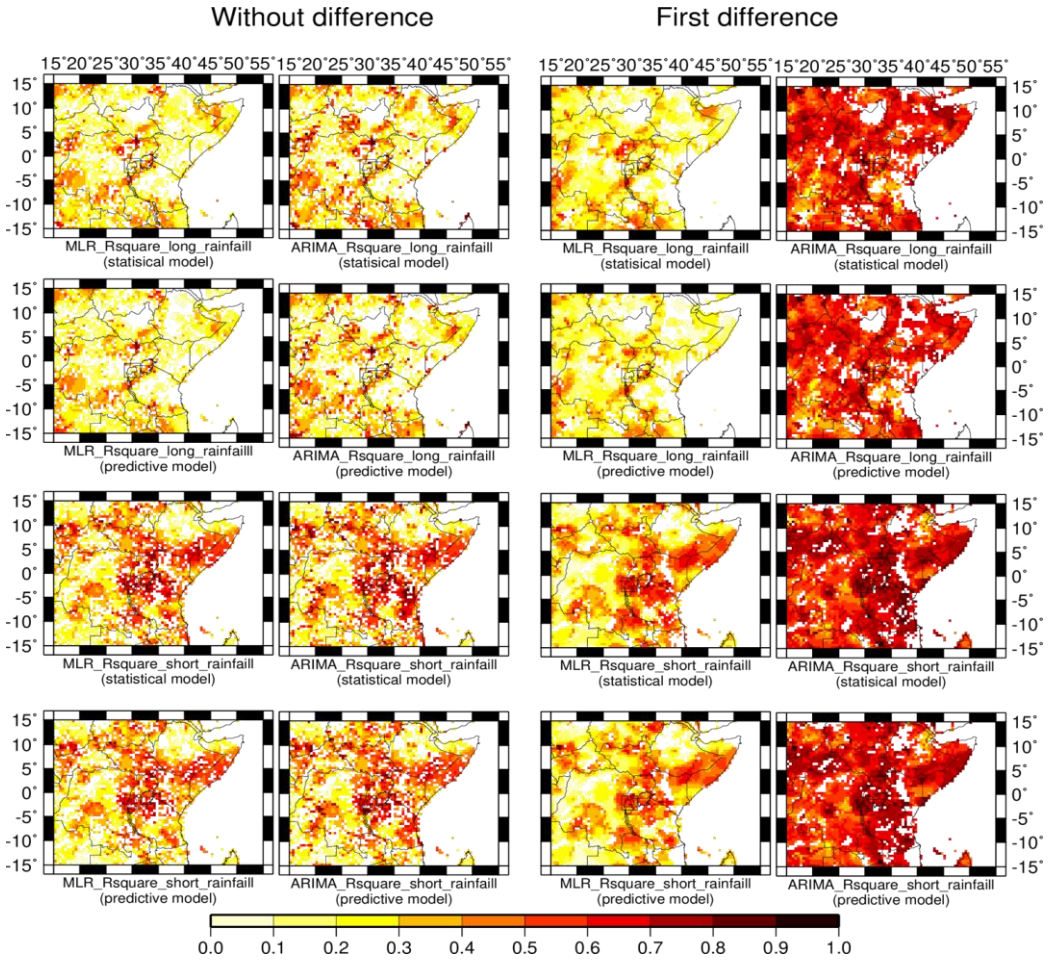


Figure 4.3:Spatial and temporal distribution of R^2 values for MLR and RARIMAE

The Figure 4.3 shows the R^2 distribution which describe the variation of precipitation in the region explained by pre-described teleconnections in chapter 3. The maps presented in figure 5 are categorised in two parts. The first one counts 8 maps in first two columns. The R-squared in these maps are computed based on the original data for both statistical model and predictive model in long rainfall season as well as in short rainfall season. The second one comprises also 8 maps in two last columns. The R^2 in these maps are computed based on the differentiated data for both statistical model and predictive model in long rainfall season as well as short rainfall season. In the second column for this category an improvement in R^2 values is observed in most of grid boxes of the region compared to the R^2 values in the first column. And this is due to the presence autocorrelation in observations for time series after the first difference.

4.1.3.2 Prediction performance of seasonal rainfall models explained by RMSE

The RMSE is the standard deviation of the residuals (prediction errors). Residuals are a measure of how far from the regression line data points are. Therefore, RMSE is a measure of how spread out these residuals are. In other words, it tells us how concentrated the data is around the line of the best fit.

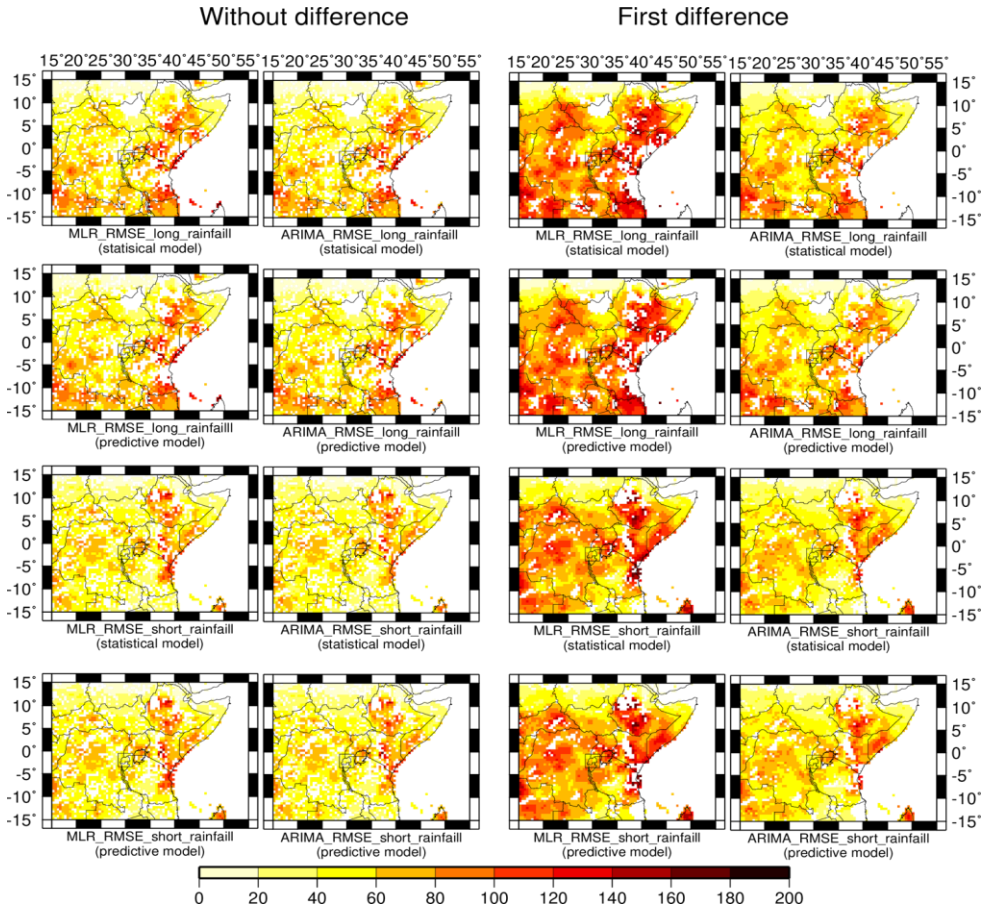


Figure 4.4:Spatial and temporal distribution of RMSE values for MLR and RARIMAE

The Figure 4.4 presents the RMSE maps which describe how far from the fitted line data points are for precipitation in the region impacted by pre-described teleconnection in chapter 3. The maps presented in Figure 4.4 are categorised in two parts. The first one comprises 8 maps in first two columns. The RMSE in these maps are computed based on the original data for both statistical model and predictive model in long rainfall season as well as short rainfall season. The second one comprises 8 maps in two last columns. The RMSE in these maps are computed based on the differentiated data for both statistical model and predictive model in long rainfall season as well as short rainfall season. An improvement in RMSE is observed for differenced

data in most of grid boxes of the region and this is due to the presence of autocorrelation in observations for time series after the first difference.

4.1.3.3 Seasonal rainfall models validation in each grid box

Once we are done with training our model, we cannot assume that it is going to work well on data that it has not seen before. In other words, we can't be sure that the model will have the desired accuracy and variance in prediction environment. We need some kind of assurance of the accuracy of the predictions that our model is putting out. For this, we need to validate our model. This process of deciding whether the numerical results quantifying hypothesized relationships between variables, are acceptable as descriptions of the data, is known as validation.

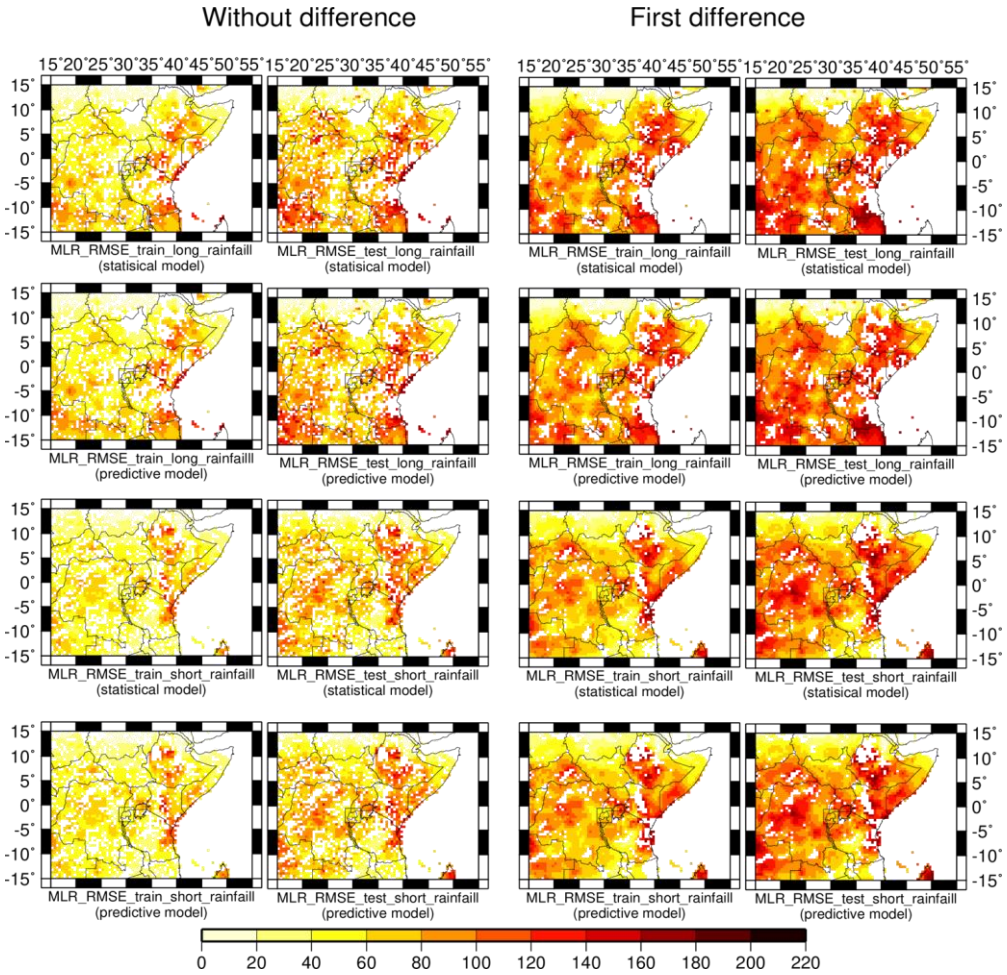


Figure 4.5: Spatial and temporal distribution of RMSE for training and Testing for MLR model

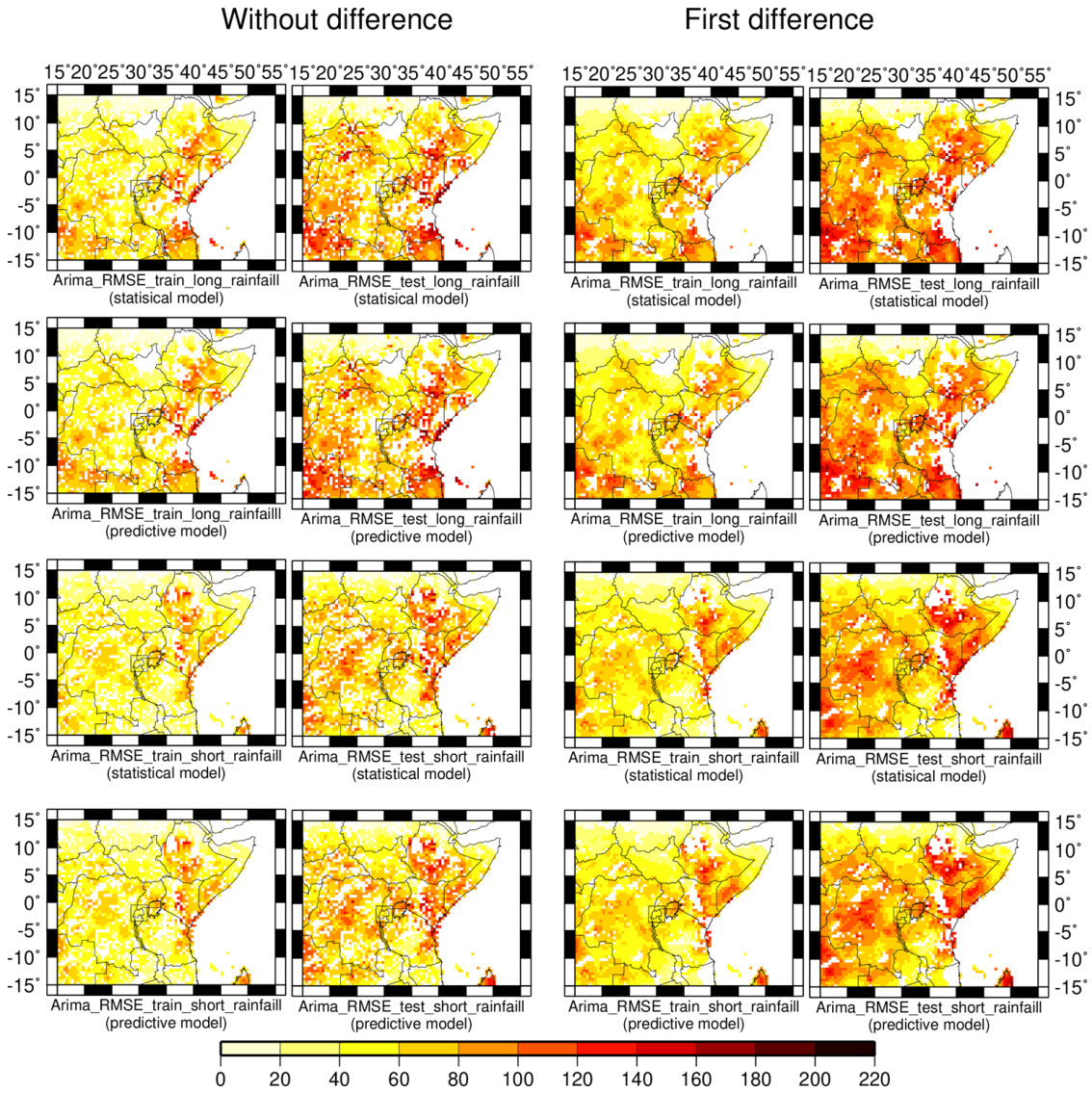


Figure 4.6: Spatial and temporal distribution of RMSE for training and Testing for RARIMAE model

The Figure 4.5 and Figure 4.6 present maps in which the RMSE for training and testing data are presented for both MLR and RARIMAE model respectively. The developed RARIMAE has shown low values of RMSE compared to the developed MLR model. Low values of this parameter in all grid box indicate a good prediction performance of the developed RARIMAE model. Once the model was developed for the calibration period, validation tests were developed with the same model inputs sets using RARIMAE analysis. In the validation period, the developed model showed an increase in RMSE compared to the calibration period for all grid boxes for both RARIMAE and MLR model.

4.1.3.4 Improvement in R-squared and Reduction in RMSE by RARIMAE

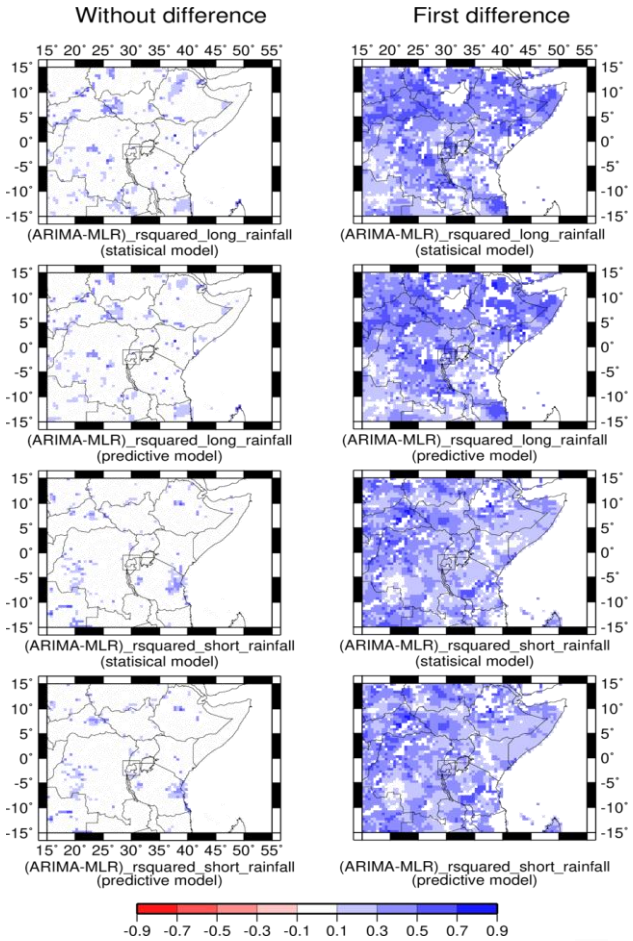


Figure 4.7: The difference between MLR-R^2 values and RARIMAE-R^2 values

Figure 4.7 shows the difference between explained variances (R^2) of monthly precipitation during 1949-2000 of MLR model and RARIMAE model. The values in Figure 4.7 are obtained by subtracting R^2 explained by RARIMAE model from R^2 explained by MLR model. Positive value in Figure 4.7 indicates that R^2 is improved by RARIMAE. For original data (left column), most of the grid box present the difference between explained variances for the two model which is equal or almost equal to zero. This is an indication that the observations in time series data in these grid boxes are non-autocorrelated. Therefore, many grids' boxes present an increase of R^2 less than 30%. For differenced data, where the autocorrelation within observation of residuals is present, an increase in R^2 for most of grid box is more than 30% and go up 70.

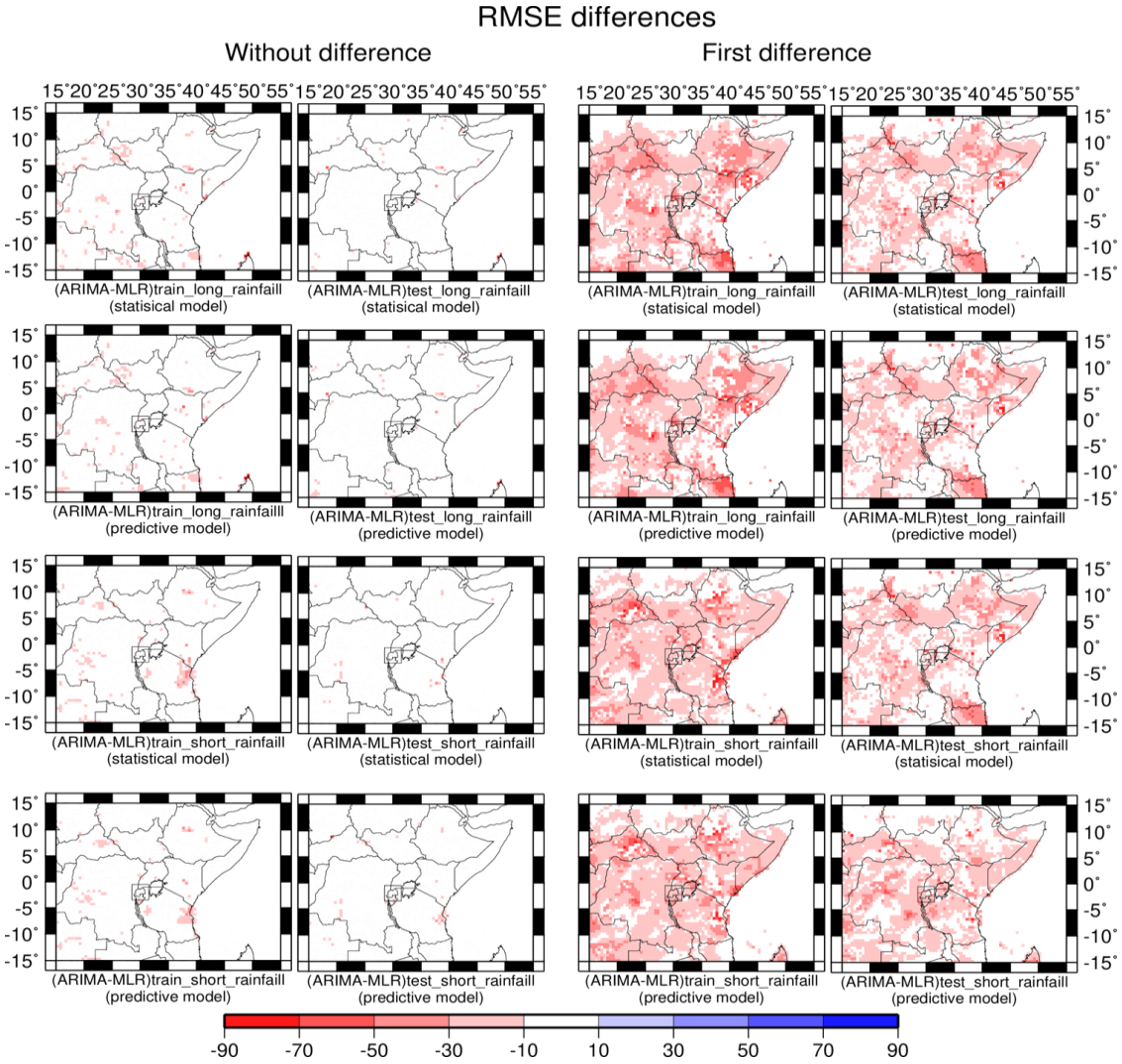


Figure 4.8:The difference between MLR-RMSE values and RARIMAE-RMSE values

The Figure 4.8 shows the differences in RMSE of monthly precipitation during 1949-2000 of MLR and RARIMAE model. The values in Figure 4.8 are obtained by subtracting RMSE explained by MLR model from RMSE explained RARIMAE model. Negative value in Figure 4.8 indicates that RMSE is reduced by RARIMAE for training and testing data sets. For original data (2 columns in the left), most of the grid boxes present the difference between RMSE for the two model which is equal or almost equal to zero. This is an indication that the observations in time series data in these grid boxes are non-autocorrelated. Therefore, many grids' boxes present a reduction of RMSE less than 30 mm/season. For differentiated data (2 columns in the right), where the autocorrelation within observation of residuals is present, an increase in RMSE for most of grid box is more than 30 mm/season and go up to 90 mm/season.

4.2 Seasonal rainfall Models in East Africa

To perform residuals analysis for MLR and RARIMAE models, a single time series was constructed by averaging monthly precipitation in the selected grid box for each season (Figure 4.9). The area which is considered cover a longitude of 31.5°E to 41°E and a latitude of 3.5°S to 0.5°S. This region covers central east of DRC, north of Burundi, south of Uganda, Rwanda, north of Tanzania and south of Kenya.

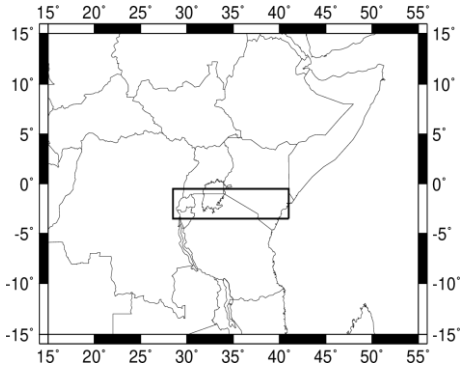


Figure 4.9: East African region of averaging monthly rainfall time series

Statistic	Long rainfall season	Short rainfall season
Observation	52	52
Mean	384.5	254.6
Median	397.8	237.1
Standard Deviation	64.3	79.82
Minimum	249.8	139.5
Maximum	545.0	503.3
Skewness	0.10	1.13
Kurtosis	-0.37	0.83
Coefficient of Variation (%)	16.7	31.35

Table 4.1:Summary for rainfall original time series data

The descriptive statistics of the seasonal rainfall time series are presented in Table 4.1. The highest mean and median values in the period of 52 years were observed in long rainfall season. The Mean, min, max, standard devition (Std) and median have the units corresponding to the units of metheorological varialbe (mm); skewness and kurtosis are non-demensional.

The parameters of skewness and kurtosis of the analysed time series give information about differences in their statistical distributions. For the long rainfall season, precipitation time series is characterised by positive skewness (0.10) and small and negative kurtosis (-0.37), which inform us that this distribution is nearly symmetrical. A different distribution shape can be observed for precipitation time series during the short rainfall season, with positive skewness (1.13) and low kurtosis values (0.83). This means that this distribution is also nearly symmetrical. The summary or descriptive statistics of seasonal rainfall in Est Africa in Table 4.1 indicates that the short rainfall season data are more heterogenous (with 31.5% of coefficient of variation) than that for the long rainfall season (with 16.7% of coefficient of variation).

4.2.1 Regression models for forecasting long rainfall season in East Africa

The variables considered for regression analysis are described in chapter 3. After the development of 4 leading variables with three months' time steps for each original teleconnection, the total number of independent variables increases up to 72. The multiple linear regression analysis was carried out by considering all teleconnections as predictors while the long rainfall season time series data from CRU is considered as predictand. The first step was to select teleconnections which are significantly correlated with rainfall time series at 5% level of significance.

To overcome the multi-collinearity problem, one of the measures was to drop the unimportant variables *i.e.* the variables which are explaining less variations in dependent variables have need to be drop from the model and the dropping of variable was done through the collinearity and stepwise VIF selection method. Finally, stepwise regression analysis was carried out to fit the model.

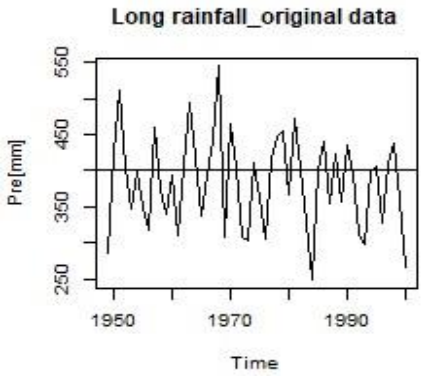


Figure 4.10: Long rainfall season time series data

4.2.1.1 Estimation and significance check of model parameters

Statistical model				
Variable	Coefficient	Std. Error	t test	Probability
Constant	362.90	11.58	31.33	<2e-16
SLP_PC19	21.49	8.19	2.63	0.0115
DMI_LR9	-62.81	2.80E-06	-2.61	0.0121
Predictive model				
Variable	Coefficient	Std. Error	t test	Probability
Constant	362.90	11.58	31.33	<2e-16
SLP_PC19	21.49	8.19	2.63	0.0115
DMI_LR9	-62.81	2.80E-06	-2.61	0.0121

Table 4.2:Estimates of regression model for long rainfall original time series data

Table 4.2 shows that the unexplained or non-significant variables are dropped from the model so that one can get maximum error degrees of freedom. In this stepwise regression analysis, we obtained a totally of two significant independent variables for both statistical and predictive model. These variables are the Indian Ocean first principal component Sea Level Pressure nine months lead (SLP_PC19) and the Dipole Mode Index nine months lead (DMI_LR9). After getting regression parameters the next step is residuals of MLR model analysis (Figure 4.11).

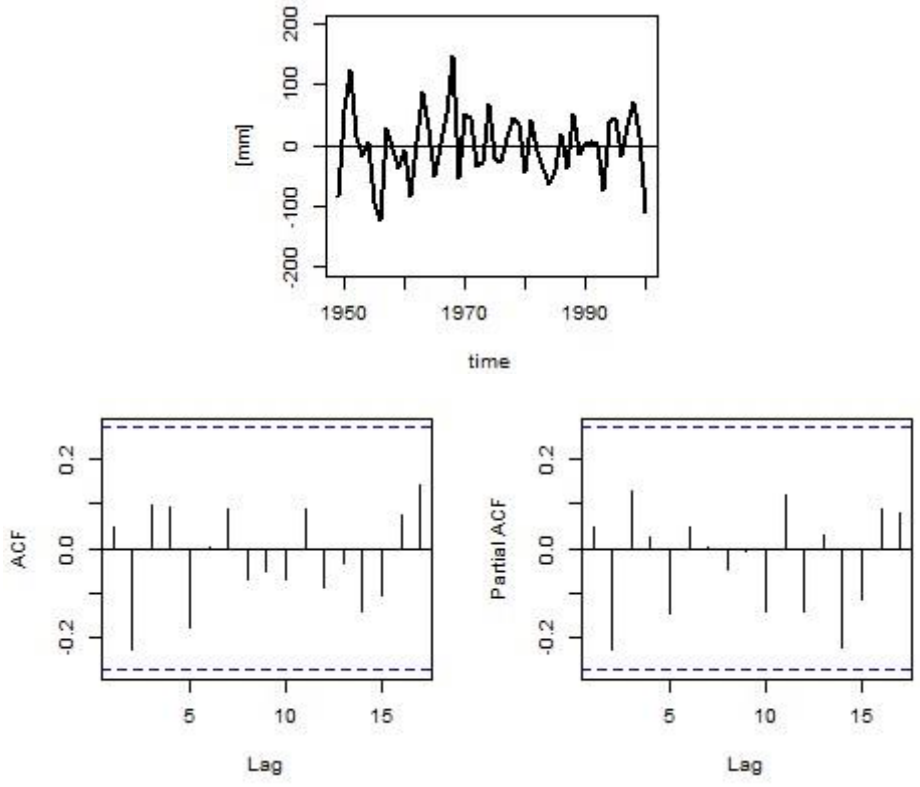


Figure 4.11:Diagnostic of residual plots for MLR

The residuals time series and the Autocorrelation function (ACF) and partial autocorrelation function (PACF) plots are presented in Figure 4.11. Based on the three plots one can say that the test confirms the non-presence of autocorrelation pattern in series, all spikes do not show an autocorrelation pattern in residuals. This is a good sign when the diagnostic of every model is done. In time series Analysis auto correlation is useful because its presence tells you important things about the variable and potential problems with your model. But When using Ordinary Least Square (OLS) to estimate a model (like MLR model) auto correlation in the residual terms violates one of the Gauss–Markov conditions (that the errors are independent). This condition is necessary for making OLS estimates minimum variance ("best") among the class of linear unbiased estimators. So, it is very useful in modelling to have solid evidence to suggest that the error is random and hence not predictable. Otherwise, there would be some better model we can build. Based on that the next step is to check if there exist any autocorrelation within time series data and with residuals computed after getting the model.

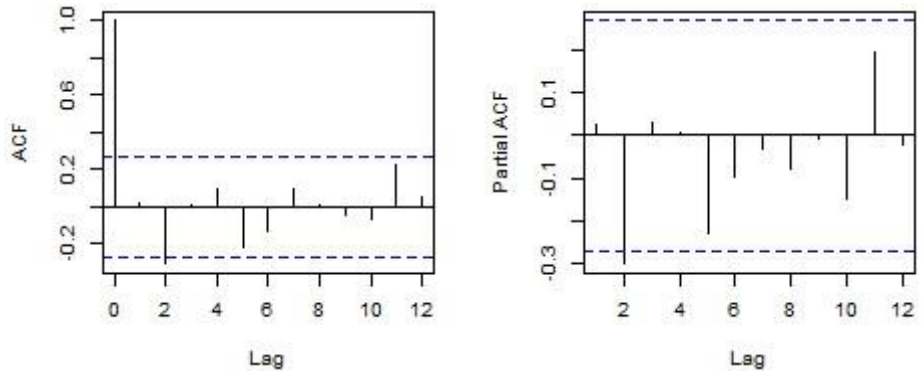


Figure 4.12:ACF and PACF plots for long rainfall season time series data

Based on ACF and PACF plots obtained in Figure 4.12, one can say that the test confirms the non-presence of autocorrelation pattern in series especially at lag 1. Only one spike at lag 2 shows an autocorrelation pattern but that can prevent that the time series is considered as white noise. Once the test of autocorrelation pattern in series is already done, then next step is to test for stationarity by Augmented Dickey-Fuller Test (ADF).

4.2.1.2 Stationary test of long rainfall season time series

By considering that we are comparing the forecasting performance of a linear model and a non-linear model, an important consideration when estimating a regression with ARMA

errors is that all the variables in the model must first be stationary. Thus, we first have to check that Y_t and all of the predictors $(X_{1,t}, \dots, X_{k,t})$ appear to be stationary. If we estimate the model when any of these are non-stationary, the estimated coefficients will not be consistent estimates (and therefore may not be meaningful).

We therefore first difference the non-stationary variables in the model. It is often desirable to maintain the form of the relationship between Y_t and the predictors, and consequently it is common to difference all the variables if any of them need differencing. The resulting model is then called a "model in differences," as distinct from a "model in levels," which is what is obtained when the original data are used without differencing.

If all the variables in the model are stationary, then we only need to consider ARMA errors for the residuals. It is easy to see that a regression model with ARIMA errors is equivalent to a regression model in differences with ARMA errors. For example, if a regression model with ARIMA (1,1,1) errors is differenced we obtain the model

$$Y_t' = \beta_1 X_{1,t}' + \cdots + \beta_k X_{k,t}' + \eta_t'$$

$$(1 - \emptyset_1 B)\eta_t' = (1 - \theta_1 B)\varepsilon_t$$

Where $Y_t' = Y_{t-1}, X_t' = X_{t,i} - X_{t-1,i}$ and $\eta_t' = \eta_t - \eta_{t-1}$, which is a regression model in differences with ARMA errors.

	ADF Test			
	Test Statistics	**Lag order**	**p-value**	**Comment**
Predictand	-3.4183	6	0.06262	Non-significant
SLP_PC19	-2.1236	6	0.5247	Non-significant
DMI_LR9	-3.5773	6	0.04347	Significant

Table 4.3:Unit Root and Stationarity tests for the time series involved in long rainfall model

The Table 4.3 depicts the ADF Test. The null hypothesis stipulate that the series is unit root non-stationary, and the alternative hypothesis stipulate that the series is unit root stationary. The probability value of the predictand (rainfall time series) (0.06262) and SLP_PC19 time series (0.5247) are greater than the level of significance at 5% indicating strong evidence against the null hypothesis. Once at least one of the time series under consideration is found to be non-stationary then one goes for differencing of all series to make it stationary.

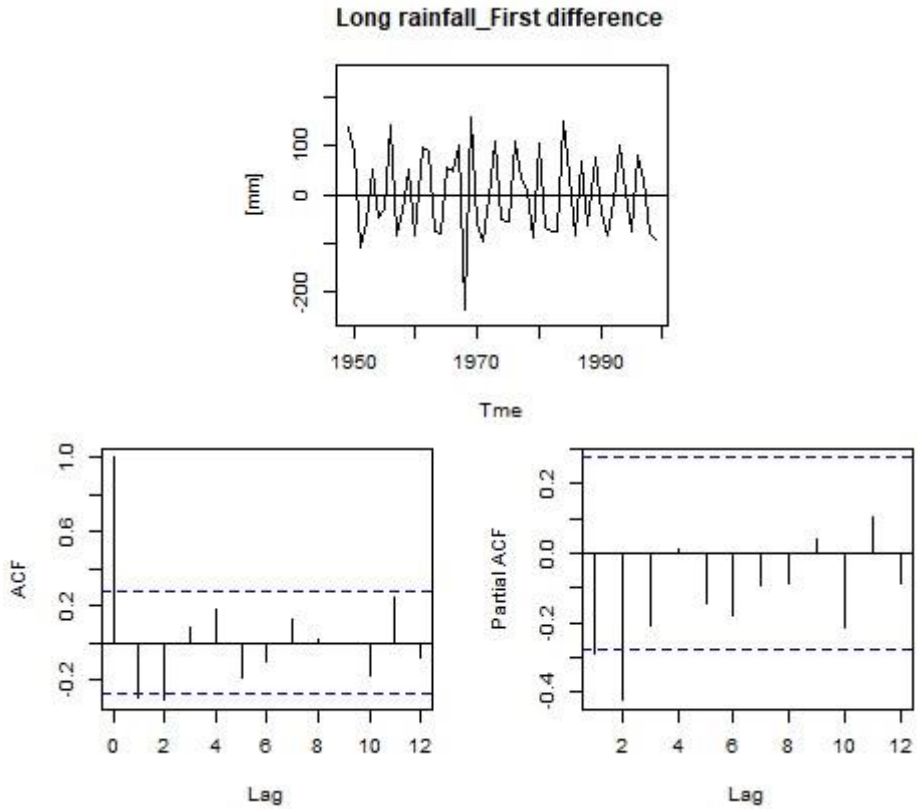


Figure 4.13: Autocorrelation check for white noise of first differenced long season rainfall time series

In this step, the residual is tested for evaluation purposes and goodness of the fit statistics is provided. Figure 4.13 shows the ACF and PACF for the residuals of MLR models and significant correlation are appear at a lag 1 and lag2. This show significant correlation values, confirms that the residuals for this model are not random. Once the test confirms the presence of autocorrelation pattern in series, then next step is to test for stationarity.

	ADF Test			
	Test Statistics	**Lag order**	**p-value**	**Comment**
Predictand	-4.1254	6	0.01106	Significant
SLP_PC19	-3.9185	6	0.01992	Significant
DMI_LR9	-3.6285	6	0.03929	Significant

Table 4.4: Unit Root and Stationarity tests for the time series involved in long rainfall model (after first difference)

The Table 4.4 shows that all-time series involved in constructing a multiple linear regression model for both statistical and predictive model are stationarity since the calculated probability value is less than the level of significance (5%).

4.2.1.3 Estimation and significance check of model parameters

Statistical and Predictive model				
Variable	Coefficient	Std. Error	t test	Probability
SLP-PC19	30.137	9.031	3.337	0.00162
DMI_LR9	-118.210	33.983	-3.479	0.00107

Table 4.5:Estimates of MLR model after first difference

Table 4.5 shows that the estimated coefficients are significantly different from zero. After getting regression parameters the next step is residuals analysis of multiple linear regression model. The model validation is concerned with checking the residual of the model to determine if the model contains any systematic pattern which can be removed to improve on the selected model.

4.2.1.4 Diagnostic checking process for the estimated model

	Original data			First differenced data		
	Summary statistics			Summary statistics		
Test Type	X-squared	Df	p-value	X-squared	Df	p-value
Box-Pierce and Ljung-Box	0.11576	1	0.7337	4.1132	1	0.04255

Table 4.6:Box-Pierce and Ljung-Box Test for long rainfall season model

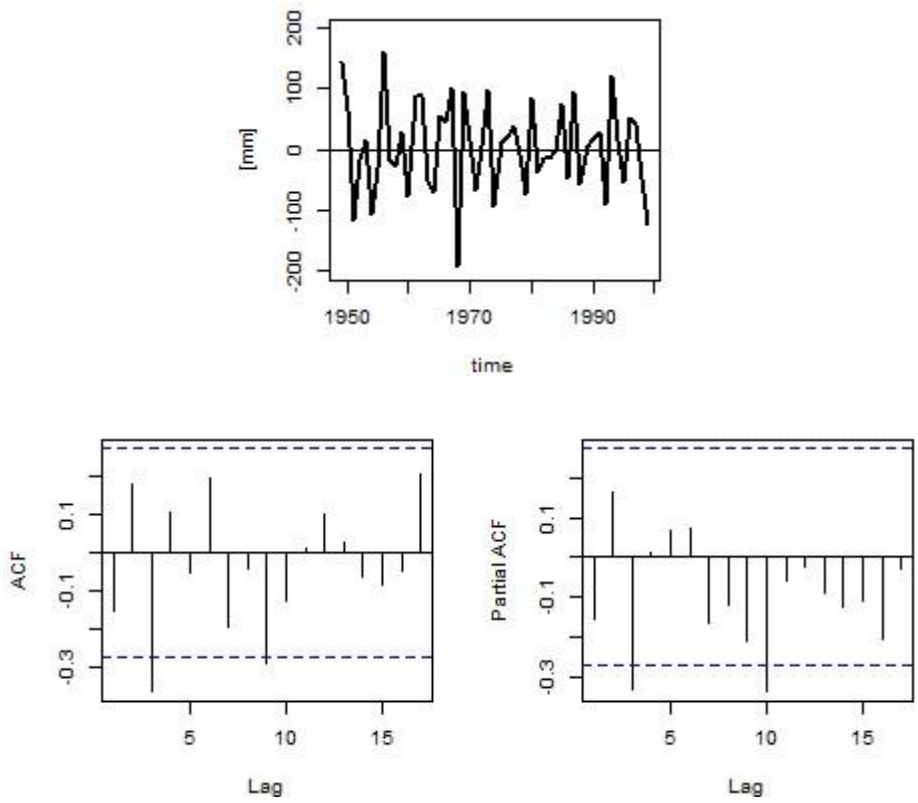


Figure 4.14:Diagnostic of residual plots for MLR (predictive model)

The Figure 4.14 shows a diagnostic of a MLR model which involves checking of model residuals. For the Figure 4.14 on the top, an inspection of the residuals time series plot shows some outliers. In Figure 4.14 in the bottom, the ACF and PACF of residuals shows apparent departure from the model assumptions at lag1 and lag2. For the MLR model constructed based on original data in Table 4.6, the Q statistics is never significant at the lags shown. The Ljung-Box test for this model gives a chi-squared value of 0.11576, leading to a p- value of 0.7337. The bell-shape feature is clearly noticed in Appendix E.1; indicating that the residuals are normally distributed for differentiated data. This is approved by Jarque Bera test results in Appendix E.2. In addition, all series in this model should be differentiated in order to meet the stationarity assumption. For differentiated data, the Ljung-Box test for this model gives a chi-squared value of 4.1132, leading to a p- value of 0.04255, a further indication that the model has not captured the dependence in the time series.

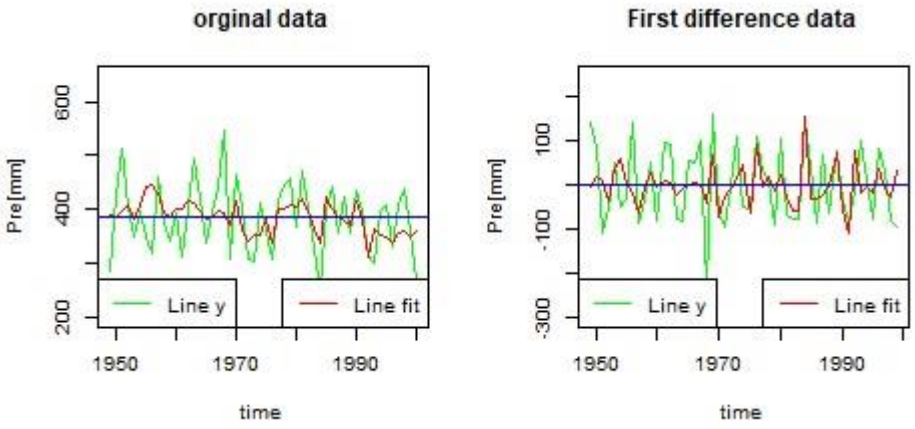


Figure 4.15:Comparison graph of observed short rainfall season vs Predicted short rainfall season (Statistical and Predictive model)

To evaluate the multiple linear regression model for long rainfall season performance in East Africa the prediction performance was performed. Figure 4.15 shows a comparison between the real values and the ones resulted from the developed MLR model for the period between 1949 and 2000. For MLR model on differentiated data, we obtain the graphical plot for the actual pitch series versus the predicted pitch series and from the visual inspection of the plot it is quite evident that the chosen model is good enough as the predicted. The correlation coefficient (r) for models developed based on time series original data and first differentiated data are 0.45 and 0.53 respectively. Same case for both models the RMSE were found to be 56.66 and 74.2 respectively.

4.2.2 Regression models forecasting short rainfall season in East Africa

The system which impacts the short rainfall season in East Africa and the variables involved in the construction of short rainfall season model are described in chapter three. As the same case as for long rainfall season model, the development of leading variables with three months' time steps was done and the total number of independent variables becomes 72. The multiple linear regression analysis was carried out by considering all teleconnections as predictors and long rainfall season time series data from CRU as predictand. The first step was to select teleconnections which are significantly correlated with rainfall time series at 5% level of significance. To overcome the multi-collinearity problem, one of the measures was to drop the unimportant variables *i.e.* the variables which explaining less variations in dependent variables need to be drop from the model, the dropping of variable was done through the collinearity and

stepwise VIF selection. Finally, the stepwise regression analysis was carried out to fit the model.

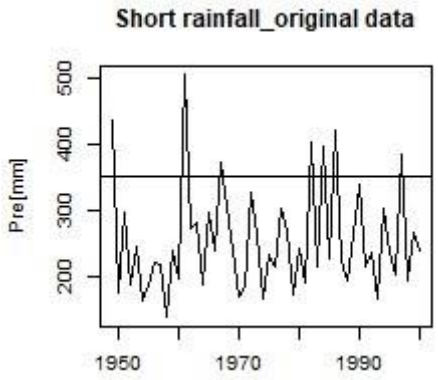


Figure 4.16:Time series for short rainfall time series

4.2.2.1 Estimation and significance check of model parameters

Statistical model				
Variable	Coefficient	Std. Error	t test	Probability
Constant	251.134	9.465	24.532	< 2e-16
SOI_SR3	-33.682	10.319	-3.264	0.00203
SAOD_SR0	47.331	17.480	2.708	0.00936
SST_PC13	-17.067	9.668	-1.765	0.08386
Predictive model				
Variable	Coefficient	Std. Error	t test	Probability
Constant	251.00	10.31	24.355	<2e-16
SOI_SR3	-25.46	10.53	-2.418	0.0194
SST_PC13	-16.43	10.71	-1.535	0.1313

Table 4.7:Estimates of regression model for short rainfall original time series data

The table 4.7 shows that the unexplained or non-significant variables are dropped from the model so that one can get maximum error degrees of freedom. After stepwise regression analysis, we obtained in total three independent variables for statistical model which are the Southern Oscillation Index three months lead (SOI_SR3), the South Atlantic Ocean Dipole current time series (SAOD_SR0) and the first principal component of Indian Sea Surface Temperature three months lead (SST_PC13). For predictive model two independent variables (SOI_SR3, SST_PC13) are selected. Among three variables entered in construction of multiple linear regression model for short rainfall season two of them are significant (SOI_SR3, SAOD_SR0) and the remaining one is not significant (SST_PC13) at 5% level of significant).

After getting regression parameters the next step is residuals analysis for multiple linear regression model analysis.

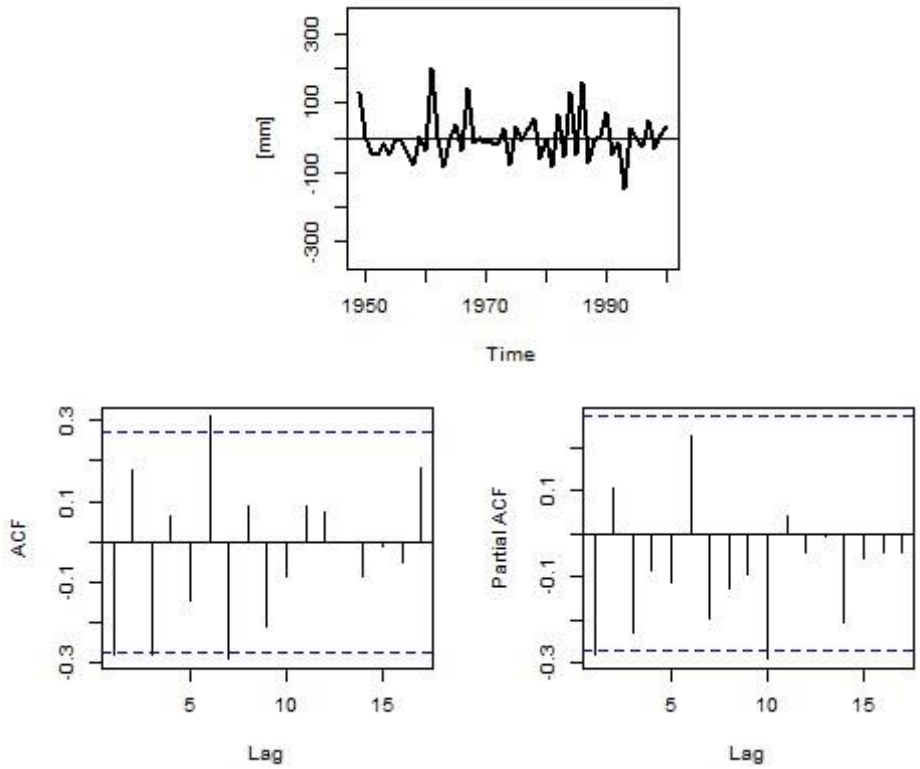


Figure 4.17:Diagnostic of residual plots for MLR (Statistical model)

Based on ACF and PACF plots obtained in Figure 4.17, one can say that the test confirms the presence of autocorrelation pattern in residuals series. This shows significant correlation values and confirms that the residuals for this model are not random, which means that the model is not a good fit for the series and essential components have been omitted from the models. Even if it is like that, the most important aspect when we are in time series analysis is the check of stationarity for the time series data.

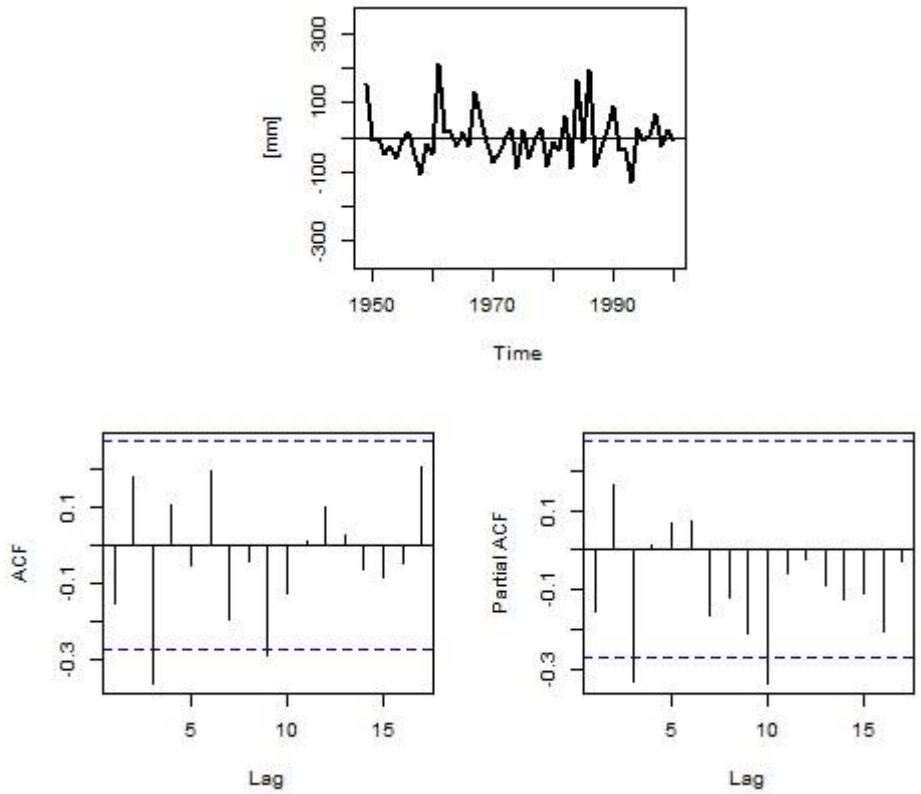


Figure 4.18:Diagnostic of residual plots for MLR (predictive model)

Based on ACF and PACF plots obtained in Figure 4.18, one can say that the test confirms the non-presence of autocorrelation pattern in series especially at lag 1. Only two spikes at lags 3 and 8 show an autocorrelation pattern but that can prevent that the time series is considered as white nose. Once the test of autocorrelation pattern in series is already done, then next step is to test for stationarity by ADF test.

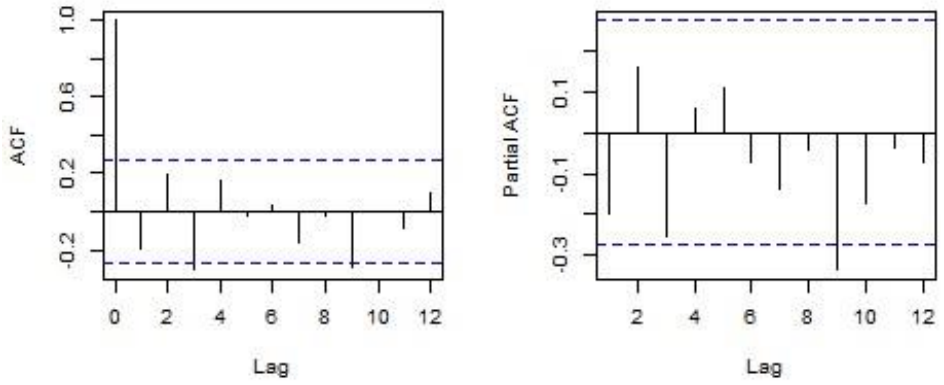


Figure 4.19:ACF and PACF plots for short rainfall season time series

Based on ACF and PACF plots obtained in Figure 4.19, one can say that the test confirms the non-presence of autocorrelation pattern in series especially at lag 1. Two spikes at lag 3 and lag 9 show an autocorrelation pattern and this is enough to suspect that the time series is considered is not white nose. Once the test of autocorrelation pattern in series is already done, then next step is to test for stationarity by ADF test.

4.2.2.2 Stationary test of short rainfall season time series data

In this step all-time series variables which are involved in model's construction for both statistical model and predictive model are checked for stationarity using ADF test. The results from the test are presented in table 4.8.

	ADF Test			
	Test statistics	**Lag order**	**p-value**	**Comment**
Predictand	-2.7374	4	0.2778	Non-significant
SOI_SR3	-2.4753	4	0.3832	Non-significant
SAOD_SR0	-2.8712	4	0.2240	Non-significant
SST_PC13	-2.3983	4	0.4142	Non-significant

Table 4.8: Unit Root and Stationarity tests for the time series involved in short rainfall model

From Table 4.8, it is shown that the time series are non-stationary. The probability value of short rainfall time series (0.2778), SOI_SR3 time series (0.3832), SADO_SR0 time series (0.2240) and SST_PC13(04142) are greater than the level of significance at 5% indicating strong evidence against the null hypothesis. Once at least one of the time series under consideration is found to be non-stationary then one goes for differencing of all series to make it stationary.

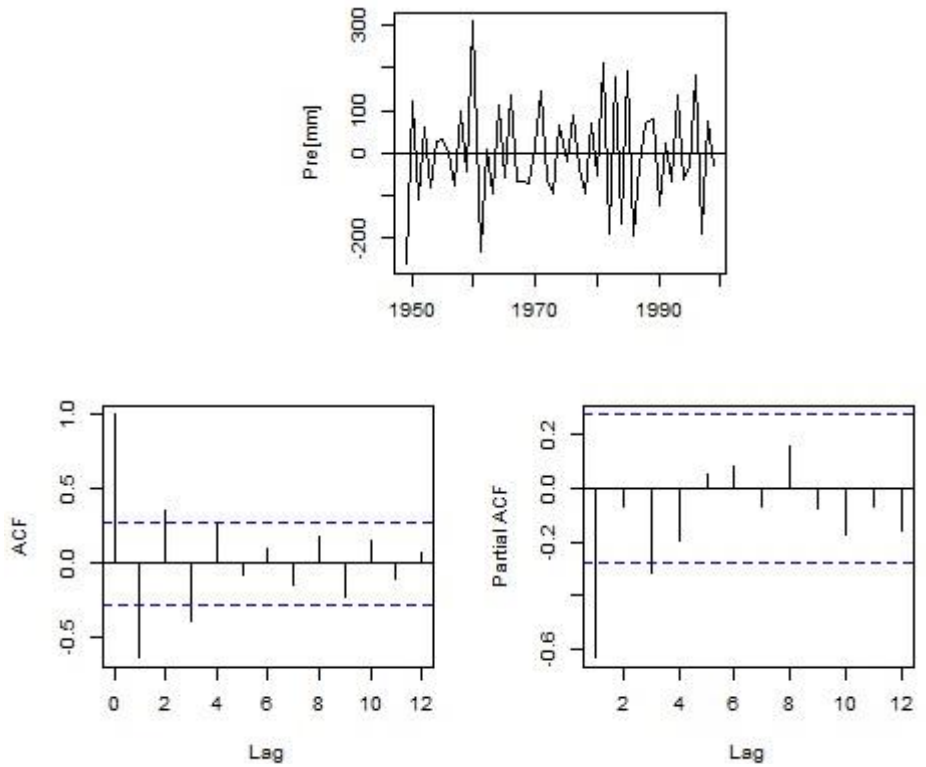


Figure 4.20:Autocorrelation check for white noise of first differenced short rainfall season time series

Based on ACF and PACF plots obtained in Figure 4.20, one can say that the second difference time series confirms the presence of autocorrelation pattern in series and the next step is to check stationarity using ADF test.

	ADF Test			
	Test statistics	**Lag order**	**p-value**	**Comment**
Predictand	-5.0588	4	0.0100	Significant
SOI_SR3	-3.828	4	0.0238	Significant
SAOD_SR0	-4.2275	4	0.0100	Significant
SST_PC13	-5.3555	4	0.0100	Significant

Table 4.9: Unit Root and Stationarity tests for the time series involved in short rainfall model (after first difference)

The Table 4.9 shows that all the time series involved in constructing a multiple linear regression model for both statistical and predictive model are stationarity at the first difference since the calculated probability values are less than the level of significance (5%).

4.2.2.3 Estimation and significance check of model parameters

Statistical model				
Variable	Coefficient	Std. Error	t-test	Probability
SOI_SR3	-44.741	11.769	-3.802	0.000406
SAOD_SR0	24.761	20.758	1.193	0.238796
SST_PC13	-6.0080	11.838	-0.508	0.614081
Predictive model				
Variable	Coefficient	Std. Error	t-test	Probability
SOI_SR3	-34.47	10.73	-3.211	0.00234
SST_PC13	-6.500	12.61	-0.516	0.60845

Table 4.10:Estimates of MLR model after first difference

Table 4.10 shows that after stepwise regression analysis, we obtained in total three significant independent variables for the statistical model and two significant variables for the predictive model. Among three variables entered in construction of multiple linear regression statistical model for short rainfall season one of them is significant (SOI_SR3) and the remaining two are not significant (SST_PC13 and SAOD_SR0) at 5% level of significant. For predictive model only SOI_SR3 is significant. The model validation is concerned with checking the residual of the model to determine if the model contains any systematic pattern which can be removed to improve on the selected model.

4.2.2.4 Diagnostic checking process for the estimated model

		Original data			**First differenced data**		
		Summary statistics			Summary statistics		
	Test Type	X-squared	df	p-value	X-squared	df	p-value
Statistical model	Box-Pierce and Ljung-Box	4.096	1	0.04298	19.058	1	1.268e-05
Predictive model	Box-Pierce and Ljung-Box	1.1961	1	0.2741	17.485	1	2.896e-05

Table 4.11:Box-Pierce and Ljung-Box Test for short rainfall season model

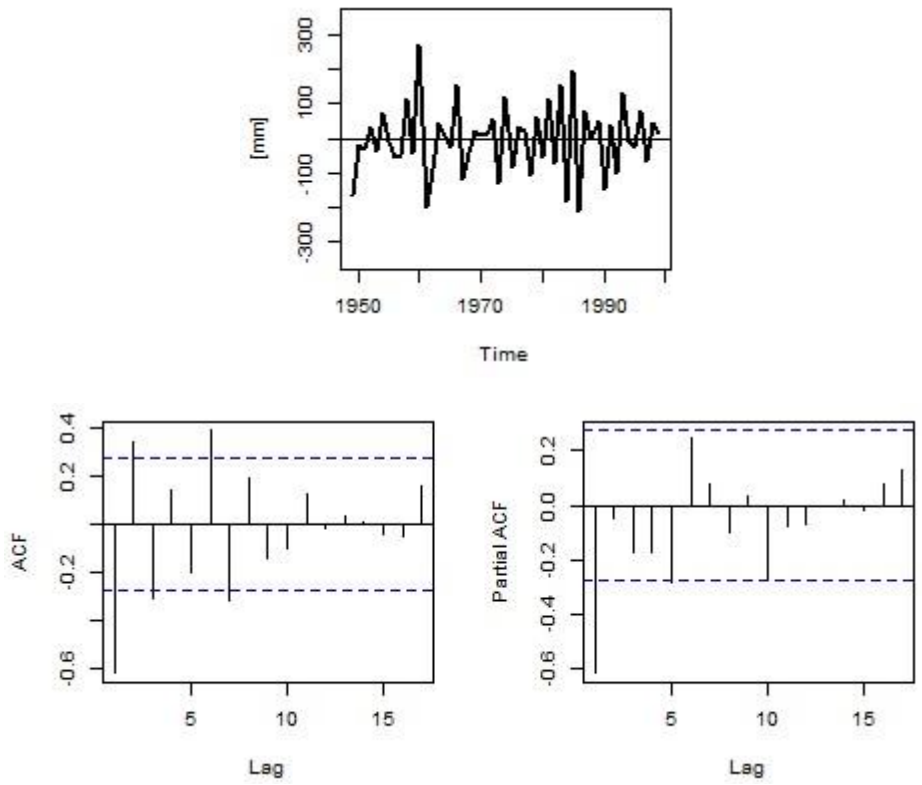


Figure 4.21:Diagnostic of a multiple linear regression model for short rainfall season

The ACF of the residuals shows apparent departure from the model assumptions in lag1, lag2 and lag3 in Figure 4.21. For the multiple linear regression model constructed based on original data, the Q statistics is significant at the lags shown. The bell-shape feature is clearly noticed in Appendix F.1; indicating that the residuals are normally distributed for differentiated data. This is approved by Jarque Bera test results in Appendix F.2. In Table 4.11, the Ljung-Box test for the statistical model gives a chi-squared value of 4.096, leading to a p- value of 0.04298. This shows that the fitted short rainfall season statistical model provides not good fit for the entire time series. For differentiated data, the Ljung-Box test for this model gives a chi-squared value of 19.058, leading to a p- value of 1.268e-05, a further indication that the model has not captured the dependence in the time series.

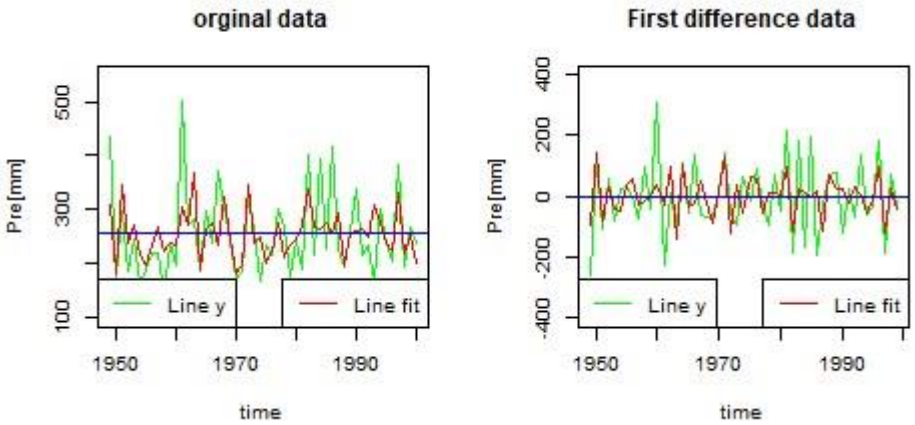


Figure 4.22:Comparison graph of observed short rainfall season vs Predicted short rainfall season (Statistical model)

To evaluate the multiple linear regression model for short rainfall season performance in East Africa the prediction performance was performed. Figure 4.22 shows a comparison between the real values and the ones resulted from the developed MLR model for the period between 1949 and 2000. For MLR model on differentiated data, we obtain the graphical plot for the actual pitch series versus the predicted pitch series and from the visual inspection of the plot it is quite evident that the chosen model is good enough as the predicted. The correlation coefficient (r) for models developed based on time series original data and first differentiated data are 0.56 and 0.55 respectively. Same case for both models the RMSE were found to be 65.28 and 104.04 respectively.

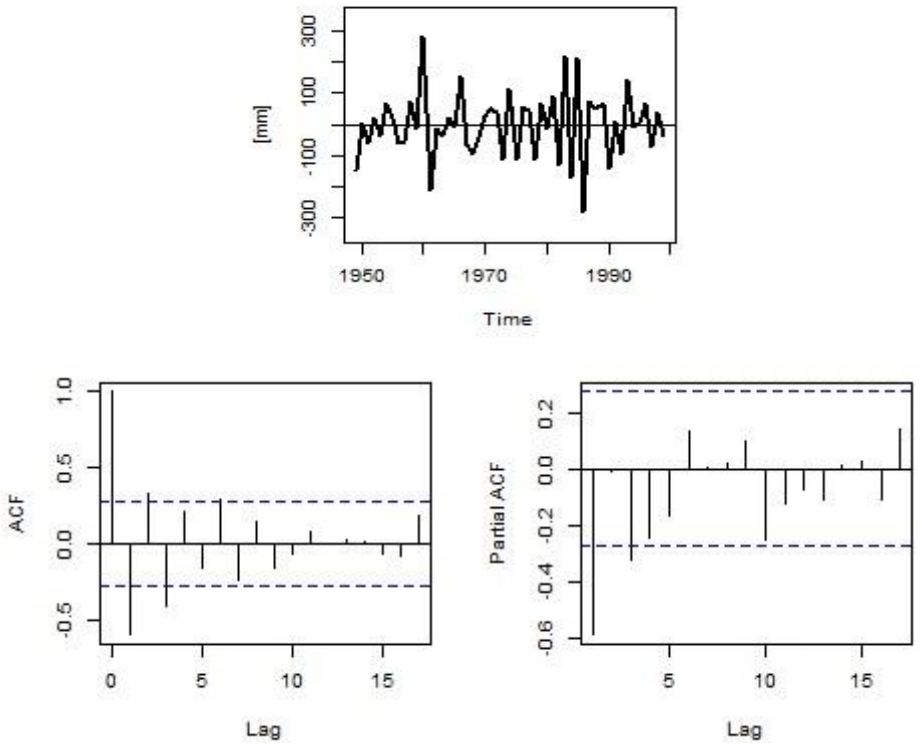


Figure 4.23:Diagnostic of a MLRM for short rainfall season

The ACF of the residuals shows apparent departure from the model assumptions in lag1 and lag2 in Figure 4.23. For the multiple linear regression model constructed based on original data, the Q statistics is never significant at the lags shown. The bell-shape feature is clearly noticed in Appendix G.1; indicating that the residuals are normally distributed for differentiated data. This is approved by Jarque Bera test results in Appendix G.2. In Table 4.11, The Ljung-Box test for this model gives a chi-squared value of 1.1961, leading to a p- value of 0.2741. This shows that the fitted long rainfall season model provides good fit for the entire time series. But this is not the case because all series in this model should be differentiated in order to meet the stationarity assumption. For differentiated data, the Ljung-Box test for this model gives a chi-squared value of 17.485, leading to a p- value of 2.896e-05, a further indication that the model has not captured the dependence in the time series.

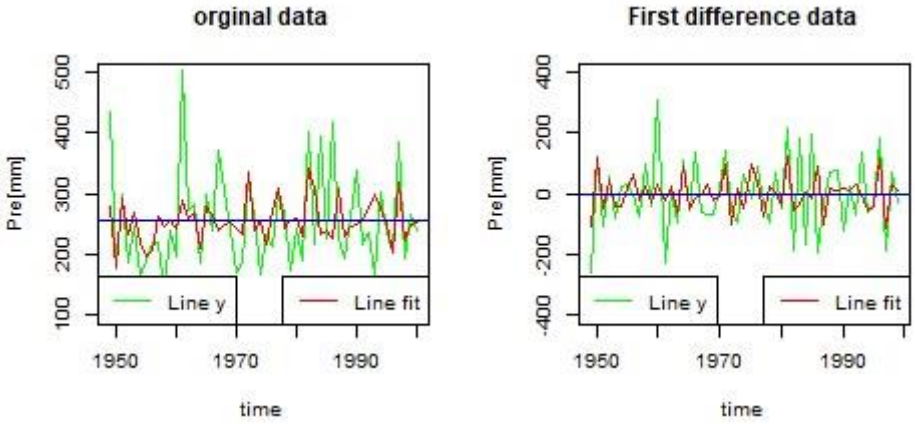


Figure 4.24:Comparison graph of observed short rainfall season vs Predicted short rainfall season (Predictive model)

To evaluate MLR model for short rainfall season performance in East Africa the prediction performance was performed. Figure 4.24 shows a comparison between the real values and the ones resulted from the developed MLR model for the period between 1949 and 2000.The correlation coefficient (r) for models developed based on time series original data and first differentiated data are 0.43 and 0.48 respectively. Same case for both models the RMSE were found to be 71.4 and 106.18 respectively.

4.2.3 RARIMAE model for forecasting long rainfall season in East Africa

The principal step in Box-Jenkins ARIMA model building is identification of the model. Different orders of Autoregressive (AR) and Moving Average (MA) parameters p and q are considered and combination of the order which yields maximum log-likelihood and lowest values of Akaike Information Criteria (AIC) and Bayesian Information Criteria (BIC) are considered as final model orders. The main task of automatic ARIMA forecasting (Hyndman and Khandakar (2008)) is selecting appropriate model order, that is the value p, q, d . If d is known, we can select the order of p, q via information criterion such as AIC:

$$AIC = -2\log(L) + 2(p + q + k)$$

Where $k = 1$ if $c \neq 0$ and 0 otherwise, and L is the maximum likelihood of the model fitted on the differenced data.

4.2.3.1 Long rainfall season model identification process

The model identification process is where the form and order of tentative models are basically selected. The form and order of these models are picked from the sample autocorrelation function and partial autocorrelation function of the observed series. However, such observed series must be stationary before tentative models are selected.

Models	Log likelihood	AIC	BIC
ARIMA(2,0,2) with non-zero mean	-279.84	575.6812	591.2911
ARIMA(0,0,0) with non-zero mean	**-283.71**	**575.4227**	**583.2277**
ARIMA(1,0,0) with non-zero mean	-283.64	577.277	587.0333
ARIMA(0,0,1) with non-zero mean	-283.54	577.0773	586.8335
ARIMA(0,0,0) with zero mean	-362.91	731.8153	737.6691
ARIMA(1,0,1) with non-zero mean	-282.90	577.7955	589.5030

Table 4.12:Estimated candidate ARIMA models for long rainfall time series

The Table 4.12 gives the maximum likelihood estimates, their AIC and BIC for ARIMA (0,0,0) model based on automatic ARIMA forecasting process. The estimated model is a "Regression with ARIMA (0,0,0) errors" which indicates no autoregressive or moving average pattern in the residuals. We can also see this by looking at an ACF plot of the residuals (Figure 4.11). Once the model order was determined then, next step is to go for parameter estimation of the model by maximum likelihood estimation method which is the second step in Box-Jenkins ARIMA model building procedure. The results of parameter estimation of ARIMA (0,0,0) are given in Table 4.13.

4.2.3.2 Estimation and significance check of model parameters

Coefficients	Estimate	Standard error	p-value
Intercept	362.9043	11.2444	2.2e-16
SLP_PC19	21.4984	7.9496	0.006844314
DMI_LR9	-62.8151	23.4012	0.007268809

Table 4.13:Estimates of RARIMAE model (Original data)

Table 4.13 shows that there is no difference between MLRM estimated parameters (Table 4.2) and the ones estimated by RARIMAE (Table 4.13). Therefore, we obtain two highly significant independent variables for both statistical and predictive model which have participated in model construction. After getting RARIMAE model parameters the next step is the model residuals analysis (Figure 4.25).

4.2.3.3 Diagnostic checking process for the estimated model

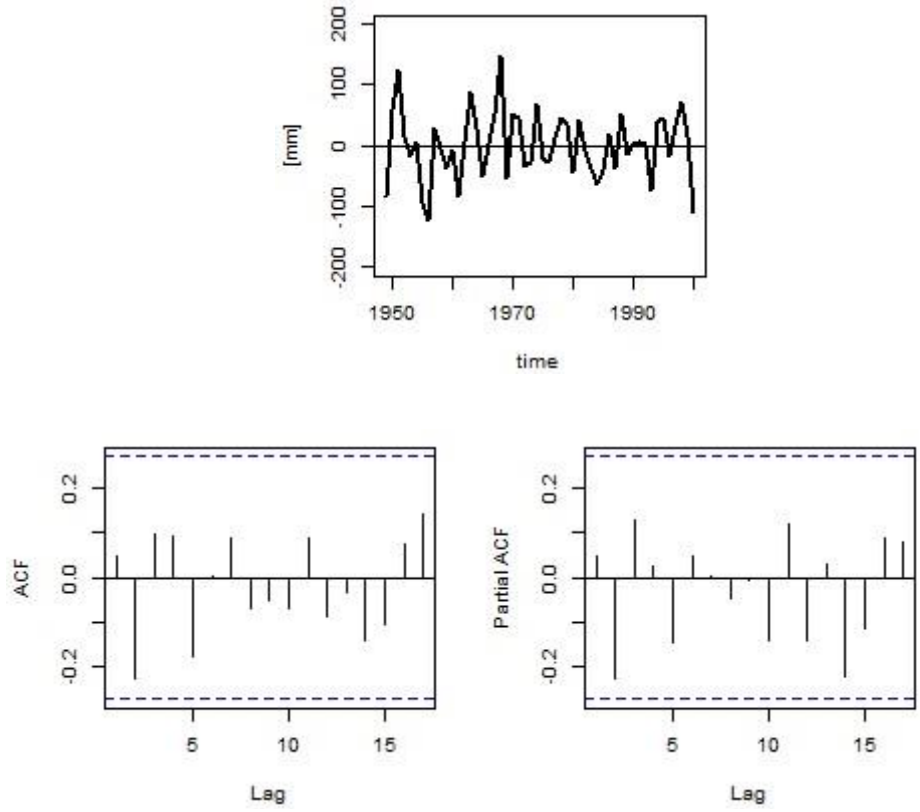


Figure 4.25:Diagnostic of residual plots for RARIMAE

Based on ACF and PACF plots obtained in Figure 4.25, one can say that the test confirms the non-presence of autocorrelation pattern in series. This is a good sign when the diagnostic of every model is done but the most important aspect when we are in time series analysis is the check of autocorrelation in the time series data. After running a regression with ARIMA errors by use of differentiated data the candidate ARIMA models are found in Table 4.14.

Models	Log likelihood	AIC	BIC
ARIMA(0,0,0) with non-zero mean	-292.03	592.0639	599.7912
ARIMA(1,0,0) with non-zero mean	-289.53	589.0601	598.7193
ARIMA(0,0,0) with zero mean	-292.03	590.0689	595.8644
ARIMA(2,0,0) with non-zero mean	-282.89	577.7704	589.3614
ARIMA(3,0,0) with non-zero mean	-281.25	576.5067	590.0295
ARIMA(4,0,0) with non-zero mean	-281.19	578.3864	593.8410
ARIMA(3,0,0) with zero mean	**-281.26**	**574.5191**	**586.1100**
ARIMA(2,0,0) with zero mean	-282.89	575.7744	587.4335
ARIMA(4,0,0) with zero mean	-281.20	576.4038	589.9266

Table 4.14: Estimated candidate ARIMA models for long rainfall time series (first differenced data)

The table 14 gives the maximum likelihood estimates, their AIC and BIC for ARIMA (3,0,0) model based on automatic ARIMA forecasting. The estimated model is a "Regression with ARIMA (3,0,0) errors" which indicates the presence of autoregressive pattern in the residuals. We can also see this by looking at an ACF plot of the residuals (Figure 4.13). ARIMA (3,0,0) means that the predicted value for the next rainfall season depending on the 3 seasonal data of rainfall before, 0 seasonal data of rainfall earlier error.

4.2.3.4 Estimation and significance check of Tentative model parameters

Coefficients	Estimate	Standard error	p-value
θ_1	-0.6700	0.1512	9.407521e-06
θ_2	-0.6922	0.1527	5.842725e-06
θ_3	-0.2801	0.1504	6.258247e-02
SLP_PC19	25.2124	8.0376	1.708049e-03
DMI_LR9	-97.9991	33.6040	3.542144e-03

Table 4.15:Estimates of RARIMAE model (first differentiated data)

From Table 4.15, all coefficients of estimated parameters are significantly different from zero. There is a stationarity in θ_1, θ_2 and θ_3 as the absolute values of their estimates are far from 1. After getting model parameters the next step is residuals analysis for RARIMAE model analysis.

4.2.3.5 Diagnostic checking process for the estimated model

	Original data			**First differenced data**		
	Summary statistics			Summary statistics		
Test Type	X-squared	Df	p-value	X-squared	Df	p-value
Box-Pierce and Ljung-Box	10.338	14	0.7371	8.4027	14	0.8673

Table 4.16:Box-Pierce and Ljung-Box Test for RARIMAE long rainfall season model

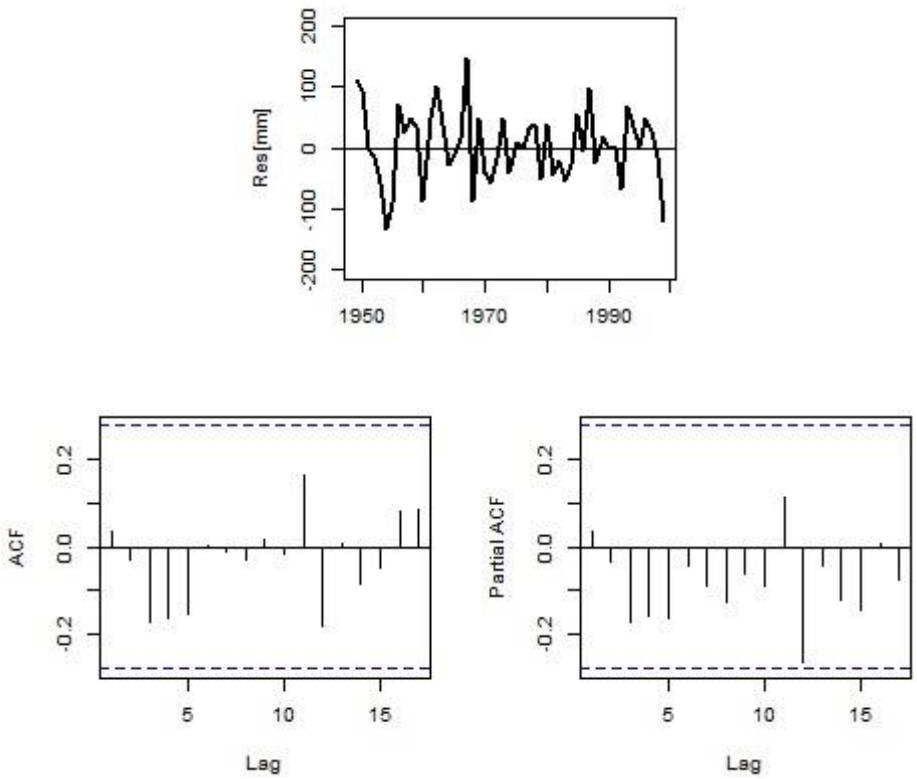


Figure 4.26:Diagnostic of residual plots for RARIMAE (first difference data)

From theory, it is expected that $\frac{\mathrm{N}}{4} = \frac{52}{4} = 13$ autocorrelation functions of residuals out of which less than 5% spikes should be noticed for the residuals to be accepted as a white noise. However, from Figure 4.26, almost all the spikes of the ACF and PACF plots all lie within the confidence bounds suggesting that the residuals are white noise. For differentiated data, the normal Q-Q plot seems good because most of the dataset lie on the straight line (Appendix H). The bell-shape feature is clearly noticed in Appendix H.1, indicating that the residuals are normally distributed, and this is approved by Jarque Bera test results in Appendix H.2. A further analysis was conducted to ascertain the certainty of the residuals being white noise. From table 4.16 a Box-Ljung test was reported a $\chi^2 = 10.338(\mathrm{df} = 14)$ with a large $\mathrm{p} - \mathrm{value} = 0.7371$,suggesting that the residuals from a model computed based on original data are white noise. For differenced data, a Box-Ljung test was reported a $\chi^2 = 8.4028(\mathrm{df} = 14)$ with a large $\mathrm{p} - \mathrm{value} = 0.8673$,suggesting that the residuals from the model are also white noise.

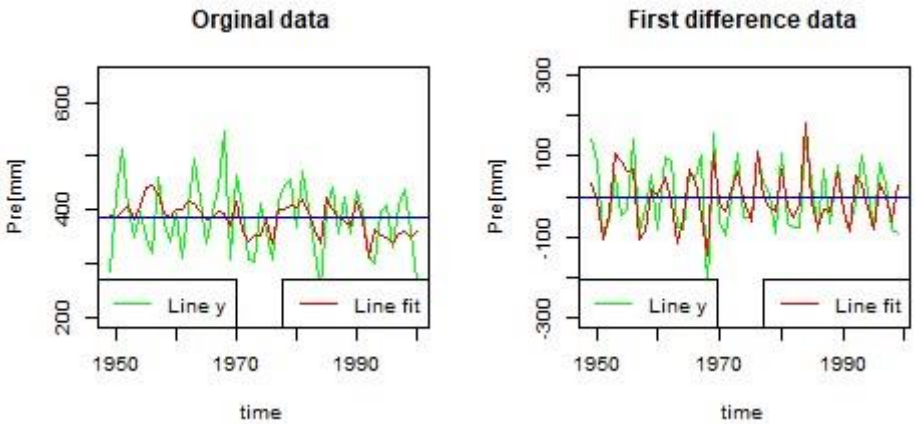


Figure 4.27:Observed long rainfall season vs Predicted long rainfall season

To evaluate the multiple linear regression model for long rainfall season performance in East Africa the prediction performance was performed. Figure 4.27 shows a comparison between the real values and the ones resulted from the developed ARIMA model for the period between 1949 and 2000. ARIMA (3,0,0) model on differentiated data, we obtain the graphical plot for the actual pitch series versus the predicted pitch series and from the visual inspection of the plot it is quite evident that the chosen model is good enough as the predicted. The correlation coefficient (r) for models developed based on time series original data and first differentiated data are 0.45 and 0.73 respectively. Same case for both models the RMSE were found to be 56.66 and 59.47 respectively.

4.2.4 RARIMAE model for forecasting short rainfall season in East Africa

The same steps for model construction are followed as it is done for long rainfall season. The only difference between these two seasons is that the number of variables selected for model prediction in statistical model and predictive model are different while for long rainfall season same variables are selected for both statistical and predictive model.

4.2.4.1 Short rainfall season RARIMAE statistical model

Before performing a RARIMAE statistical model 72 variables were candidates in model construction. After variables selections based on criteria described in chapter 3 of this study, a RARIMAE model is developed using a combination of exogenous variables used to develop a multiple linear regression model in the same season (Table 4.7). Like other ARIMA models, RARIMAE models follows step in Box-Jenkins ARIMA model building, which is consist by model identification, parameters estimation and diagnostic and checking.

4.2.4.1.1 Model identification for short rainfall season RARIMAE statistical model

The form and order of these models are picked from the sample autocorrelation function and partial autocorrelation function of the observed series. However, such observed series must be stationary before tentative models are selected.

Models	Log likelihood	AIC	BIC
ARIMA(0,0,0) with non-zero mean	-291.08	592.1601	601.9163
ARIMA(1,0,0) with non-zero mean	-288.48	588.9513	602.6588
ARIMA(0,0,1) with non-zero mean	-288.75	589.499	603.2064
ARIMA(0,0,0) with zero mean	-362.62	733.2368	741.0418
ARIMA(2,0,0) with non-zero mean	-288.27	590.5327	604.1914
ARIMA(1,0,1) with non-zero mean	**-287.31**	**588.6174**	**602.2762**
ARIMA(2,0,1) with non-zero mean	-287.25	590.5054	606.1153
ARIMA(1,0,2) with non-zero mean	-287.23	590.4506	606.0605
ARIMA(0,0,2) with non-zero mean	-288.74	591.4889	605.1476
ARIMA(1,0,1) with zero mean	-294.91	601.8265	613.5340

Table 4.17:Estimated candidate ARIMA models for short rainfall time series (Original data)

The Table 4.17 gives the maximum likelihood estimates, their AIC and BIC for ARIMA (1,0,1) model based on automatic ARIMA forecasting. The estimated model is a "Regression with ARIMA (1,0,1) errors" which indicates the presence of autoregressive and moving average pattern in the residuals. We can also see this by looking at an ACF plot of the residuals (Figure 4.17). Once the model order was determined then, next step is to go for parameter estimation of the model by maximum likelihood estimation method which is the second step in Box-Jenkins ARIMA model building procedure. The results of parameter estimation of ARIMA (1,0 ,1) are given in Table 4.18.

4.2.4.1.2 Estimation and significance check of Tentative model parameters

Coefficients	Estimate	Standard error	P-value
θ_1	-0.8742	0.1192	2.193801e-13
$\emptyset_1$	0.6806	0.1816	1.789004e-04
Intercept	251.2248	7.5807	0.000000e+00
SOI_SR3	-29.0495	9.0147	1.270919e-03
SAOD_SR0	47.8195	15.1314	1.576163e-03
SST_PC13	-22.1195	8.3709	8.231476e-03

Table 4.18:Estimates of RARIMAE model (Original data)

From table 4.18, all coefficients of estimated parameters are highly significantly different from zero. There is no stationarity in θ_1 as the absolute value of its estimate is not far from 1. After getting regression parameters the next step is residuals analysis for RARIMAE model analysis.

4.2.4.1.3 Diagnostic checking process for the estimated model

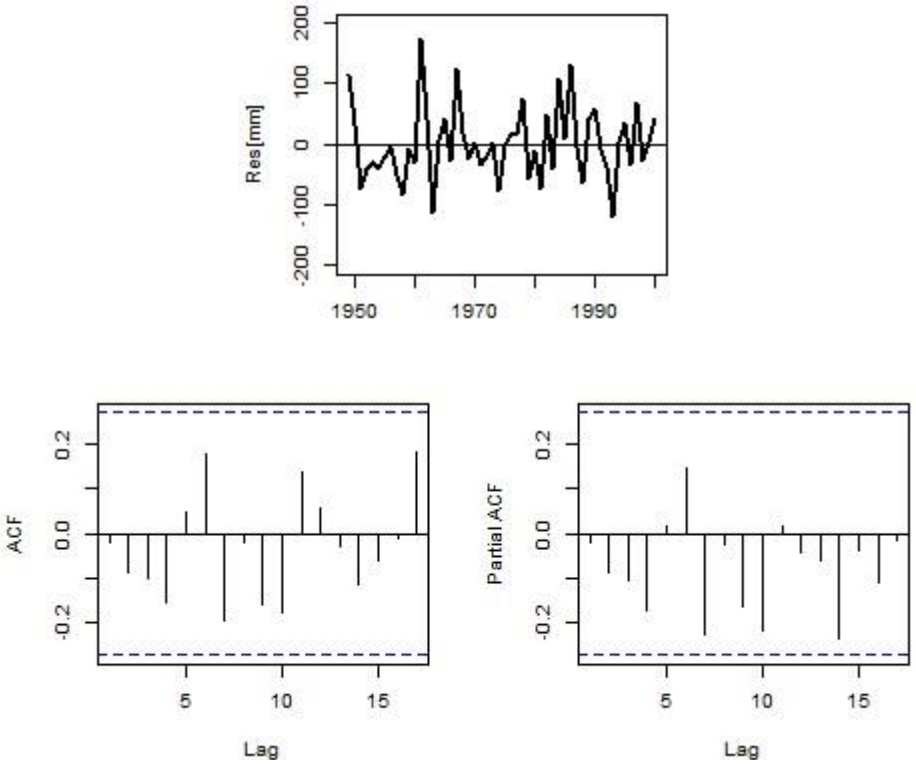


Figure 4.28:Diagnostic of residual plots for RARIMAE

Models	Log likelihood	AIC	BIC
ARIMA(0,0,0) with non-zero mean	-307.75	625.4914	635.2476
ARIMA(1,0,0) with non-zero mean	-292.95	597.904	609.6115
ARIMA(0,0,0) with zero mean	-307.76	623.5134	631.3183
ARIMA(2,0,0) with non-zero mean	-292.56	599.1157	612.7744
ARIMA(1,0,0) with zero mean	**-292.97**	**595.9311**	**605.6873**
ARIMA(2,0,0) with zero mean	-292.57	597.1394	608.8469

Table 4.19:Estimated candidate ARIMA models for short rainfall time series (differenced data)

The Table 4.19 gives the maximum likelihood estimates, their AIC and BIC for ARIMA (1,0,0) model based on automatic ARIMA forecasting. The estimated model is a "Regression with ARIMA (1,0,0) errors" which indicates the presence of autoregressive pattern in the residuals. We can also see this by looking at an ACF plot of the residuals (Figure 4.21).

ARIMA (1,0,0) means that the predicted value for the next year depending on the data 1 rainfall season before, 0 seasonal rainfall earlier error.

4.2.4.1.4 Estimation and significance check of Tentative model parameters

Coefficients	Estimate	Standard error	P-value
θ_1	-0.6935	0.1022	1.137357e-11
SOI_SR3	-28.9779	9.7741	3.029093e-03
SAOD_SR0	30.4844	16.1511	5.910024e-02
SST_PC13	-22.6784	9.0045	1.178359e-02

Table 4.20:Estimates of RARIMAE model (First differentiated data)

From Table 4.20, all coefficients of estimated parameters are significantly different from zero. There is a in θ_1 stationarity as the absolute values of its estimate is far from 1. After getting regression parameters the next step is residuals analysis for RARIMAE model analysis.

4.2.4.1.5 Diagnostic checking process for the estimated model

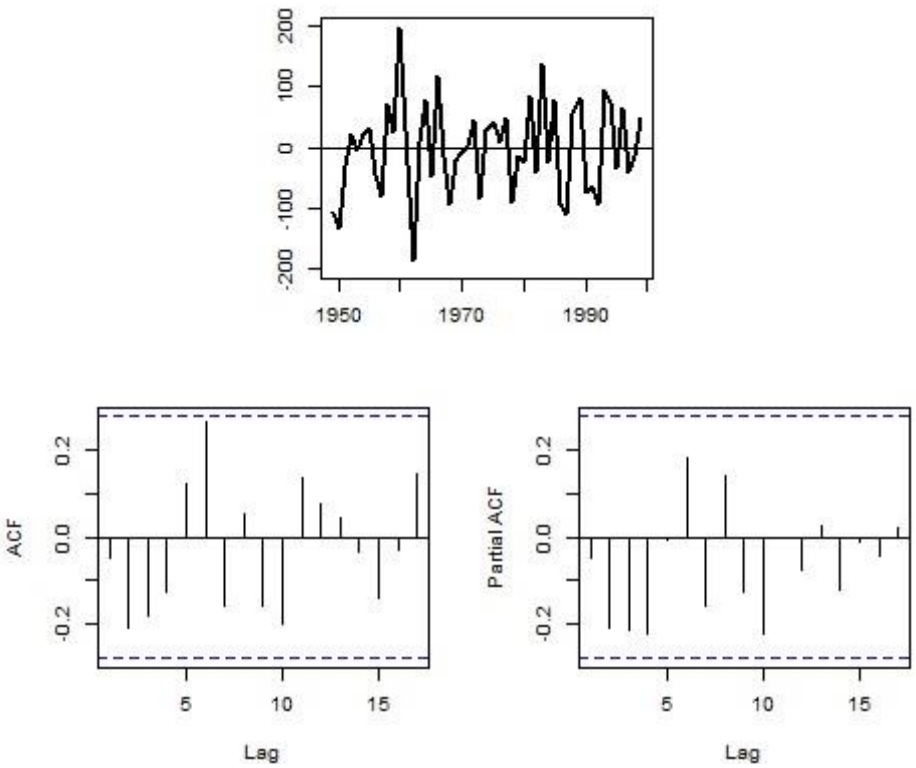


Figure 4.29:Diagnostic of residual plots for RARIMAE (first difference data)

4.2.4.2 Short rainfall season RARIMAE predictive model

Before performing a RARIMAE statistical model 58 variables were candidates in model construction. After variables selections based on criteria described in chapter 3 of this study, a RARIMAE model is developed using a combination of exogenous variables used to develop a multiple linear regression model in the same season (Table 4.7). Like other ARIMA models, RARIMAE models follows step in Box-Jenkins ARIMA model building, which is consist by model identification, parameters estimation and diagnostic and checking.

4.2.4.2.1 Model identification for short rainfall season RARIMAE predictive model

The form and order of these models are picked from the sample autocorrelation function and partial autocorrelation function of the observed series. However, such observed series must be stationary before tentative models are selected.

Model	Log likelihood	AIC	BIC
ARIMA(2,0,2) with non-zero mean	-292.29	600.5713	616.1813
ARIMA(0,0,0) with non-zero mean	-295.74	599.4701	607.2751
ARIMA(1,0,0) with non-zero mean	-295.03	600.0500	609.8063
ARIMA(0,0,1) with non-zero mean	-295.2	600.3942	610.1504
ARIMA(0,0,0) with zero mean	-362.63	731.2688	737.1225
ARIMA(1,0,1) with non-zero mean	**-292.91**	**597.8108**	**609.5182**
ARIMA(2,0,1) with non-zero mean	-292.52	599.0310	612.6897
ARIMA(1,0,2) with non-zero mean	-292.32	598.6377	612.2964
ARIMA(0,0,2) with non-zero mean	-294.82	601.6364	613.3438
ARIMA(2,0,0) with non-zero mean	-294.09	600.1852	611.8927
ARIMA(1,0,1) with zero mean	-299.48	608.9657	618.7219

Table 4.21:Estimated candidate ARIMA models for short rainfall time series (Original data)

The Table 4.21 gives the maximum likelihood estimates, their AIC and BIC for ARIMA (1,0,1) model based on automatic ARIMA forecasting. The estimated model is a "Regression with ARIMA (1,0,1) errors" which indicates no autoregressive or moving average pattern in the residuals. We can also see this by looking at an ACF plot of the residuals (Figure 4.18). Once the model order was determined then, next step is to go for parameter estimation of the model by maximum likelihood estimation method which is the second step in Box-Jenkins ARIMA model building procedure. The results of parameter estimation of ARIMA (1,0 ,1) are given in Table 4.22.

4.2.4.2.2 Estimation and significance check of Tentative model parameters

Coefficients	Estimate	Standard error	p-value
θ_1	-0.8856	0.1082	2.220446e-16
$\emptyset_1$	0.7204	0.1530	2.484536e-06
Intercept	251.5127	8.6454	0.000000e+00
SOI_SR3	-19.8939	9.6982	4.023665e-02
SST_PC13	-20.1352	9.5143	3.431848e-02

Table 4.22:Estimates of RARIMAE model (Original data)

From table 4.22, all coefficients of estimated parameters are significantly different from zero. There is a in θ_1 no stationarity as the absolute values of its estimate is not far from 1. After getting regression parameters the next step is residuals analysis for RARIMAE model analysis.

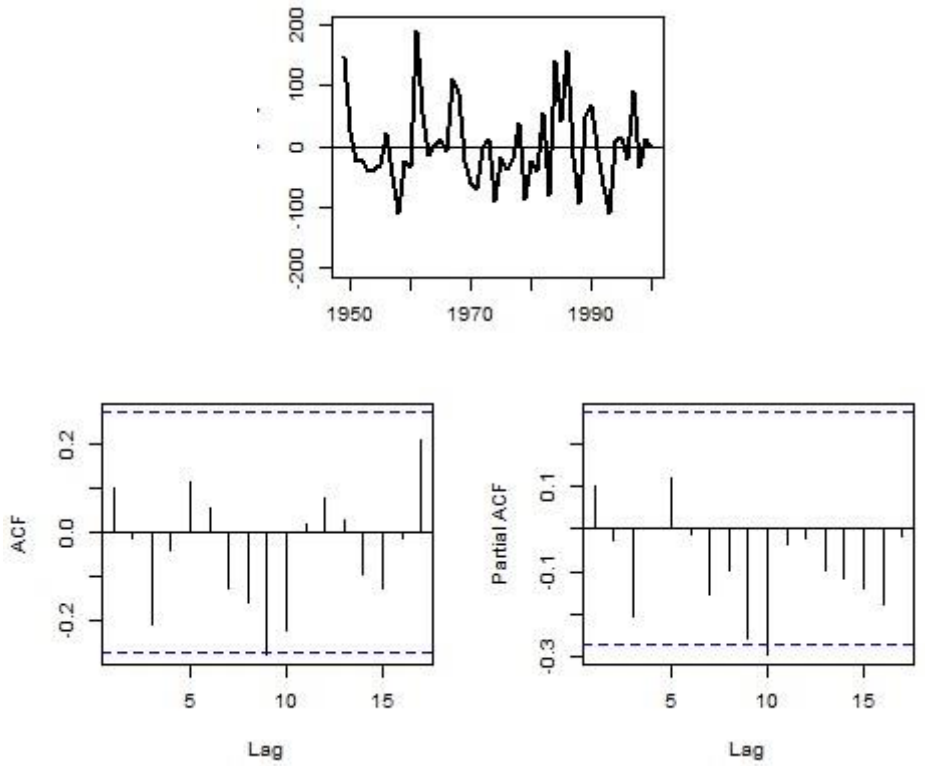


Figure 4.30:Diagnostic of residual plots for RARIMAE

Model	Log likelihood	AIC	BIC
ARIMA(0,0,0) with non-zero mean	-310.26	628.5204	636.2477
ARIMA(1,0,0) with non-zero mean	-296.85	603.698	613.3571
ARIMA(0,0,0) with zero mean	-310.29	626.5729	632.3684
ARIMA(2,0,0) with non-zero mean	-296.8	605.6019	617.1928
ARIMA(1,0,0) with zero mean	**-296.88**	**601.7615**	**609.4888**
ARIMA(2,0,0) with zero mean	-296.83	603.6661	613.3253

Table 4.23:Estimated candidate ARIMA models for short rainfall season (differenced data)

The Table 4.23 gives the maximum likelihood estimates, their AIC and BIC for ARIMA (1,0,0) model based on automatic ARIMA forecasting. The estimated model is a "Regression with ARIMA (1,0,0) errors" which indicates the presence of autoregressive pattern in the residuals. We can also see this by looking at an ACF plot of the residuals (Figure 4.23). ARIMA (1,0,0) means that the predicted value for the next year depending on the data 1 rainfall season before, 0 seasonal rainfall earlier error.

4.2.4.2.3 Estimation and significance check of model parameters

Coefficients	Estimate	Standard error	P-value
θ_1	-0.6790	0.1068	2.054799e-10
SOI_SR3	-15.9068	9.6743	1.001298e-01
SST_PC13	-24.3021	9.8737	1.384348e-02

Table 4. 24:Estimates of RARIMAE model (First differentiated data)

From table 4.24, all coefficients of estimated parameters are significantly different from zero. There is a in θ_1 stationarity as the absolute values of its estimate is far from 1. After getting regression parameters the next step is residuals analysis for RARIMAE model analysis.

4.2.4.2.4 Diagnostic checking process for the estimated statistical and predictive model

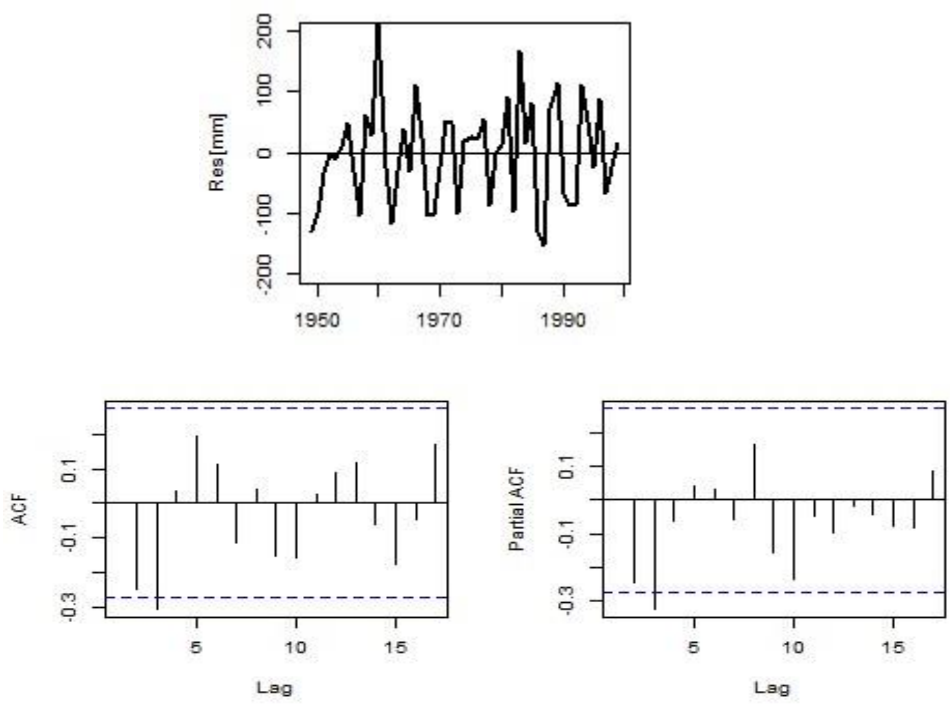


Figure 4.31:Diagnostics of ARIMA (1,0,0) fit on the first differenced data

		Original data			First differenced data		
		Summary statistics			Summary statistics		
	Test Type	X-squared	Df	p-value	X-squared	Df	p-value
Statistical model	Box-Pierce and Ljung-Box	26.29	16	0.0501	17.077	16	0.3806
Predictive model	Box-Pierce and Ljung-Box	22.993	16	0.1139	18.196	16	0.3126

Table 4.25:Box-Pierce and Ljung-Box Test for short rainfall season model

Diagnosing an ARIMA model is a crucial part of the model-building process and involves analysing the model residuals. A residual is the difference, or error, between the observed value and the model-predicted value. In this step, the residual is tested for evaluation purposes and goodness of the fit statistics is provided. Figure 4.29 and Figure 4.31 shows the ACF and PACF for the residuals of ARIMA (1,0, 0) models for both statistical and predictive model for short rainfall season. For figure 30 insignificant correlations appears at all lags, while for Figure 4.30 only one significant correlation appears at lag 2.

However, From Figure 4.28 and Figure 4.30, almost all the spikes of the ACF and PACF plots all lie within the confidence bounds suggesting that the residuals are white noise. For differentiated data, the normal Q-Q plot seems ok because most of the dataset lie on the straight line (Appendix I and J). The bell-shape feature is clearly noticed in Appendix I.1 and Appendix J.1; indicating that the residuals are normally distributed. This is approved by Jarque Bera test results in Appendix I.2 and Appendix J.2. A further analysis was conducted to ascertain the certainty of the residuals being white noise. Table 25 shows that the Box-Ljung testn reported a $\chi^2 = 17.077(\mathrm{df} = 16)$ with a large $\mathrm{p} - \mathrm{value} = 0.3806$ for statistical model, suggesting that the residuals from a model computed based on original data are white noise. For predictive model, a Box-Ljung test was reported a $\chi^2 = 18.196(\mathrm{df} = 16)$ with a large $\mathrm{p} - \mathrm{value} = 0.3126$,suggesting that the residuals from the model also are white noise.

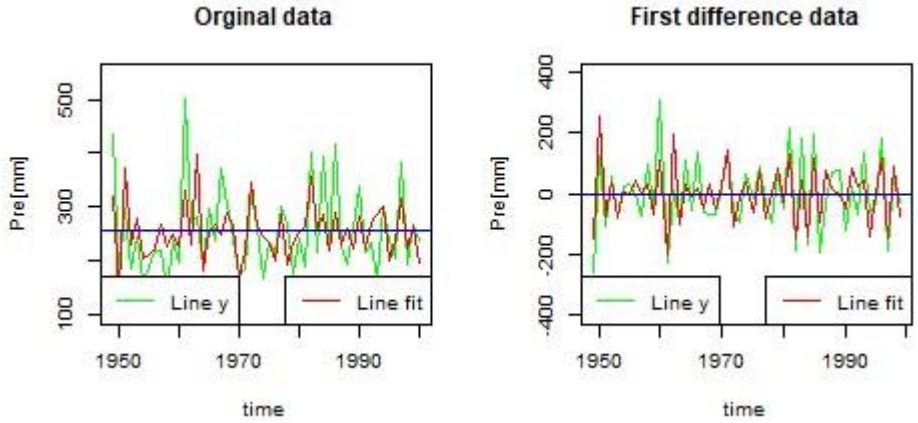


Figure 4.32:Observed vs Predicted short rainfall season (statistical model)

To evaluate the Regression with ARIMA errors model for short rainfall season (Statistical model) performance in East Africa to capture prediction performance was performed. Figure 4.32 shows a comparison between the real values and the ones resulted from the developed ARIMA model for the period between 1949 and 2000. ARIMA (1,0,0) model on differentiated data, we obtain the graphical plot for the actual pitch series versus the predicted pitch series and from the visual inspection of the plot it is quite evident that the chosen model is good enough as the predicted. The correlation coefficient (r) for models developed based on time series original data and first differentiated data are 0.64 and 0.78 respectively. Same case for both models the RMSE were found to be 60.56 and 75.11 respectively.

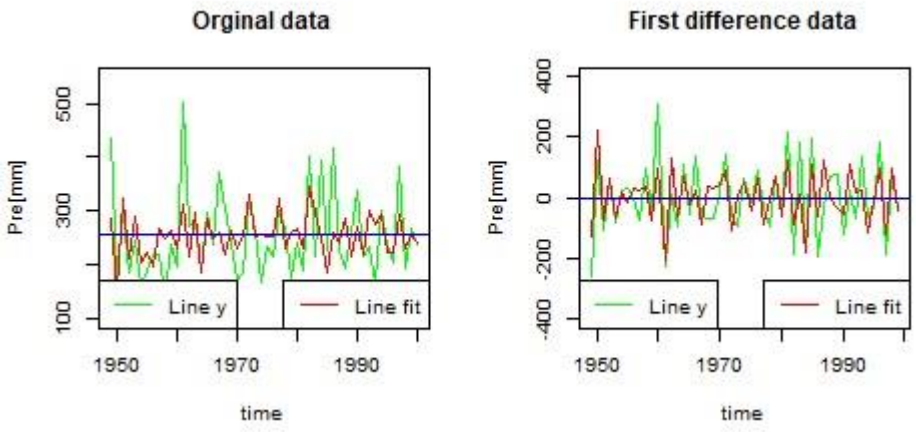


Figure 4.33:Observed vs Predicted short rainfall season (Predictive model)

To evaluate the Regression with ARIMA errors model for short rainfall season (Predictive model) performance in East Africa to capture prediction performance was performed. Figure 4.33 shows a comparison between the real values and the ones resulted from the developed

ARIMA model for the period between 1949 and 2000. ARIMA (1,0,0) model on differentiated data, we obtain the graphical plot for the actual pitch series versus the predicted pitch series and from the visual inspection of the plot it is quite evident that the chosen model is good enough as the predicted. The correlation coefficient (r) for models developed based on time series original data and first differentiated data are 0.52 and 0.74 respectively. Same case for both models the RMSE were found to be 67.47 and 81.14 respectively.

4.2.5 Comparison of overall prediction accuracy of the models under study

This study presents the inclusion of exogenous variables in the ARIMA model (termed as Regression with ARIMA errors) and showed a good performance for RARIMAE errors for seasonal rainfall prediction in East Africa. The inclusion of an exogenous variable is only possible if these predictors show a significant correlation with the dependent variable. For the first part of this study, a study on spatial and temporal analysis on seasonal rainfall situation was done with the help of maps in which 3569 grid box were analysed. The correlation analyses were conducted between 72 teleconnections (for statistical model) and 58 teleconnections (for predictive model) and seasonal rainfall time series in each grid points.

To understand the effectiveness of the Regression with ARIMA errors models, its statistical parameters were compared with developed MLR model for the same regions. Temporal and spatial presentation with the help of maps showed low value for RMSE and improved values of R-squared all computed using regression RARIMAE errors (Figure 4.7,4.8). From such comparison, a significant rise in R-squared, a decrease of RMSE and a decrease in MAE values were observed in RARIMAE models for both short rainfall and long rainfall season averaged time series. In terms of reliability, RARIMAE outperformed its MLR counterparts with better efficiency and accuracy (Figure 4.15,4.22 ,4.24,4.27,4.32 and 4.33). For this section the overall prediction accuracy of all the models under study has been discussed. The prediction accuracy is measured in terms of R-squared, RMSE and MAE as discussed in chapter three.

			Original data		First differentiated data	
Season	Model	Criteria	MLR	RARIMAE	MLR	RARIMAE
Long rainfall season	Statistical and predictive model	R-squared	0.208	0.208	0.274	0.536
		RMSE	56.659	56.659	74.234	59.466
		MAE	45.569	45.569	59.895	47.524
Short rainfall season	Statistical model	R-squared	0.318	0.413	0.301	0.614
		RMSE	65.285	60.566	101.04	75.118
		MAE	46.988	45.807	79.137	60.718
	Predictive model	R-squared	0.184	0.272	0.228	0.551
		RMSE	71.398	67.465	106.17	81.141
		MAE	51.128	51.187	81.470	66.169

Table 4.26: Comparison of forecasting performance of all models for long and short rainfall season time series

The developed model can predict long rainfall season for 9 months in advance for the region with SLP and DMI as predictors. For the averaged long rainfall time series, the above-mentioned model has been fitted and modelling and forecasting performance has been assessed in terms of their prediction ability measured by model errors under both original and differentiated data set. For the original data sets, RARIMAE model performed as the MLR model (Table 26) as R-squared, RMSE and MAE of RARIMAE model for both models are equal, and this is due to the residuals from multiple linear regression model which are white noise. For the differentiated data, the RARIMAE model performed better as compared to regression analysis in both data set (Table 4.26) as R-squared is higher while RMSE and MAE of RARIMAE model is lower as compared to regression model.

The developed predictive model can also predict short rainfall season for 3 months in advance for the region with SOI and the first principal component of SST(SST_PC1) as predictors, while the statistical model adds on these two variables SAOD (with current time series) as the third predictor. For the averaged short rainfall time series, the above-mentioned model has been fitted, modelling and forecasting performance has been assessed in terms of their prediction ability measured by model errors under both original and differentiated data set. For both the original data and differentiated data, the RARIMAE model performed better as compared to regression analysis in statistical and predictive model (Table 4.26) as R-squared is higher while RMSE and MAE of RARIMAE model is lower as compared to regression model. This is due to the existence of autocorrelation in residuals from MLR M performed based on original data as well as on differentiated data.

Such capability of the model to predict long rainfall season up to 9 months in advance and short rainfall season up to 3 months in advance has also justified by several studies. All these studies considered lagged climate indices to forecast East African seasonal rainfall as the current study did. Chena and Georgakakos (2015) conducted a study on Seasonal prediction of East African rainfall. This study compares several forecasting methods using SST anomalies to predict East African rains with various lead time. It has shown positive evidence to use climate indices to predict long rainfall season in several months in advance (up to 11 months) and short rainfall season in few months in advance (up to 3 months). They conclude that unlike the results of the short rains, the optimal lead times for the long rains are consistently longer than for the short rains.

However, the developed MLR models have some limitation as they were not able to predict all extreme cases. Investigating the existing nonlinear relationship between rainfall and climate drivers can improve the better understanding of the trend, and associated variabilities that the developed models failed to address. Possible analysis approaches can be considered for developing a model which captures nonlinear and linear component. Since rainfall is a complex mechanism, any linear or non-linear model by itself, might not be able to predict or capture all the extreme cases. The regression with ARIMA errors model residuals can be used to explain the non-linear relationship, where the combined output of both MLR and non-linear models can be used for improving forecasting.

By application of OLS (Ordinary least squares) method on long rainfall season and short rainfall season for both statistical model and predictive model, it is shown that there exists autocorrelation of residual at the first difference. This support the statement saying that one of the major problems encountered while using time series data is Autocorrelation. Consequences of autocorrelation are: (a) the estimates of the parameters do not have the statistical bias. In other words, even when the residuals are serially correlated the parameter estimates of OLS are statistically unbiased, in the sense that their expected value is equal to the true parameter. (b) With autocorrelation values of disturbance term, the OLS variances of the parameter estimates are likely to be larger than those of other econometric methods. (c) The variance of the random term may be seriously underestimated if the u are uncorrelated. And (d) if the values of u's are uncorrelated, the predictions based on the OLS estimate will be inefficient. Therefore, whenever the data suffer from autocorrelation, we can go for MLR with ARIMA error, the

ARIMA error part is more to correct the autocorrelation thereby improving the variance and productiveness of the model.

CONLUSION

This study investigated the influence of climate drivers on East African rainfall variability and developed the forecast models to predict seasonal long and short rainfall using the RARIMAE and MLR models. In this attempt, climate drivers were used as transfer functions, while all other previous attempts considered conventional time series models only. From a statistical perspective, it was evident that the predictability performance of the RARIMAE model is much higher than the MLR models. For short rainfall season the models are capable of predicting rainfall 3 months in advance while for the long rainfall season they can predict the rainfall only 9 months in advance.

The developed RARIMAE models have shown a strong correlation (r) as well as minimum errors. It was also observed that all these RARIMAE models were successful in predicting seasonal rainfall in East Africa and their ability to predict long rainfall season in advance of 9 months and short rainfall in 3 months has strengthened their acceptability. From the stakeholder's perspective, such flexibility offered in the developed model has greater importance, as a timely prediction can help in strategic decision making and reducing associated risks and damage potentials. Overall, the SLP_PC19 – DMI_LR9 model for the long rainfall season and SST_PC13–SOI_SR3 model for the short rainfall season predictive model showed exceptional performance with good prediction accuracy and can be recommended for future rainfall prediction in East Africa.

However, some aspects should be given attention. Regression with ARIMA errors could be quite complex. It is important to look for the most parsimonious model. As regarded the data, it is clear that the number of potential variables tested in our models is large but the number of selected variables in our models is limited. Nevertheless, the combination of regression model and an ARIMA error structure gives an acceptable fit, even without the non-selected elements. The effect of these omitted factors, but explicitly tested, are reflected in the error terms. Also, adding more explanatory variables brings more multicollinearity into the model.

BIBLIOGRAPHY

Ali, A., Amani, A., Diedhiou, A., and Lebel T. (2005). Rainfall in the Sahel. Part II: Evaluation of Rain Gauge Networks in the CILSS Countries and Objective Intercomparison of Rainfall Products. Journal of Applied meteorology, 44:1707-1722.doi:1175/JAM2305.1.

Anyah, R.O. and Semazzi, F.H.M. (2007). Variability of East Africa rainfall based on multiyear RegCM3 simulations. *Int.J.Climat*.,Behera,357-371.

Anyamba, E.K. (1984). Some aspects of the origin of rainfall deficiency in East Africa. *Proc. Of the WHO regional scientific conf* On GATE, WAMEX and tropical Meteorology, Dakar, Senegal,110-112.

Anyamba, E.K. (1993). Short term climate variability in East Africa. *Proc. First International Conference of Africa Meteor. Soc*., Nairobi Kenya,224-235.

Bergmeir, C., Hyndman,R.J., and Koo,B. (2018). A Note on the validity of cross-validation for evaluating Autoregressive Time series Prediction. *Computational Statistics and Data Analysis*, 120:70-83.

Birkett, C., Murtugudde, R., and Allan, T. (1999). Indian Ocean climate event brings floods to East Africa's lakes and the Sudd Marsh. *Geophys. Res.Lett*.26:1031-1034.

Black, E., Slingo, J., and Sperber, K.R. (2003). An observational study of the relationship between excessively strong short rains in coastal east Africa and Indian Ocean SST. Mon. Weather. Rev. 131:74-94.

Bowden, J. and Semazzi, F.H.M. (2007). Empirical analysis of intraseasonal climate variability over the Greater Horn of Africa. J. Climate 20(23):5715-5731.

Box, G.E.P. and Jenkins, G.M. (1970). Time series analysis, Forecasting and control, Holden-Day, San Francisco, CA.

Box, G.E.P. and Jenkins, G.M. (1979). Time Series Analysis: Forecasting and Control; Revised Ed.; Holden-Day: San Francisco, CA, USA.

Broad, K. and Agrawala, S. (2000). The Ethiopia food crisis-Uses and limits of climate forecasts. *Science* 289:1693-1694.

Brohan, P., Kennedy, J.J., Harris I., et al. (2006). Uncertainty estimates in regional and global observed temperature changes: A new data set from 1850. Journal of Geophysical Research 111:D12106.doi:10.1029/2005JD006548.

Camberlin, P. and Okoola R.E. (2003). The onset and cessation of the "long rains" in Eastern Africa and their interannual variability. *Theory. App. Climatol.*75:43-54.

Camberlin, P. and Philippon, N. (2002). The East African March–May rainy season: Associated atmospheric dynamics and predictability over the 1968–97 period. J. Climate, 15:1002–1019.

Chadsuthi, S., Modchang, C., Lenbury, Y., Iamsirithaworn, S., and Triampo, W. (2012). Modeling seasonal leptospirosis transmission and its association with rainfall and temperature in Thailand using time–series and ARIMAX analyses. *Asian Pac. J. Trop. Med*., 5:539–546.

Chen, C.J. and Georgakakos, A.P. (2014). Hydroclimatic forecasting using sea surface temperatures – Methodology and application for the southeast U.S. *Clim.Dyn*.42:2955-29822, doi:10:1007/s00382-013-1908-4.

Chen, C.J. and Georgakakos, A.P. (2015). Seasonal prediction of East Africa rainfall. *International Journal of Climatology,*35:2698-2723.

Cryer, J.D. and Chan, K.S. (2008). Time Series Analysis: With Applications in R; Springer Science & Business Media: Berlin, Germany.

Davis, F.E. and Harrell, G.D. (1942). Relation of weather and its distribution to corn yield. *U.S. Department of Agriculture, Technical Bulletin,* 806.

Deryng, D., Conway, D., Ramankutty, N., et al. (2014). Global crop yield response to extreme heat stress under multiple climate change futures. Environmental Research Letters 9:034011.doi:10.1088/1748-9326/9/3/034011.

Di Falco S., Yesuf, M., Kohlin, G., and Ringler, C. (2012). Estimating the impact of climate change on agriculture in low-income countries: Household level evidence from the Nile basin, Ethiopia, Environmental and Resource Economics, vol.52 no.4, pp.457 – 478.

Dinku, T., Ceccato, P., Grover-Kopec, E., et al. (2007). Validation of satellite rainfall products over East Africa' s complex topography. *International Journal of Remote Sensing* ,28:1503-1526.doi:10.1080/01431160600954688.

Diro, G. T., Grimes, D. I. F., and Black, E. (2008): Seasonal forecasting of Ethiopian spring rains. Meteor. Appl., 15: 73–83, doi:10.1002/met.63.

Doocy, S., Daniels, A., Murray, S., and Kirsch T.D. (2013). The human impact of floods: a historical review of events 1980-2009 and systematic literature review. *PLOS Curr.Disas*.,doi:10.1371/currents.dis.f4deb457904936b07c09daa98ee8171a.

Draper, N.R. and Smith, H. (1998). Applied Regression analysis, Wiley Series in Probability and Mathematical Statistics, 171-172.

Ermert, V., Fink, A.H., Morse, A.P., and Paeth, H. (2012). The impact of regional climate change on malaria risk due to greenhouse forcing and land-use changes in tropical Africa. Environmental health perspectives 120:77-84.doi:10.1289/ehp.1103681.

Fan, J., Shan, R., Cao, X., and Li, P. (2009). The analysis to tertiary-industry with ARIMAX model. *J. Math. Res*., 1, 156.

Fisher, R.A. (1925). The influences of rainfall on the yield of wheat at Rothamsted. *Philosophical Transaction of Royal Society of London*, Series B, 213: 89-142.

Folland, C., Owen, J., Ward, M.N., and Colman A.W. (1991). Prediction of seasonal rainfall in the Sahel region of Africa using empirical and dynamical methods. *J.Forecasting*, 10:2-56.

Funk, C., Senay, G., Asfaw, A., Virdin J., Rowland, J., Michaelsen, J., Eilerts, G., Korecha, D. and Choularton, R. (2005). Recent drought tendencies in Ethiopia and equatorial-subtropical eastern Africa. Famine Early Warning System Network (FEWS NET) Special Report, USAID, Washington,DC.

Gaardbo Kuhn, K., Campbell-Lendrum, D.H., and Davies, C.R. (2002). A Continental Risk Map for Malaria Mosquito (Diptera: Culicidae) Vectors in Europe. Journal of Medical Entomology 39:621-630.doi:10.1603/0022-2585-39.4.621.

Gbobaniyi, E., Sarr, A., and Sylla, M.B. (2014). Climatology, annual cycle and interannual variability of precipitation and temperature in CORDEX simulations over West Africa. *International Journal of Climatology*, 34:2241-2257.doi:10.1002/joc.3834.

Gilbert, W. (1931). On Periodicity in Series of Related Terms, Proceedings of the *Royal Society of London, Series A*, 131: 518-532.

Haile, G.G., Tang, Q., Sun, S., Huang, Z., Xhang, X., and Liu, X. (2019). Drought in East Africa: Causes, impacts and resilience. *Earth - Science Reviews*, doi.org/10.1016/j.earscirev.2019.04.015, 193,146-161.

Han, P., Wang, P.X., and Zhang, S.Y.(2010). Drought forecasting based on the remote sensing data using ARIMA models. *Math. Comput. Model*. 51:1398–1403.

Harris, I., Jones, P.D., Osborn, T.J., and Lister, D.H. (2014a). Updated high-resolution grids of monthly climatic Observations - the CRU TS3.10 Dataset. International Journal of Climatology, 34:623-642.doi:10.1002/joc.3711.

Harris, R. and Sollis, R. (2003). Applied time series modelling and forecasting. Chichester, UK: John Wiley & Sons.

Harris, R.M.B., Grose, M.R., Lee G., et al (2014b). Climate projections for ecologists. Wiley Interdisciplinary Reviews: Climate Change 5:621-637.doi:10.1002/wcc.291.

Hastenrath, S. (2007). Circulation mechanisms of climate anomalies in East Africa and the equatorial Indian Ocean. *Dyn.Atmos.Oceans*, 43:25-35.

Hastenrath, S., Polzin, D., and Mutai C. (2007). Diagnostic the 2005 drought in equatorial East Africa. *J.Clim*.20:4628-4637.

Hastenrath, S., Polzin, D., and Mutai C. (2010). Notes and correspondence: Diagnostic the 2005 drought in equatorial East Africa during boreal autumn 2005-08. *J.Clim*.23:813-817.

Hastenrath, S., Polzin, D., and Camberlin, P. (2004). Exploring the predictability of the 'short rains' at the coast of East Africa. Int. J. Climatol., 24:1333–1343, doi:10.1002/joc.1070.

Huang, B., Banzon, V.F., Freeman, E., Lawrimore, J., Liu,W., Peterson,T,C.,Smith,T.M., Thorne,P.W., Woodruff,S.D., and Zhang,H.M. (2014). Extended Reconstructed Sea Surface Temperature version4 (ERSST.v4):Part I. Upgrades and intercomparisons, J.Climate, 28: 911-930.

Huffman, G.J., Adler, R.F., Bolvin, D.T., and Gu G. (2009). Improving the global precipitation record: GPCP Version 2.1. Geophysical Research Letters, 36:L17808. doi:10.1029/2009GL040000.

Hurrell, J.W. (1995). Decadal Trends in the North Atlantic Oscillation: Regional Temperatures and Precipitation. Science.269(5224):676 – 679.doi:10.1126/science.269.5224.676.

Hurrell, J.W., Kushnir,Y., Ottersen, G., and Visbeck, M.(2003). The North Atlantic Oscillation: Climatic Significance and Environmental Impact. American Geophysical Union. ISBN 9780875909943.

Hyndman, R. J. and Khandakar, Y. (2008). Automatic time series forecasting: The forecast package for R. *Journal of Statistical Software*, *27*(1),122.

Ilunga, L., Muhire, I., Mbaragijimana, C. (2004). Pluviometric seasons and rainfall in Rwanda. *Geo-Ec-Trop* 28:61-68.

Indeje, M., Semazzi, F.H.M., and Ogallo,L.J. (2000). ENSO signals in East African rainfall and their prediction potentials. *Int. J. Climatol*. 20:19- 46.

Ismail, Z., Yahya, A., and Shabri, A. (2009). "Forecasting Gold Prices Using Multiple Linear Regression Method" in American Journal of Applied Sciences. 6(8): 1509-1514.

Jacob, D., Elizalde, A., Haensler, A., et al (2012). Assessing the Transferability of the Regional Climate Model REMO to Different COordinated Regional Climate Downscaling EXperiment (CORDEX) Regions. *Atmosphere*, 3:181-199.doi:10.3390/atmos3010181.

Jalalkamali, A., Moradi, M., and Moradi, N. (2015). Application of several artificial intelligence models and ARIMAX model for forecasting drought using the Standardized Precipitation Index. *Int. J. Environ. Sci. Technol*., 12: 1201-1210.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2014). *An Introduction to Statistical Learning: With Application in R* (Springer Publishing Company, Incorporated).

Jury, M.R., Pathack, B., and Parker, B. (1999). Climatic determinants and statistical prediction of tropical cyclone days in the Southwest Indian Ocean. *J.Climate*, 12:1738-1746.

Kim, J., Waliser, D.E., Mattmann, C.A., et al (2013). Evaluation of the CORDEX-Africa multi-RCM hindcast: systematic model errors. *Climate Dynamics*, 42:1189-1202.doi:10.1007/s00382-013-1751-7.

Kizza, M., et al. (2009). Temporal rainfall variability in the Lake Victoria Basin in East Africa.

Kumar, K.K., Soman, M.K., and Kumar, K.R. (1995). Seasonal forecasting of Indian summer monsoon rainfall: A review. *Weather*, 50:449-467.

Ling, A., Darmesah, G., Chong, K., and Ho, C. (2019). Application of ARIMAX Model to Forecast Weekly Cocoa Black Pod Disease Incidence. *Math. Stat*., 7, 29–40.

Liu, W., Huang, B., Thorne, P.W., Banzon, V.F., Zhang, H.M., Freeman, E., Lawrimore,J., Peterson, T.C., Smith, T.M., and Woodruff, S.D. (2014). Extended Reconstructed Sea Surface Temperature version4 (ERSST.v4): Part II. Parametric and structural uncertainty estimations, *J.Climate*, 28:931-951.

Ljung, G.M. and Box, G.E. (1978). On a measure of lack of fit in time series models. *Biometrika*, 65:297-303.

Lobell, D.B. and Field, C.B. (2007). Global scale climate – crop yield relationships and the impacts of recent warming. *Environmental Research Letters,* 2:014002.doi:10.1088/1748-9326/2/1/014002.

Lyon, B. and DeWitt, D.G. (2012). A recent and abrupt decline in the East African long rains.*Geophys.Res.Lett*. 39: L02702, doi:10.1029/2011GL050337.

Mitchell, T.D. And Jones, P.D. (2005). An improved method of constructing a database of monthly climate observations and associated high-resolution grids. International Journal of Climatology, *25*: 693-712.doi:10.1002/joc.1181.

Momani, P. and Naill, P.E. (2009). Time series analysis model for rainfall data in Jordan: Case study for using time series analysis. *Am J.Environ. Sci.*,5,599.

Montgomery, D.C., Peck, E.A., and Vining, G.G. (2003). Introduction to Linear Regression Analysis, 3rd Ed.John Wiley and Sons (Asia) Pvt.Ltd.

Mutai, C.C., Ward, M.N., and Colman, A. W. (1998). Towards the prediction of the East Africa short rains based on sea-surface temperature–atmosphere coupling. Int. J. Climatol., 18:975–997.

Mutai,C.C. and Ward,M.N.(2000). East African rainfall and the tropical circulation/conviction on intra-seasonal to interannual timescales. *J.Climate* ,13:3915-3939.

Mwale, D. and Gan, T.Y.(2005). Wavelet analysis of variability, teleconnectivity, and predictability of the September–November East African rainfall. J. Appl. Meteor., 44:256–269, doi:10.1175/JAM2195.1.

Neter, J., Kutner, M., Nachtsheim, C., and Wasserman,V.(1996). *Applied Linear Stochastic Models*. McGraw-Hill/Irwin: New York, NY.

New, M., Hulme, M., and Jones, P. (1999). Representing Twentieth-Century Space – Time Climate Variability. Part I: Development of a 1961 – 90 Mean Monthly Terrestrial Climatology. Journal of Climate, 12:829 – 856.

Newbold, P. and Granger, C.W.J. (1974). Experience with forecasting Univariate time series and the combination of forecasts, *Journal of the Royal Statistical Society, Australia*, 137: 131-165.

Nicholson, S.E. (2019). A review of climate dynamics and climate variability in Eastern Africa. In *the Limnology, Climatology and Paleoclimatology of the East African Lakes*, Johnson TC, Odada E (eds). Gordon & Breach: Toronto,25-56.

Nicholson, S.E., Some, B., McCollum, J., et al (2003a). Validation of TRMM and other Rainfall estimates with high-Density Gauge Dataset for West Africa. Part I: Validation of GPCC rainfall Product and Pre-TRMM Satellite and Blended Products. *Journal of Applied Meteorology*, 42: 1337-1354.

Nicholson, S.E., Some, B., McCollum, J., et al (2003b). Validation of TRMM and other Rainfall estimates with high-Density Gauge Dataset for West Africa. Part II: Validation of TRMM rainfall Products. *Journal of Applied Meteorology*, 42: 1355-1368.

Nikulin, G., Jones, C., and Giorgi, F. (2012). Precipitation Climatology in an Ensemble of CORDEX-Africa Regional Climate Simulations. *Journal of Climate*, 25:6057-6078.doi:10.1175/JCLI-D-11-00375.1.

Nkrintra, S., Balaji, R., Martyn, C., et al. (2005). "Seasonal Forecasting of Thailand Summer Monsoon Rainfall", in International Journal of Climatology, Vol. 25, Issue 5, American Meteorological Society, 2005, pp. 649-664.

Ntale, H. K., Gan T. Y., and Mwale D. (2003). Prediction of East African seasonal rainfall using simplex canonical correlation analysis. J. Climate, 16, 2105–2112.

Nyakwanda, W. (2003). Climate risk and vulnerability in Kenya. Proc. *The sixth Kenya Meteor. Soc. Workshop on Meteor*. Research, Applications and Services, Mombasa, Kenya, 29 September to 3 October 2003.pp35-45.

Omondi, P., Ogallo., Anyah, R., Muthama, J.M., and Ininda J. (2013). Linkages between global sea surface temperatures and decadal rainfall variability over the Eastern Africa region. *Int.J.Climatol*.33:2082-2104.

Otok, B.W. and Suhartono, F. (2009). Development of rainfall forecasting model in Indonesia by using ASTAR, transfer function, and ARIMA methods. *Eur. J. Sci. Res*.,38, 386–395.

Owiti, Z., Ogallo, L.A., and Mutemi, J. (2008). Linkages between the Indian Ocean Dipole and east African seasonal rainfall anomalies. J. Kenya Meteor. Soc. 2:3-17.

Owiti, Z.O. (2005). Use of the Indian Ocean Dipole indices as predictor east African rainfall anomalies. MSc Thesis, Department of Meteorology, University of Nairobi, 130.

Paeth, H. and Diederich, M. (2011). Postprocessing of simulated precipitation for impact research in West Africa. Part II: A weather generator for daily data. Climate Dynamics *36*:1337-1348.doi:10.1007/s00382-010-0840-0.

Paeth, H., Born, K., Podzun, R., and Jacob, D. (2005). Regional dynamical downscaling over West Africa: model evaluation and comparison of wet and dry years. Meteorologische Zeitschrift 14:349-367.doi:10.1127/0941-2948/2005/0038.

Paeth, H., Hall, N.M.J., Gaertner, M.A., at al (2011b). Progress in regional downscaling of west African precipitation. Atmospheric Science Letters, 12:75-82. doi:10.1002/asl.306.

Panga, P.A. (2021). Forecasting Rainfall in Tanzania Using Time Series Approach Case Study: Dar es Salaam. J Climatol Weather Forecast ,9: 272.

Panitz, H.J., Dosio, A., Büchner, M., et al. (2013). COSMO-CLM (CCLM) climate simulations over CORDEX-Africa domain: analysis of the ERA-Interim driven simulations at 0.44^0 and 0.22^0 resolution. Climate Dynamics,42:3015-3030.doi:10.1007/s00382-013-1834-5.

Pankratz, A. (2012). Forecasting with dynamic regression models.

Perker, D.J., Good E., and Chadwick, R. (2011). Reviews of observational data available over Africa for monitoring, attribution and forecast evaluation. Hadley Centre Technical Note HCTN86.

Peter, D. and Silvia, P. (2012). ARIMA vs. ARIMAX–which approach is better to analyze and forecast macroeconomic time series. In Proceedings of the 30th International Conference Mathematical Methods in Economics, Karviná, Czech Republic, 11–13.

Peterson, T.C., Karl, T.R., Jamason, P.F., et al. (1998). First difference method : Maximizing station density for the calculation of long-term global temperature change. Journal of Geophysical Research 103:25967.doi:10.1029/98JD01168.

Philippon, N., Camberlin, P. and Fauchereau, N. (2002). Empirical predictability study of October–December East African rainfall. Quart. J. Roy. Meteor. Soc., 128, 2239–2256.

Polhl, B. and Camberlin, P. (2006). Influence of the Madden-Julian Oscillation of the East African rainfall: II. March-May season extremes and interannual variability. *Q.J.R.Meteorol.Soc*.132:2541-2558.

Sachs, L. and Hedderich, J.(2006). *Angwandte Statistik. Methodensammlung mit R*. Springer, Berlin, Heidelberg, New York , 12 edition.

Saigal, S. and Mehrotra, D. (2012). Performance comparison of time series data using predictive data mining techniques. *Adv. Inf. Min*. 4:57–66.

Saji, N.H., Goswani, B.N., Vinayachandran, P.N., and Yamagata, T. (1999). A dipole mode in the tropical Indian Ocean. Nature 401:360-363.

Singh, J., Knapp, H.V., Arnold, J., and Demissie, M.(2005). Hydrological modeling of the Iroquois river watershed using HSPF and SWAT 1. *J. Am. Water Resour. Assoc*. 41:343–360.

Slutzky, E. (1937). The Sommation of Random Causes as the Source of Cyclic Processes, *Econometrica, 5*: 105-146.

Stepherson,D.B., Wanner, H., Brönnimann,S., and Luterbacher,J. (2003). The History of Scientific Research on the North Atlantic Oscillation, in The North Atlantic Oscillation: Climatic Significance and Environmental Impact, edited by Hurrel, J.W.,Kushnir, Y., Ottersen, G., and Visbeck, M.,pp37 – 50, American Geophysical Union, Washington, DC,doi:10:1029/134GM02.

Storch, H.V. and Zwiers, F.W.(2004). Statistical analysis in climate research. Cambridge University Press, Cambridge.

Sylla, M.B., Giorgi, F., Coppola, E. and Mariotti, L. (2012). Uncertainties in daily rainfall over Africa: assessment of gridded observation products and evaluation of a regional climate model simulation. *International Journal of Climatology*, 33:1805-1817. doi:10.1002/joc.3551.

Tao, F., Yokozawa, M. and Zhang, Z. (2008). Climate – crop yield relationships at provincial scales in China and the impacts of recent climate trends. Climate Research 38:83-94.doi:10.3354/cr00771.

Terzi, Ö. (2012). "Monthly Rainfall Estimation Using Data-Mining Process" in Hindawi Publishing Corporation Applied Computational Intelligence and Soft Computing. Volume 2012, 6.

Trenberth, K.E., Jones, P.D., Ambenje, P., et al (2007). Observations: Surface and Atmospheric Climate Change. In:Solomon S.,Qin D., Manning M., at al (eds) Climate Change 2007: The Physical Science Basis. Contribution of the Working Group I to the Fourth Assessment Report of the IPCC. Cambridge University Press, Cambridge, United Kingdom and New York, USA, pp 235-336.

Tularam, G. (2010). Relationship between El Niño southern oscillation index and rainfall (Queensland, Australia). *Int. J. Sustain. Dev. Plan*.,5:378–391.

Ummenhofer, C.C., Gupta, A.S., Enlgand, M.H., and Reason, C.J.C. (2009). Contributions of Indian Ocean seas surface temperatures to enhanced East African rainfall.*J.Clim*.22:993-1013.

Van den Bossche, F., Wets, G., and Brijs, T. (2004). A regression model with ARMA errors to investigate the frequency and severity of road traffic accidents. Proc. 83rd Annual Meeting of the Transportation Research Board, Washington, DC, Transportation Research Board of the National Academies, TRB2004-001658.

Verdin, J., Funk, C., Senay, G., Choularton, R. (2005). Climate science and famine early warning.*Phil.Trans.R. Soc.B* 360B:2155-2168.

Weeks, W. and Boughton,W. (1987). Tests of ARMA model forms for rainfall-runoff modelling. *J. Hydrol.*,91:29–47.

Wilks, D.S. (2011). Statistical Methods in the Atmospheric Sciences. Academic press.

Williams, A.P. And Funk, C. (2011). Westward extension of the warm pool leads to a westward extension of the Walker circulation, drying eastern Africa. *Clim.Dyn*,37:2417-24:35, doi:10.1007/s00382-010-0984-y.

Wolter, K. (1987). The Southern Oscillation in surface circulation and climate over the tropical Atlantic, Eastern Pacific, and Indian Oceans as captured by cluster analysis. J. Climate Appl. Meteor., 26, 540-558.

Wolter, K. and Timlin, M.S. (1993). Monitoring ENSO in COADS with a seasonally adjusted principal component index. Proc. of the 17th Climate Diagnostics Workshop, Norman, OK, NOAA/NMC/CAC, NSSL, Oklahoma Clim. Survey, CIMMS and the School of Meteor., Univ. of Oklahoma, 52-57.

Yaffee, R.A. and McGee, M. (2000). An introduction to tome series analysis and forecasting: With application of SAS and SPSS.

Yaglom, A.M. (1955). Correlation theory of processes with stationary increments of order *n*. *Matematicheskii Sbornik*. 37, 141. (English translation in American Mathematical Society. Translation Series. 2: 8–87, 1958).

Yule, G.U. (1927). On a Method of Investigating Periodicities in Disturbed Series, with Special Reference to Wolfer's Sunspot Numbers Philosophical Transactions of the Royal Society of London, Series A, 226: 267-298.

Zaw, W.T. and Naing, T.T. (2008). "Empirical Statistical Modeling of Rainfall Prediction over Myanmar" in World Academy of Science, Engineering and Technology.,10-270.

Zebiak, S.E. (2003). Research Potential for Improvement in Climate Prediction. *Bull. Amer Meteorol*. Soc., 84:1692:1696.

Zhang, G.P. (2003). Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing*,50, 159–175.

Zhang, Q., Körnich, H., and Holmgren, K. (2012). How well do reanalyses represent the southern Africa precipitations? Climate Dynamics, 40:951-962.doi:10.1007/s00382-012-1423-z.

APPENDICES

Appendix A

PC-№	Actual	3 months lead	6 months lead	9 months lead	12 months lead
1	22.49	29.12	31.19	19.44	20.00
2	8.41	7.44	6.31	8.14	7.68
3	6.98	6.80	5.81	7.10	7.03
4	5.56	5.01	4.35	5.61	5.63
5	4.77	4.28	4.21	4.59	4.71
Total	47.91	52.65	51.87	44.88	45.05

The first five principal component analysis modes of long rainfall season sea surface temperature

Appendix B

PC-№	Actual	3 months lead	6 months lead	9 months lead	12 months lead
1	31.72	20.54	22.49	29.12	31.19
2	6.71	8.34	8.41	7.44	6.31
3	5.82	7.09	6.98	6.80	5.81
4	4.38	5.67	5.56	5.01	4.35
5	3.75	4.52	4.77	4.28	4.21
Total	52.30	46.16	47.91	52.65	51.87

The first five principal component analysis modes of short rainfall season sea surface temperature

Appendix C

PC-Nº	**Actual**	**3 months lead**	**6 months lead**	**9 months lead**	**12 months lead**
1	16.28	21.73	19.20	21.85	16.13
2	12.38	15.18	12.22	12.27	12.26
3	9.22	10.58	7.75	9.82	9.99
4	6.74	6.51	7.04	6.67	6.63
5	5.81	5.82	5.88	5.21	5.77
Total	50.43	59.82	52.09	55.82	50.78

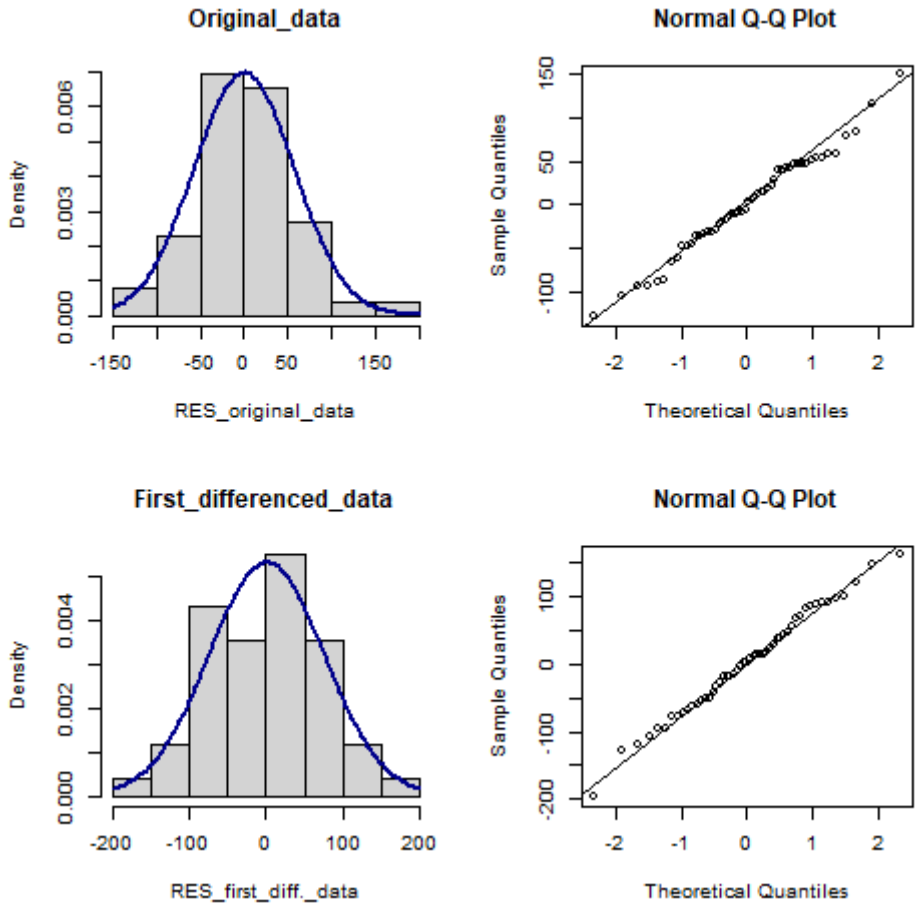
The first five principal component analysis modes of long rainfall season sea level pressure

Appendix D

PC-Nº	**Actual**	**3 months lead**	**6 months lead**	**9 months lead**	**12 months lead**
1	19.81	21.91	16.28	21.73	19.20
2	12.14	12.37	12.38	15.18	12.22
3	7.51	10.03	9.22	10.58	7.75
4	6.87	6.78	6.74	6.51	7.04
5	5.82	5.14	5.81	5.82	5.88
Total	52.15	56.23	50.43	59.82	52.09

The first five principal component analysis modes of short rainfall season sea level pressure

Appendix E

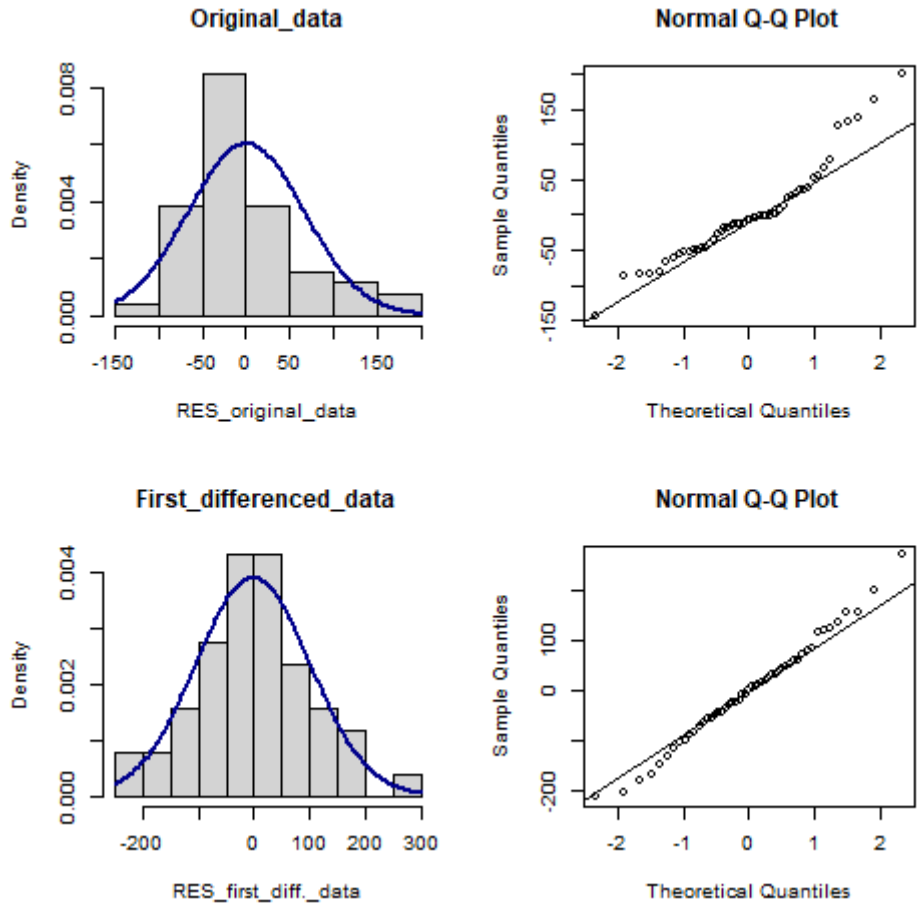


Appendix E.1: Residual's normality test for MLR long rainfall season

	Original data			**First differenced data**		
	Summary statistics			Summary statistics		
Test Type	X-squared	Df	p-value	X-squared	Df	p-value
Jarque Bera Test	0.008663	2	0.9957	0.23039	2	0.8912

Appendix E.2: Residuals Jarque Bera Test for long rainfall MLR model

Appendix F

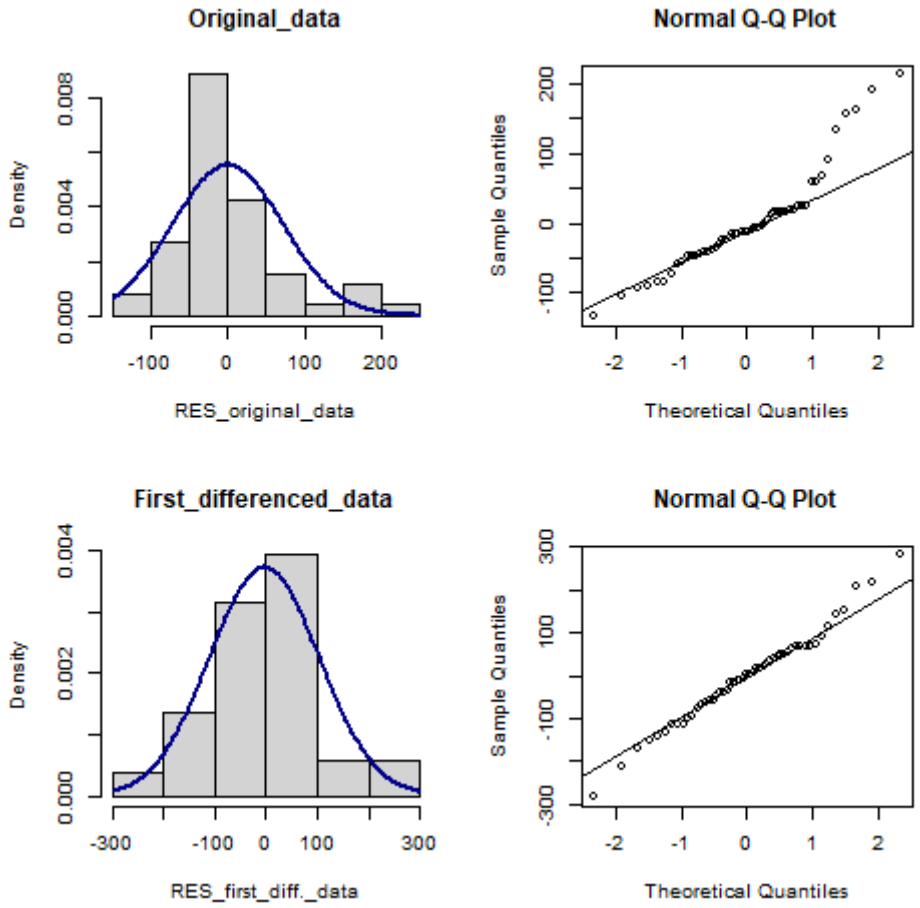


Appendix F.1: Residual's normality test for MLR short rainfall season (Statistical model)

	Original data			**First differenced data**		
	Summary statistics			Summary statistics		
Test Type	X-squared	Df	p-value	X-squared	Df	p-value
Jarque Bera Test	10.808	2	0.0044	0.22006	2	0.8958

Appendix F.2: Residuals Jarque Bera Test for short rainfall MLR statistical model

Appendix G

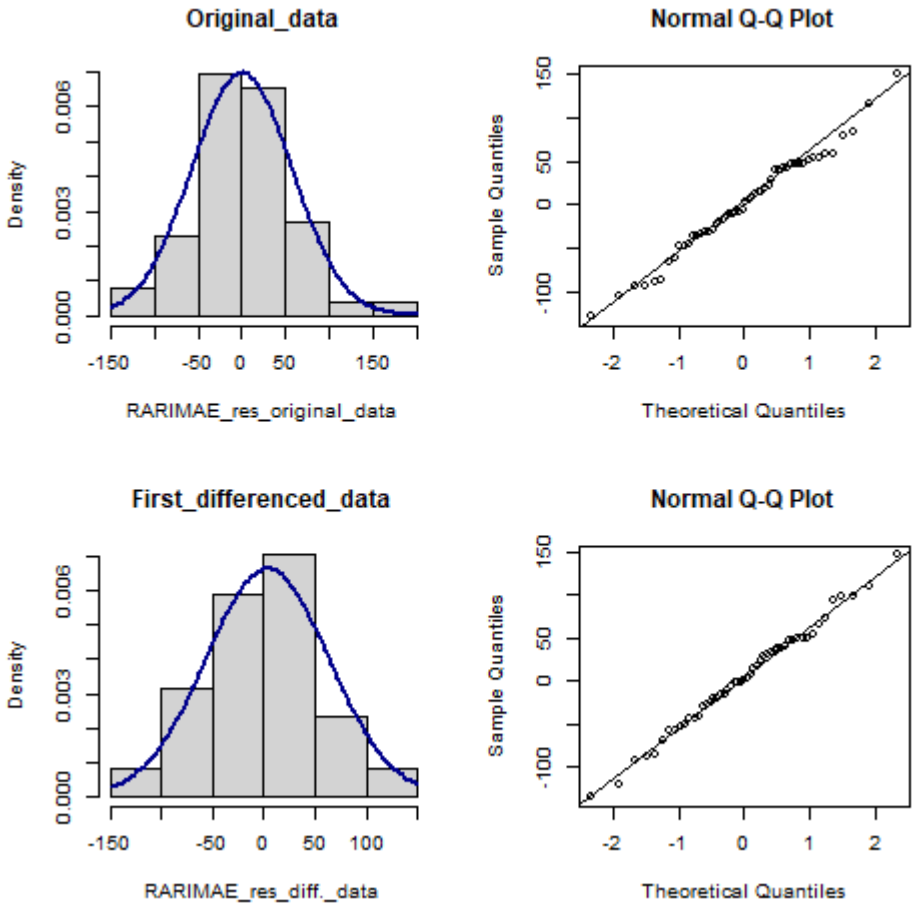


Appendix G.1: Residual's normality test for MLR short rainfall season (Predictive model)

	Original data			**First differenced data**		
	Summary statistics			Summary statistics		
Test Type	X-squared	Df	p-value	X-squared	Df	p-value
Jarque Bera Test	15.176	2	0.0005	0.63429	2	0.7282

Appendix G.2: Residuals Jarque Bera Test for short rainfall MLR predictive model

Appendix H

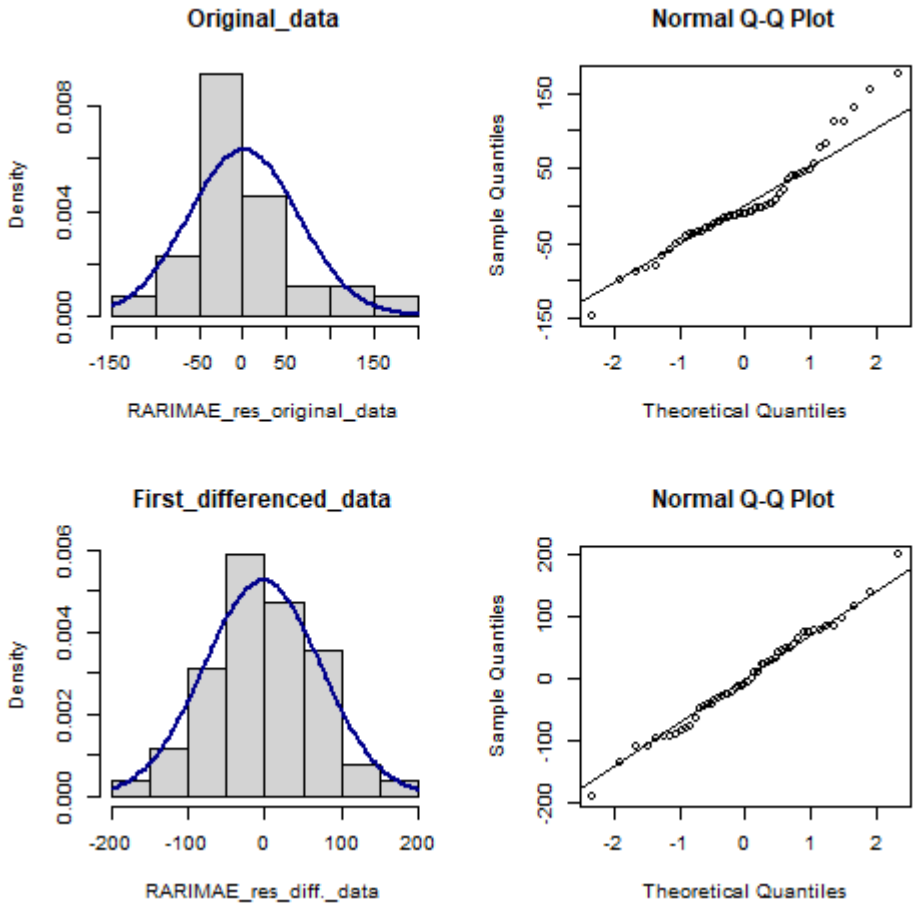


Appendix H.1: Residual's normality test for RARIMAE long rainfall season

	Original data			**First differenced data**		
	Summary statistics			Summary statistics		
Test Type	X-squared	Df	p-value	X-squared	Df	p-value
Jarque Bera Test	0.008663	2	0.9957	0.080654	2	0.9605

Appendix H.2: Residuals Jarque Bera Test for long rainfall RARIMAE model

Appendix I

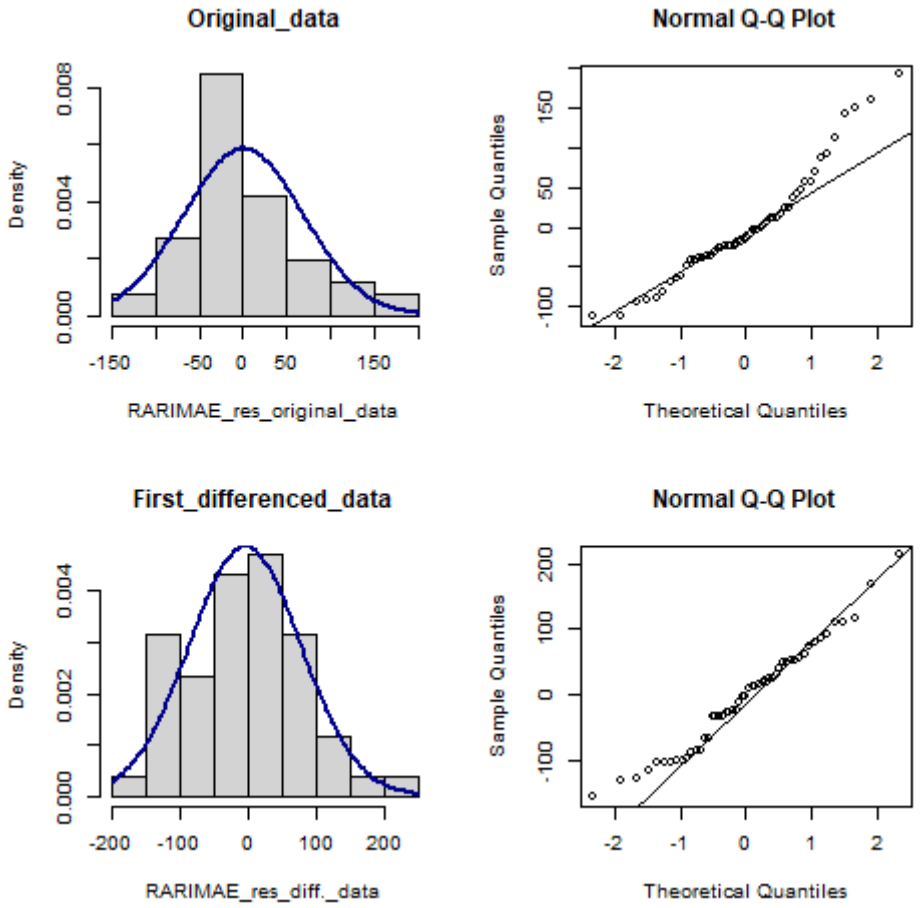


Appendix I.1: Residual's normality test for RARIMAE Short rainfall season (statistical model)

	Original data			**First differenced data**		
	Summary statistics			Summary statistics		
Test Type	X-squared	Df	p-value	X-squared	Df	p-value
Jarque Bera Test	3.6497	2	0.1612	0.072548	2	0.9644

Appendix I.2: Residuals Jarque Bera Test for short rainfall RARIMAE statistical model

Appendix J



Appendix J.1: Residual's normality test for RARIMAE Short rainfall season (predictive model)

	Original data			**First differenced data**		
	Summary statistics			Summary statistics		
Test Type	X-squared	Df	p-value	X-squared	Df	p-value
Jarque Bera Test	6.9583	2	0.0308	0.82344	2	0.6625

Appendix J.2: Residuals Jarque Bera Test for short rainfall RARIMAE predictive model