KYAMBOGO UNIVERSITY

A Time Series Forecasting Model for Rainfall in Kasese district, Uganda using the SARIMA approach

Kaluya Joshua

22/U/GMAPM/254/PE

A Dissertation Submitted to the Directorate of Research

and Graduate Training in Partial Fulfillment for the

Requirements of the Award of the Degree of Master of Science in

Applied Mathematics of Kyambogo University

July, 2025

Declaration

I, Kaluya Joshua hereby declare that this dissertation is my own work and has not been previously submitted for any qualification to any university or institution of higher learning.

Signature: . . . .

Date: . . . .

Approval by Supervisors

This dissertation was examined and approved for the award of the degree, with the guidance and approval of the following supervisors.

1. Dr. Awichi Richard

   Signature: ....

   Date: ....

2. Dr. Mukalazi Herbert

   Signature: ....

   Date: ....

Dedication

I dedicate this dissertation to my dear parents: Wampande George William Thomas and Namayiga Tapenensi. I also dedicate it to my beloved wife: Namulinda Rebecca Kaluya and to my brother Kategere James.

Acknowledgment

I would like to express my sincere gratitude to my supervisors: Dr Awichi Richard and Dr Mukalazi Herbert for their indispensable guidance, encouragement, and support throughout the entire process of this research. Their expertise, insightful feedback, and unwavering commitment have been so instrumental in shaping the direction and outcomes of this dissertation. I deeply appreciate Kyambogo University for providing the necessary resources and facilities that helped in the smooth progress of this research. The academic environment, University library have really been of great impact in this research.

I also extend my heartfelt appreciation to my family for their unconditional love, understanding, and encouragement. Their belief in my abilities has been a great source of strength and motivation.

I am also grateful to my friends and colleagues for their encouragement, constructive discussions, and moral support throughout this journey. Their fellowship and companionship have made the research journey enjoyable and memorable.

Finally, I would like to express my gratitude to all the lecturers who generously shared their time and insights, without whom this research would have not been possible.

Contents

List of Figures

List of Tables

List of Acronyms

UNMA	Uganda National Meteorological Authority
LVB	Lake Victoria Basin
NWP	Numerical Weather Prediction
WRF	Weather Research and Forecasting
GFS	Global Forecast System
MLE	Maximum Likelihood Estimation
AIC	Akaike Information Criterion
BIC	Bayesian Information Criterion
LR	Linear Regresion
SLR	Simple Linear Regression
MLR	Multiple Linear Regression
GLMs	Generalised Linear Models
DLMs	Dynamic Linear Models
IQR	Interquartile Range
SARIMA	Seasonal AutoRegressive Integrated Moving Average
SARIMAX	Seasonal AutoRegressive Integrated Moving Average with eXogenous Variables
ARIMA	AutoRegressive Integrated Moving Average
AR	AutoRegressive
MA	Moving Average
ACF	Autocorrelation Function
PACF	Partial Autocorrelation Function
ADF	Augmented Dickey-Fuller
ME	Mean Error
MAE	Mean Absolute Error

MSE	Mean Squared Error
RMSE	Root Mean Squared Error
MASE	Mean Absolute Scaled Error
WRF	Weather Research and Forecasting
ANN	Artificial Neural Networks

Abstract

The study developed a time series forecasting model aimed at predicting monthly rainfall in Kasese district, Uganda. Rainfall patterns play a critical role in agriculture, water management, and disaster preparedness in the region. Using data from January 1960 to December 2023 obtained from the Uganda National Meteorological Authority (UNMA), the study employed a Seasonal Autoregressive Integrated Moving Average (SARIMA) model to analyze and predict rainfall trends.

Preliminary analysis revealed that the data was non-stationary, as confirmed by the Augmented Dickey-Fuller (ADF) test (test statistic = -8.56, p-value = 0.01), and showed significant autocorrelation based on the ACF and PACF plots. The analysis showed that the months of March–April–May (MAM) and September–October–November (SON), with a mean monthly rainfall of approximately 118.03 mm (95% confidence interval: 111.32–124.74 mm) receive the highest amount if rainfall. In contrast, the driest months—January, February, June, and July—had a mean monthly rainfall of about 47.53 mm (95% confidence interval: 43.02–52.04 mm).

A SARIMA(3,1,1)(1,0,0)[12] model was selected based on the Akaike Information Criterion (AIC = 7948.98) and the Bayesian Information Criterion (BIC = 7976.55). A 12 - month seasonal model was used to capture monthly rainfall variations throughout the year, despite Uganda's two main rainy seasons. The model demonstrated good predictive accuracy, achieving a MAE of 42.5962, RMSE of 54.972, and MASE of 0.8376. Residual analysis confirmed that the model adequately captured the seasonal and trend components without significant autocorrelation.

The study concluded that the SARIMA model provided reliable short-term forecasts of monthly rainfall in Kasese, supporting agricultural planning and disaster risk reduction. The research recommended adopting of the model by local authorities, for short - term forecasts to improve agricultural planning, and also hold workshops and training sessions to educate the people on how to use and interpret the model forecasts and predict rainfall trends.

Chapter 1: Introduction

1.1 General introduction

Kasese is one of the regions in Uganda that are vulnerable to floods and landslides. These devastating natural disasters claim lives, destroy peoples homes, damage infrastructures and disrupt development in the region, which calls for an urgent need for mitigation strategies. Therefore, conducting research on rainfall forecasting for the region, contributes to the solutions to address the above challenges. When there are reliable rainfall predictions, then proactive strategies will be encouraged, which ultimately addresses the challenges posed by varying patterns of rainfall in the region.

1.2 Background

Rainfall forecasting is crucial globally, which helps in effective agricultural planning, disaster preparedness, and water resource management (Agyekum, 2022). Across the world, precise rainfall predictions help mitigate the effects of extreme weather events, droughts, and floods, ensuring better resource management in regions that depend heavily on rain for agriculture. In Africa, unpredictable rainfall patterns have a significant impacts on agricultural productivity, food security, and economic stability (Rankoana, 2020). Many countries in Sub-Saharan Africa are particularly vulnerable to rainfall variability due to their reliance on rain for farming.

Recent reports by the World Bank (2021) and the World Meteorological Organization (Kennedy et al., 2021), showed that the global average annual rainfall is approximately 990 mm, although it varies greatly by region. In Uganda, the national average annual rainfall ranges between 1,200 mm and 1,500 mm depending on location, with higher amounts in mountainous areas and lower in the north-east (Nsubuga et al., 2014; World Bank, 2020). For rain-fed agriculture, studies have indicated that most staple crops require a minimum of about 500–800 mm of well-distributed rainfall per year to achieve sustainable yields (Mekouar, 2018). However, Uganda has experienced increasing variability in rainfall patterns over the past two decades, with some regions receiving shorter rainy seasons and more intense downpours, raising concerns about agricultural productivity and disaster risk (Nsubuga & Rautenbach, 2018; Oriangi et al., 2024). These trends underscore the need for improved rainfall forecasting models to better support planning and resilience efforts.

Rainfall forecasting also plays a crucial role in East Africa by enabling communities to prepare for and respond to varying weather conditions effectively. However, many countries like Uganda face challenges in dealing with variable seasonal rainfall patterns and extreme weather conditions, such as landslides and floods. These irregular patterns and severe events stress the need for reliable rainfall forecasts to protect lives and properties.

In Uganda, the agricultural sector plays a vital role in the countrys' economy, and the ability to predict rainfall patterns is crucial for maintaining productivity. Reliable prediction of seasonal rains, especially in regions with complex terrain and climate patterns like Kasese (Momani, 2009) is very important. Kasese is particularly vulnerable to floods and land slides. Therefore, accurate rainfall forecasting plays a vital role in the agricultural sector of Kasese, where floods caused by shallow soils and limited water infiltration capacity require precise predictions to mitigate disaster risks and optimize agricultural productivity. The region's susceptibility to floods due to shallow soils and limited water infiltration capacity necessitates reliable prediction of extreme events (Alhassoun, 2011; Tibara et al., 2022). Existing forecasting methods, while helpful, often have limitations in capturing the complex dynamics of rainfall patterns in the region (Momani, 2009; Ali et al., 2021). This research focused on enhancing the existing SARIMA (Seasonal Autoregressive Integrated Moving Average) model to improve rainfall forecasting in Kasese. The SARIMA model holds promise for this research due to its specific strengths. Firstly, the model has demonstrated effectiveness in analysing historical rainfall data, which formed the foundation of this study (Momani, 2009). Secondly, the model's ability to account for seasonality is well-suited for Kasese, which experiences distinct seasonal rainfall patterns.

The Seasonal Autoregressive Integrated Moving Average (SARIMA) model is a powerful technique that is used in time series forecasting to capture and predict the seasonal patterns in the data. It's an extension of the ARIMA models (Box et al., 2015). It comprises of the seasonal autoregressive (SAR) and moving average (SMA) terms together with their corresponding non-seasonal components i.e, autoregressive (AR) and moving average (MA) components. Additionally, it integrates both the non-seasonal differencing (d) and the seasonal differencing (D) which helps to make the data stationary. By incorporating seasonal and non-seasonal components along with differencing, the SARIMA model effectively captures the complex seasonal variations and long-term trends present in the data. This enabled generation of reliable predictions of rainfall amounts, timing and intensity. SARIMA models effectively analyse time series data by capturing complex patterns such as trends, cycles and seasonality.

Improved rainfall forecasts hold significant potential benefits for Kasese. Farmers can optimize planting schedules, irrigation practices, and resource allocation based on reliable forecasts. Timely knowledge of potential extreme rainfall events allows authorities and communities to implement flood mitigation measures and ensure safety. Reliable forecasts also guide water management strategies for optimal and sustainable water usage throughout the year.

Despite the existence of various rainfall forecasting methods, the limitations in capturing the complex rainfall dynamics of the region necessitate further study. This research proposed and justified the refining and optimizing of the SARIMA model as a potential solution to improve the accuracy and reliability of rainfall forecasts in Kasese, Uganda. The anticipated benefits across agriculture, disaster preparedness, and water resource management highlight the importance of pursuing this research and its potential positive impact on the region.

1.2.1 Overview of Rainfall Patterns in Uganda

Uganda experiences a bimodal rainfall pattern in most regions, characterized by two rainy seasons: March to May (MAM) and September to November (SON). However, the amount, onset, and duration of rainfall vary significantly across the country due to topographical diversity and climatic influences such as the Intertropical Convergence Zone (ITCZ) and large water bodies like Lake Victoria (Mugume et al., 2017; UBOS, 2016). Central and southern regions receive relatively more consistent rainfall, while areas such as the northeast suffer from frequent droughts and erratic rainfall patterns.

1.2.2 Specific Concerns in Kasese District

Kasese District, located in Western Uganda, lies along the Rwenzori Mountain ranges and has historically experienced intense and highly variable rainfall patterns. These variations have increased the frequency of natural disasters, including floods and landslides, particularly in mountainous sub-counties such as Kilembe and Maliba (Alhassoun & Rajeh, 2011; Tibara et al., 2022). The district's vulnerability is exacerbated by human settlement in flood-prone areas, deforestation, and limited adaptive infrastructure.

1.2.3 Impact of Erratic Rainfall

Erratic rainfall has dire consequences on agriculture, food security, infrastructure, and public health in Kasese. Sudden heavy downpours lead to flash floods, destroying crops, roads, and homes, while prolonged dry spells reduce agricultural productivity and water availability. For instance, in 2020, heavy rainfall caused the Nyamwamba River to flood, displacing thousands of residents and damaging Kilembe Mines Hospital (Wasswa & Semakula, 2022). These events highlight the vulnerability of communities whose incomes depend on rainfall and other weather patterns.

1.2.4 Importance and Challenges of Rainfall Forecasting

Accurate rainfall forecasting is essential for disaster preparedness, agricultural planning, and water resource management. It enables early warning systems that can save lives and property. However, forecasting rainfall in Uganda—especially in mountainous regions like Kasese—remains a challenge due to limited ground-based observational data, complex terrain, and insufficient integration of local knowledge with scientific modeling approaches (Pazvakawambwa, 2017; Liu et al., 2022). The development of reliable time series forecasting models tailored to specific regions is therefore a pressing need.

1.3 Statement of the problem

Rainfall forecasting in Kasese District remains a critical challenge due to the inability of traditional methods to capture the complex temporal and seasonal dynamics of local rainfall patterns. These fluctuations significantly affect sectors such as agriculture, transportation, public health, and human settlement—particularly through recurrent floods and landslides (Tibara et al., 2022). The limitations of models like Multiple Linear Regression (MLR), as noted by (Ali et al., 2021), lie in their inability to incorporate non-linear components such as trend, seasonality, and autocorrelation (Montgomery et al., 2015; Shumway & Stoffer, 2017). To overcome these gaps, this study proposed the use of a Seasonal Autoregressive Integrated Moving Average (SARIMA) model, which has been shown to handle time-dependent structures more effectively (Box et al., 2015). By tailoring the SARIMA model to rainfall data specific to Kasese, the study aims to generate more accurate forecasts that can support local planning, disaster risk reduction, and sustainable resource management.

1.4 Objectives

1.4.1 Main objective

The main objective of this study was to develop a rainfall forecasting model for Kasese district, Uganda.

1.4.2 Specific objectives

In order to achieve the general objective of this research, the following specific objectives guided the study:

(i) Decompose the time series data into trend, seasonality and residual components to reveal underlying patterns and guide model development.

(ii) Determine the optimal solution of the model parameters for the rainfall forecasting model.

(iii) Forecast monthly rainfall using the developed model.

1.5 Scope of the study

1.5.1 Content scope

In line with the specific objectives of the study, the essential rainfall parameters of interest that guided in understanding of the rainfall patterns, were: rainfall amount, intensity, duration, frequency (return period), trends and seasonal distribution.

1.5.2 Geographical scope

The study area was Kasese district in western Uganda. Kasese district is located in the extreme western part of Uganda, in the Rwenzori region along the equator. It lies approximately 343km, by road, West of Kampala, the capital and biggest city of Uganda (www.DistanceCalculator.net). It's neighbouring districts are: Kabarole district in the North East; Bundibugyo district and Bunyangabu district in the North; Rubirizi district in the South; Kamwenge district and Kitagwenda district in the East; and Democratic Republic of Congo (DRC) in the West. Its neighboured by lake Edward in the South and lake George in East. It lies between the latitudes 0^o $12'$S and 0^o 26N; longitudes 29^o $42'$E and 30^o $18'$E (Uganda & WWF, 2013; UBOS, 2021). It occupies an area of approximately 3389.8 square kilometers; of which 2911.3 square kilometers (about 86%) is dry land, 409.7 square kilometer (about 12%) is open water, and 68.8 square kilometers (2%) is covered by wetlands. (Uganda & WWF, 2013; UBOS, 2020). The map of Kasese, district is as shown in Figure 1.1.

Kasese in particular was considered due to its shallow soils and rocky surfaces, which are associated with low infiltration rate, hence being more susceptible to floods when heavy rains occur (Alhassoun & Rajeh, 2011). Also the region is dominated by huge rock particles which block the river streams and create artificial dams upstream. Thus,

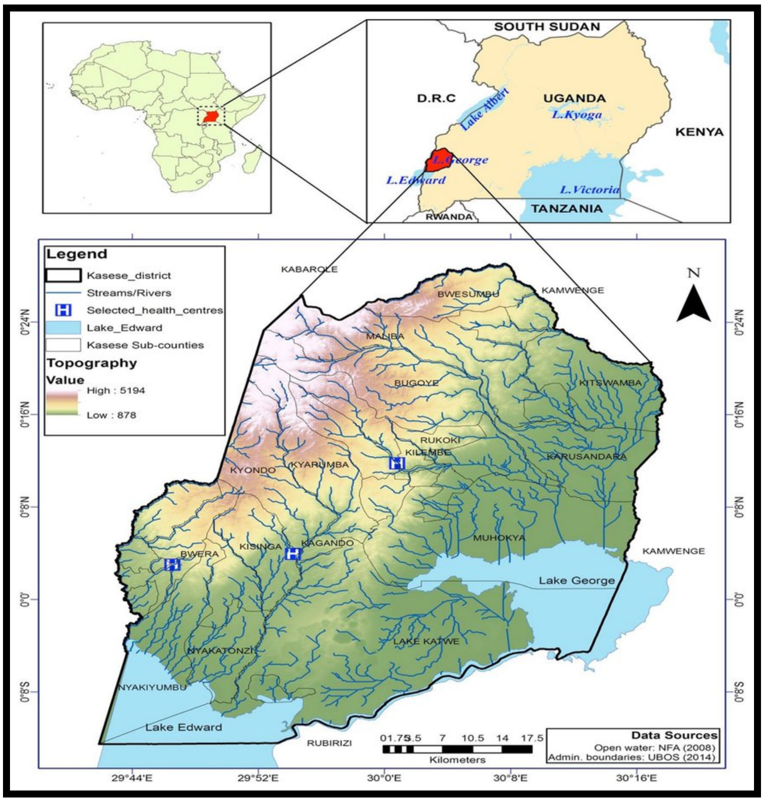


Figure 1.1: A Map of Kasese district, Uganda.

when the amount of water increases, it's most likely to cause a lot of huge destruction along it's path downstream (Tibara et al., 2022).

1.5.3 Time scope

The study used secondary historical rainfall data, obtained from the Uganda National Meteorological Authority (UNMA) for a period of 64 years starting from January 1963 to December 2023. This period provides a comprehensive overview of the rainfall patterns in Uganda.

1.6 Significance of the study

Forecasting of rainfall has a great significance in societal planning and making informed decisions. It's also crucial to many sectors like health, disaster preparedness, agriculture and transportation. Therefore, the study of developing a time series model for forecasting rainfall in Uganda will have significant implications as discussed below:

(i) Disaster preparedness: Understanding rainfall patterns will help in predicting and mitigating the effects of extreme weather events like floods, and landslides. Decision makers will also be able to allocate resources and provide prompt support early enough.

(ii) Agricultural planing: Reliable rainfall forecasts will help farmers in planning and preparing crops to cultivate, harvesting techniques, irrigation methods and pest control. This will enhance agricultural productivity and food security.

(iii) Prediction and prevention of diseases like malaria, and cholera: By carrying out analysis of historical rainfall data, the patterns and trends associated with high amounts of rainfall can be identified. Thus, developing an reliable time series model can assist in predicting of such disease outbreaks, and this will guide the health personnel in devising means for early intervention.

(iv) For effective water resource planning, management and conservation like construction of dams, reservoirs, irrigation systems and water supply infrastructures: The time series model developed will help in understanding of the seasonal rainfall patterns and trends in Kasese, Uganda which can guide the water managers to predict water levels and make informed decisions on reservoir release rates, which ensures a sustainable water supply for various purposes.

(v) Environmental management and conservation: Rainfall trends and patterns affect the health of ecosystems and the biodiversity. Therefore, understanding these patterns and trends will aid in assessing the impact of changes in climate and this will guide in designing conservation strategies to protect fragile ecosystems and vulnerable species.

(vi) Further research: The study will serve as a basis for further research investigations, providing valuable information on weather patterns, aiding in predictive analysis, and driving continuous improvements in the methods and strategies of modeling rainfall.

1.7 Justification of the study

Developing a time series forecasting model for rainfall in Kasese, Uganda is essential to address the challenges of rainfall fluctuations in the region. Since the region heavily relies on agriculture (UBOS, 2017), reliable rainfall predictions will enable the farmers to optimise the planting and harvesting times which boosts their agricultural production and incomes. The proposed model was intended to address the need for reliable forecasts to minimise the risks linked to floods and landslides that disrupts planting and harvesting time and also resulting in loss of lives and infrastructure damage. The study is relevant not only to address the challenges in the region but also contribute to a broader discussion on climate changes in similar regions.

1.8 Organisation of the report

This dissertation is organized into five chapters. Chapter One introduces the study, outlining the background, problem statement, objectives, scope, significance, and justification. Chapter Two presents the literature review related to rainfall forecasting models and highlights research gaps. Chapter Three describes the methodology used to develop and validate the SARIMA model. Chapter Four provides the presentation and discussion of results. Finally, Chapter Five summarizes the findings, draws conclusions, and offers recommendations.

1.9 Limitations of the Study

Although the study had access to a comprehensive dataset from UNMA covering over 60 years (from Jan 1960 to Dec 2023), the analysis focused only on monthly rainfall data without incorporating other climatic variables such as temperature and humidity, which could influence rainfall patterns. Additionally, the forecasting model was applied specifically to Kasese District, which may limit the generalizability of the findings to other regions with different climatic characteristics.

Chapter 2: Literature Review

2.1 Introduction

This chapter presented an in-depth insight into the current state of research about rainfall forecasting models. It was structured to address the specific objectives of the study: decomposing the time series data, determining optimal solution of the model parameters, and forecasting monthly rainfall using the developed model. The review highlighted relevant studies, methods, and models applied to similar problems, providing context for this study

2.2 Decomposing the time series data into trend, seasonality and residual Components

Understanding the components of a time series is critical in forecasting, as it helps reveal underlying patterns that guide model development. Decomposition separates data into trend, seasonal, and residual (noise) components, making it easier to identify and model recurring patterns in the data.

Studies such as (Box et al., 2015) emphasize that seasonal decomposition is essential in uncovering patterns hidden in the data, which directly supports the first specific objective of this study. Further, (Shumway & Stoffer, 2017; Montgomery et al., 2015) also highlight that decomposition enables better model specification and selection, especially when working with rainfall data showing strong seasonality, like that of Kasese district.

By focusing on decomposition, this study aligns with existing research practices to build a model that can capture trend and seasonal variations specific to the study area.

2.3 Determining optimal solution of the model Parameters

Selecting the best-fitting model parameters is key to obtaining accurate forecasts. This relates directly to the second specific objective of the study.

2.3.1 Seasonal Autoregressive Integrated Moving Average (SARIMA) model

Seasonal Autoregressive Integrated Moving Average (SARIMA) model is a time series forecasting model that is used when data exhibits some periodic patterns or seasonality. It's an extension of the ARIMA model whereby it incorporates both non-seasonal and seasonal components, which enables it to capture the repeating patterns in the data (Box et al., 2015).

By taking into account the seasonality, SARIMA models can provide more reliable and accurate predictions for a time series data that exhibits seasonal patterns. When the data exhibits seasonality, a $SARIMA(p,d,q)(P,D,Q)s$ model can be constructed.
where: s is the length of the season; p,d, and q are the non-seasonal parameters; P,D, and Q are the seasonal parameters; and then p is the order of the non-seasonal AR model; d is the non-seasonal differencing order (the number of times we need to difference the data to achieve stationarity); q is the order of the non-seasonal MA model; P is the order of the seasonal AR model; D is the seasonal differencing order; Q is the order of the seasonal MA model.

SARIMA models offer various applications in forecasting rainfall as seen below:

(i) SARIMA models can help to identify and forecast long term trends in rainfall patterns. This is important for planning and decision making in many sectors such as agriculture, and water resource management.

(ii) When seasonal and autoregressive components are modeled, SARIMA models can assist in predicting extreme rainfall events that may lead to flooding. Hence early warning systems can be put in place and the preventive measures taken.

Several studies such as (Metrine et al., 2015), applied SARIMA models to rainfall data in Kenya, finding that using criteria like AIC and BIC helped in determining optimal parameters for accurate forecasting. A *SARIMA* model they developed gave the best fit to the data and was employed to predict the mean monthly rainfall over a two-year period from January 2015 to December 2016. They suggested that the model should be updated continuously in order to get accurate forecasts after December 2016.

In another study by (Momani, 2009), a *SARIMA* model was developed based on monthly rainfall data obtained for Amman Airport Station, Jordan. However, it was noted with limitations, and recommended further testing and parameter refinement. These studies demonstrate the importance of parameter selection through statistical tests and diagnostics, which this research also followed to optimize the SARIMA model for Kasese district.

2.3.2 Testing the accuracy of the model

Evaluating the accuracy of rainfall forecasting models is vital for validating the performance and reliability of these models. This section reviewed different accuracy metrics and challenges associated with assessing model performance. Some of the metrics used to determine the accuracy of a model are: Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Mean Absolute Percentage Error (MAPE). For example, (Wangome, 2022) used these metrics to evaluate the performance of a rainfall prediction model which used short-term neural networks.

2.3.3 Stochastic models

These are mathematical models that incorporate in randomness and uncertainty to simulate possible outcomes for a certain problem. Stochastic models, such as Markov chains models, incorporate randomness and require estimation of transition probabilities and other parameters. These models use a probability distribution to represent situations that have uncertainty (Pinsky and Karlin, 2010). For example, (Kigobe et al., 2011) explored the use of generalized linear models (GLMs) in stochastic rainfall forecasting, emphasizing parameter estimation methods. They found out that the GLMs were able to reproduce the properties of rainfall. In their conclusion, the models they developed were not robust to generate rainfall at ungauged places. Further, (Biao & Alamou, 2018) applied stochastic processes to describe and analyse the daily rainfall patterns in Benin.

Their analysis showed that the likelihood of having two or three consecutive dry days is the highest compared to other transitions. They discovered that by understanding the likelihood of future rainfall, decision-makers can plan better for water resource management, thereby mitigating the risks that come with uncertain weather patterns.

Stochastic models are very useful in forecasting rainfall as discussed below:

(i) They are used to develop models e.g ARIMA models, Markov chain models among others that incorporate in spatial and temporal variations in rainfall. This helps in understanding the dynamics of rainfall over space and time which leads to more precise and reliable rainfall predictions.

(ii) Stochastic models can be used to assess the probability of extreme rainfall events happening like the floods and landslides. This is essential in managing of the floods and to plan for infrastructures like the roads and bridges.

2.3.4 Numerical Weather Prediction (NWP) models

These are computer based simulations that use mathematical equations and algorithms to process the current weather observations and then later generate forecasts for future weather conditions. These models use complex simulations with parameters derived from current weather observations. The techniques used to optimize these parameters include data assimilation and calibration. For example, (Powers et al., 2017) discussed the role of NWP models in predicting extreme rainfall events and the importance of accurate initial parameter estimation.

The NWP models take into account factors such as wind, pressure, humidity, temperature, among others to make predictions of weather conditions (Powers et al., 2017). Examples of NWP model include, among others:

(i) Weather Research and Forecasting (WRF) models, which is commonly used by meteorological agencies like the Uganda National Meteorological Authority (UNMA) and some research institutions to predict weather conditions.

(ii) Global Forecast System (GFS) which provides global weather forecasts for a period of 16 days in advance.

Numerical Weather Prediction models are helpful in predicting extreme rainfall events like landslides, floods plus other extreme weather conditions such as heavy thunderstorms and the cyclones. NWP (WRF) models are used mostly by meteorological agencies like UNMA to provide short term rainfall forecasts, which could be hourly, daily, weekly or monthly. This helps people to plan for the day or month in advance.

2.4 Rainfall forecasting

The final objective of this study focused on forecasting monthly rainfall using the developed model. Forecasting rainfall accurately is essential for effective planning, resource management, and mitigating the impacts of extreme weather events, particularly in regions like Kasese, which are vulnerable to floods and landslides. Reliable rainfall predictions can assist in agricultural planning, water resource management, and disaster preparedness. This section explored the existing research on rainfall forecasting and the various models that have been developed and applied for forecasting rainfall. By examining both traditional statistical approaches like SARIMA and more complex models such as Numerical Weather Prediction (NWP), Dynamic Linear Models (DLMs), and Linear Regression (LR) models, the section highlights how these models are employed in practical scenarios to predict short-term and long-term rainfall patterns. Each model has its strengths and limitations, but their collective goal remains the same: providing accurate and timely rainfall forecasts to mitigate the challenges posed by unpredictable weather. Various studies show successful application of SLR, MLR, and SARIMA models to forecast rainfall, although they have limitations, as will be discussed later on in the next sections.

2.4.1 Rainfall forecasting models and their applications

Linear Regression (LR) models

These models lay the basis for predictive analysis, and assist in understanding the relationship between the dependent and independent variables. A linear regression model fits a straight line which represents the relationship between the variables. It's classified into two as discussed below:

(i) Simple linear regression (SLR): It models the relationship between two variables i.e. one independent variable, which is the predictor and one dependent variable, which is the outcome. SLR models have been applied to relate rainfall with a single explanatory variable. For example; in a study by (Chattopadhyay and Chattopadhyay, 2010) a SLR was used to predict monsoon rainfall in India based on sea surface temperature anomalies, finding moderate predictive skill.

    Although SLR models are straightforward and easy to interpret, studies consistently report that their predictive power is limited when rainfall is influenced by multiple climatic factors.

(ii) Multiple linear regression (MLR) which involves multiple variables. MLR models can be used to forecast monthly precipitations by incorporating in various atmospheric features such as temperature, wind, humidity, pressure among others which can influence rainfall. Linear regression analysis form the basis for more advanced time series forecasting methods such as the autoregressive (AR), moving average (MA) and autoregressive integrated moving average (ARIMA) models (Box, 2015; Montgomery et al., 2015).

    Multiple Linear Regression (MLR) models have been widely used to capture the effects of several predictors simultaneously, which can significantly improve forecasting accuracy. (Ali et al., 2021) developed an MLR model to forecast monthly rainfall over Kasese district, Uganda, using predictors such as temperature and humidity. Their study found that MLR better captured rainfall variability as compared to a SLR. They suggested that future studies could examine additional variables which may impact rainfall outcomes. In another study by (Khan et al., 2018) a MLR was applied to forecast monsoon rainfall in Pakistan by including large-scale climate predictors such as sea surface temperature, wind anomalies, and humidity, achieving higher forecast skill. In a study by (Navid & Niloy, 2018), they used a MLR model to predict rainfall for Bangladesh and they stated that MLR is effective and still widely used for rainfall prediction in Bangladesh, supporting agricultural planning despite the rise of more advanced models. In addition, (Gnanasankaran & Ramaraj, 2020) used a MLR to forecast monthly rainfall in India based on meteorological variables like pressure, wind, temperature, speed, and humidity. They concluded that MLR approach is a practical and an easy to interpret method for rainfall prediction that supports in planning and decision-making, though further accuracy improvements could be handled by using advanced machine learning models.

However, while MLR models can incorporate multiple explanatory variables, several authors (e.g., Ali et al., (2021); Gnanasankaran & Ramaraj, (2020) noted that their effectiveness decreases when data shows strong seasonality or non-linear relationships, which often motivates the use of specialized time series models like the SARIMA models.

2.4.2 Existing research on rainfall forecasting

Various studies around the globe have been carried out on rainfall forecasting and simulation using several techniques like SARIMA approach. For instance (Aliyu et al., 2021) used SARIMA (2,0,1)(2,1,1)[12] to model monthly rainfall data in Nigeria from 1980 to 2015). They found out that the SARIMA model outperformed Bayesian Structural Time Series methods, as it exhibited lower values of RMSE, MAE and MAPE, and their model offered reliable and robust forecasts. In another study, (Moahmed & Mahgoub, 2014) they used SARIMA (0,0,5)(1,0,1)[12] to simulate monthly rainfall values from 1971 to 2010 for Gadaref region, Sudan. They concluded that, it was well-suited for forecasting and modeling rainfall, helping in water management and planning. Further, (Grevazi, 2024) fitted a SARIMA (1,1,1)(0,1,1)[12] model using monthly rainfall data for Namtumbo District, Tanzania from 2000 to 2020). The study forecasted rainfall upto 2030 and recommended the model to be used for regional decision-making on water management and farming. In a another study by (Nirmala & Sundaram, 2010) a seasonal multivariate ARIMA model was developed to help in forecasting monthly rainfall in Tamilnadu, India, for the years 1950 to 2008. They applied the Box – Jenkins methodology, and incorporated Sea Surface Temperature (SST) as an exogenous predictor in the model. Their findings showed that the inclusion of SST significantly enhanced the forecasting accuracy of monthly rainfall in Tamilnadu, suggesting that their SARIMA model is a useful tool for rainfall prediction in the monsoon-influenced regions.

Several other approaches have been used to forecast rainfall. For example; (Opio et al., 2020), simulated extreme rainfall over the lake Victoria basin (in particular for Uganda) using the the Weather Research and Forecasting (WRF) model, for 20 days. The model successfully simulated the extreme rainfall over the Uganda's western lake Victoria basin. Mugume, (2017), conducted research to evaluate the performance of Weather Research and Forecasting (WRF) model in simulation of rainfall over the western part of Uganda.

In another study by (Ronald. O, 2019), he contributed to the prediction of extreme rainfall events over the lake Victoria basin (LVB) in Uganda using Weather Research and Forecasting (WRF) model and concluded that it can be used to simulate extreme rainfall events over the western lake Victoria basin.

Further, (Tuyizere et al., 2017) applied a Markov chain model to predict rainfall in Gasabo district of Rwanda. Thus according to them, such an approach can be used to provide and communicate the desired information about changes in climate and rainfall forecasting.

These examples illustrate how forecasting models are developed and refined based on local data to provide actionable insights, a practice this study followed for Kasese district, Uganda.

2.5 Research gaps

Despite extensive research on rainfall forecasting, gaps remain, particularly in adapting models to local conditions like those in Kasese district. Previous models, such as Multiple Linear Regression (MLR) models (Ali et al., 2021), often did not fully capture non-linear components or seasonality. This study aimed to address these gaps by:

(i) Decomposing the time series data to better understand underlying patterns.

(ii) Optimizing model parameters using statistical criteria.

(iii) Developing forecasts tailored specifically to Kasese district's rainfall patterns.

By focusing on these gaps, this study aimed to contribute to the accuracy of rainfall predictions in the region by fitting a *SARIMA* model to the data.

2.6 Conclusion

In summary, the reviewed literature highlights that decomposing time series data, determining optimal solution of model parameters, and forecasting using SARIMA models are effective strategies for rainfall forecasting. It also reveals that SARIMA models have been particularly effective in capturing seasonality in monthly rainfall data. While previous studies in related regions have shown both strengths and limitations, this study builds on these findings by addressing local climatic variability in Kasese District. By applying these strategies specifically to Kasese, this study aimed to contribute to more reliable forecasts that can inform disaster preparedness, agriculture, and resource management.The identified research gaps justify the need for a locally tailored forecasting model to improve accuracy and practical usefulness.

Chapter 3: Methodology

3.1 Introduction

This chapter explored the methodology used to develop a quantitative time series forecasting model for rainfall in Kasese, Uganda. This methodology was closely aligned according to the specific objectives of the study.

3.2 Data Collection

The study used secondary historical record of monthly rainfall values over a period of 64 years, from January 1960 to December 2023, for Kasese district, Uganda. The data was got from the Uganda National Meteorological Authority (UNMA) center in Luzira, Uganda. The department maintains the data set in excel sheet format on daily, monthly as well as yearly basis. However, the study used monthly rainfall data.

3.3 Software packages used

The study used microsoft excel and *R*. *R* is a powerful tool for statistical analysis and visualization, excellent for handling large datasets and performing complex analysis. Its key strength lies in a wide range of packages available, such as *ggplot*2 for data visualization, *dplyr* and *tidyr* for data manipulation, *forecast* for time series analysis, *tseries* for statistical tests on time series data, and *caret* for machine learning. These packages enhance R's functionality, making it a powerful tool for analyzing and interpreting data, especially in time series modeling and forecasting.

3.4 Research design

The study adopted a quantitative approach, and used secondary historical monthly rainfall data for Kasese, Uganda, from January 1960 to December 2023. The data was got from Uganda National Meteorological Authority (UNMA) and was split into a training set (used to build the model) and a testing set (used to validate the forecasting model). The *SARIMA* model, selected based on data seasonality and stationarity, was applied using *R* packages like $forecast$ and $tseries$. Stationarity was tested using the ADF test, and model accuracy was assessed through $RMSE$, MAE, $MASE$, and $MAPE$. The model was used for rainfall forecasting to help in disaster preparedness and agricultural planning.

3.5 Model assumptions

Before developing the SARIMA model, two key assumptions that guided the study were checked:

(i) The rainfall time series is non-stationary: confirmed by the Augmented Dickey–Fuller (ADF) test.

(ii) The series exhibits significant autocorrelation: confirmed by ACF and PACF plots showing clear autocorrelation patterns.

These results justified applying a SARIMA model to capture seasonality and trends in the data.

3.6 Methodology for specific objective (i)

3.6.1 Data pre-processing

Rigorous data pre-processing was conducted to ensure the quality and consistency of the historical rainfall data used in the study. This involved identifying and handling missing values, detecting and addressing outliers, and examining the data for existing trends or seasonality. Further, the data was transformed into a readable format (Microsoft Excel) and then further analysis was carried out.

3.6.2 SARIMA Model Specification

The Seasonal Autoregressive Integrated Moving Average (SARIMA) model was used to forecast rainfall in Kasese district. The general form of a $SARIMA(p,d,q)(P,D,Q)_s$ model is given in Equation (3.1).

$$\Phi_P(B^s)\phi(B)\Delta^d\Delta_s^D y_t = \mu + \psi_Q(B^s)\theta(B)w_t \tag{3.1}$$

where:

$\phi(B)$ is the non-seasonal AR component given by

$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p \tag{3.2}$$

$\theta(B)$ is the non-seasonal MA component given by

$$\theta(B) = 1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q \tag{3.3}$$

$\Phi_P(B^s)$ is the seasonal AR component given by

$$\Phi_P(B^s) = 1 - \Phi_1 B^s - \Phi_2 B^{2s} - \cdots - \Phi_P B^{Ps} \tag{3.4}$$

$\psi_Q(B^s)$ is the seasonal MA component given by

$$\psi_Q(B^s) = 1 + \psi_1 B^s + \psi_2 B^{2s} + \cdots + \psi_Q B^{Qs} \tag{3.5}$$

Additional terms:

$\Delta^d = (1 - B)^d$ is the non-seasonal differencing operator.

$\Delta_s^D = (1 - B^s)^D$ is the seasonal differencing operator.

s is the length of the seasonal period.

μ is the mean term (average rainfall).

B is the back-shift operator, where $B^k y_t = y_{t-k}$.

Different SARIMA variants with various order parameters (p, d, q) and seasonal parameters $(P, D, Q)_s$ were tested to identify the most suitable model for the rainfall data in Kasese.

3.6.3 Testing for Stationarity

Before applying the SARIMA model, it was necessary to test whether the rainfall time series was stationary. Stationarity implies that the statistical properties of the series (mean, variance) remain constant over time (Box et al., 2025; Kirchgssner et al., 2012), which is a key assumption for SARIMA modeling.

Visualisation techniques, such as plotting the series and examining autocorrelation functions, were used initially. To statistically confirm stationarity, the Augmented

Dickey-Fuller (ADF) test was conducted (Mushtaq & Rizwan, 2011).

The null hypothesis (H_0) of the ADF test states that the series has a unit root (i.e., it is non-stationary). The alternative hypothesis (H_a) is that the series is stationary. The ADF test statistic is calculated as:

$$t = \frac{\hat{\alpha}}{SE(\hat{\alpha})} \tag{3.6}$$

where:

$\hat{\alpha}$ is the estimated coefficient from the regression.

$SE(\hat{\alpha})$ is the standard error of the estimated coefficient.

Decision criteria:

If $|t|$ > critical value (or if p-value < chosen significance level, e.g., 0.05), reject H_0: the series is stationary.

If $|t|$ critical value (or if p-value > significance level), fail to reject H_0: the series is non-stationary, and differencing is required.

In this study, the `adf.test()` function from the `tseries` package in R was used to conduct the ADF test and guide the differencing process.

3.7 Methodology for specific objective (ii)

3.7.1 Model identification

Here, the order of the $SARIMA(p,d,q)(P,D,Q)s$ model was identified. This was done with the help of the ACF and PACF plots. The ACF plot was used to determine the non - seasonal $MA(q)$ order and the seasonal $MA(Q)$ order; and the PACF was used to determine the non - seasonal $AR(p)$ order and also the seasonal $AR(P)$ order, while considering the significant spikes in these plots. The "$forecast$" package in **R** together with the "$auto.arima()$" function were used to determine the order of $SARIMA(p,d,q)(P,D,Q)s$ model, since only the ACF and PACF plots cannot be relied on due to seasonal variations in the data.

3.7.2 Model selection

After pre-processing the data, statistical tests and model selection criteria, such as the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC), were used to guide in the selection of the optimal model parameters.

The *BIC* is a statistical method that is used in selecting the correct model among a set of available models. It helped to balance both the goodness of fit of the model to the data and its complexity in order to avoid over-fitting of the model (Neath et al., 2012).

The formula for the Bayesian Information Criterion is given by:

$$BIC = -2\log(L) + k\log(n) \tag{3.7}$$

where:

n is the sample size,

k is the number of estimated parameters in the model,

L is the maximised value of the likelihood function of the model.

The log-likelihood is defined as:

$$\log L(\theta) = \log f(x/\theta) \tag{3.8}$$

where x is the observed data and θ are parameter values of the model that maximises the likelihood function.

For multiple independent observations, the likelihood function is the product of individual likelihoods i.e.

$$L(\theta) = \prod_{i=1}^{n} f(x_i/\theta) \tag{3.9}$$

and the maximum likelihood estimation(MLE) is got by taking the natural logarithm Or just logarithm of the likelihood function(log-likelihood) i.e.

$$\log L(\theta) = \sum_{i=1}^{n} \log f(x_i/\theta) \tag{3.10}$$

However, with the help of an in-built "$stats$" package in **R**, the values of $L(\theta; y_t), \theta$, and BIC were obtained.

The *AIC* is a mathematical method for evaluating how well a model fits the data it was generated from. In statistics, AIC is used to compare different possible models and determine which one is the best fit for the data among competing models for a fixed data set (Akaike, 2011). AIC is calculated from:

(i) the number of independent variables used to build the model.

(ii) the maximum likelihood estimate of the model (how well the model reproduces the data) (Bevans, 2022).

The AIC is a measure of goodness of fit that takes the number of fitted parameters into account and is an effective method of choosing between a given set of models (Bozdogan, 1987). The chosen model is the one that minimizes the loss of information.

The AIC is described by equation;

$$AIC = 2k - 2ln\hat{L} \tag{3.11}$$

where, $\hat{L}$ denotes the maximum log - probability of the estimated model and is the likelihood evaluated in the estimator; k is the number of estimated parameters (independent variables) in the predicted model.
The AIC scores are often shown as ΔAIC scores. ΔAIC represents the difference between the AIC values of a given model and best model (model with the smallest AIC). Therefore the model with the lowest AIC was chosen as the best model among all models specified for the data used, and with the help of an in-built "$stats$" package in **R**, AIC was obtained.

Using the above two criterion, the values of p, q, P and Q were determined by selecting the model with the lowest BIC and AIC.

3.7.3 Diagnostic Checking

Once a candidate model was selected, it underwent rigorous diagnostic checks to validate its adequacy and the following assumptions about residuals were checked:

(i) Residuals are normally distributed: tested using histogram and normal Q–Q plot.

(ii) Residuals are not autocorrelated: tested by examining ACF/PACF plots of the residuals.

(iii) The model assumed a linear relationship in parameters: verified during model specification.

This process ensured that the model captured the essential dynamics of the data without significant bias.

3.7.4 Model fitting

The aim of fitting the model was to find values of the parameter that maximises the likelihood function and measure how well the model fits the observed data. This guided the study to estimate the parameters to fit the model.

For a SARIMA model, the likelihood function is given as:

$$L(\theta|y) = (2\pi\sigma^2)^{-n/2} exp\{-\frac{1}{2\sigma^2}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2\} \tag{3.12}$$

where:

y is the entire set of observed values: $y_1, y_2, ..., y_n$

y_t is the actual observed value of y at time t (e.g., rainfall in month t)

$\hat{y}_t$ is the predicted value of y at time t

σ represents the standard deviation of the residual errors

n is the total number of observation for the data collected

θ represents the parameters of the model being developed.

The Standard Deviation of the residuals can be calculated using the formula below:

$$\sigma = \sqrt{\frac{1}{n-m}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2} \tag{3.13}$$

where m is the number of estimated parameters of the model.

Using, the "*stats*" package in **R**, the standard deviation of the residuals will be obtained. Then the "*residuals*()" function will give the residuals of the model and the "$summary()$" function will provide the detailed information about the SARIMA model, which includes the estimated coefficients obtained through MLE.

After fitting the SARIMA model, the residuals were checked if they are all errors i.e. they are not correlated with each other and if at all their mean and variance is the same throughout.

3.7.5 Model accuracy

The performance of the model was determined using the mean absolute error (MAE), mean squared error (MSE), the root mean squared error (RMSE) and the mean absolute scaled error (MASE).

Mean Absolute Error (MAE) is the average of the absolute error (Brassington, 2017) and it's formula is given as:

$$MAE = \frac{1}{n}\sum_{t=1}^{n} |y_t - \hat{y}_t| \tag{3.14}$$

where:
n is the number of observations (sample size), y_t is the actual value and $\hat{y}_t$ is the predicted value.
The Mean Squared Error (MSE) refers to the average of the squared difference between the observed value and the predicted value (Brassington, 2017). The formula for the mean squared error is given as:

$$MSE = \frac{1}{n}\sum_{t=1}^{n} (y_t - \hat{y}_t)^2 \tag{3.15}$$

where y_t is the observed value, $\hat{y}_t$ is the corresponding predicted value of y_t and n is the number of observations (sample size) (Brassington, 2017).

The Root Mean Squared Error (RMSE) is the square root of MSE (Brassington, 2017).i.e.

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2} \tag{3.16}$$

The $MASE$ compares the mean absolute error of the model to the mean absolute error of the naive model. If the $MASE$ value is less than 1, then the model outperforms the naive model. If it's greater than 1, then the naive model performs better than forecasting model.

The $MASE$ is computed using the formula given in equation (3.11)

$$MASE = \frac{\frac{1}{n}\sum_{t=1}^{n}|y_t - \hat{y}_t|}{\frac{1}{n-1}\sum_{t=2}^{n}|y_t - y_{t-1}|} \tag{3.17}$$

where:

n is the number of observations

y_t is the observed value at time t

$\hat{y}_t$ is the forecasted value at time t

y_{t-1} is the actual value at the previous observation.

Basing on the above metrics, a model with a lower $MAE, MSE, MASE$ and $RMSE$ was selected as the best model, which was then used to make the predictions. From **R**, using the "$accuracy()$" function from the "$forecast$" package, the above metrics were obtained.

3.8 Methodology for specific objective (iii)

3.8.1 Forecasting

The developed *SARIMA* model was used to make forecasts for period 2011 to 2023. To determine the accuracy of the model, the forecasted values were compared with the actual values in the test data set. The metrics: Mean Absolute Error (MAE), Mean Squared Error (MSE) and the Root Mean Squared Error (RMSE) were used to choose the best model.

3.8.2 Model refining

To improve on the models' accuracy and reliability so that it best captures the underlying patterns in the data, model refining was carried out considering the following steps:

(i) Residuals were re - examined to ensure that they were all errors, so that the model fits well to the data.

(ii) The ACF and PACF plots were used to identify any more patterns or correlations in the residuals.

(iii) Adjusting both the non - seasonal parameters p,d,q and the seasonal parameters P,D,Q based on the diagnostic results was done. The main aim was to select optimal parameters for the best model performance. To achieve this, techniques such as AIC and BIC were used.

(iv) Updating the model with the new parameters.

(v) New forecasts were made and compared with the previous ones. The performance of the refined model was determined.

(vi) The metrics MAE, MSE, and RMSE were used to evaluate the performance of the model.

The above steps were repeated to achieve the best possible model. The refined SARIMA model was then selected as the final model for forecasting.

Chapter 4: Presentation and Discussion of Results

4.1 Introduction

This chapter presents the findings of the study and discusses their implications, structured according to the three specific objectives of the study:

(i) Decomposition of the time series data into trend, seasonality, and residual components to reveal underlying patterns;

(ii) Determining the optimal solution of the model parameters for the rainfall forecasting model; and

(iii) Forecasting monthly rainfall using the developed model.

4.2 Objective (i): Decomposition of the time series data

To achieve the first specific objective, the monthly rainfall data from January 1960 to December 2023 was thoroughly examined and decomposed into its components. Figure 4.1 shows the original time series plot, revealing trends and seasonal patterns. Figure 4.2 shows the box plot representing rainfall variability and potential outliers, highlighting months with typically higher or lower rainfall, from Jan 1960 to Dec 2023. The data was then decomposed into observed, trend, seasonal, and random components as shown in Figure 4.3. The observed part shows the original time series data, showing the overall rainfall patterns over the period. The trend component showed long-term fluctuations, while the seasonal component captured predictable annual cycles. The random component represented irregular variations. These decomposed parts provided insight into the rainfall dynamics in Kasese district, fulfilling the first objective.

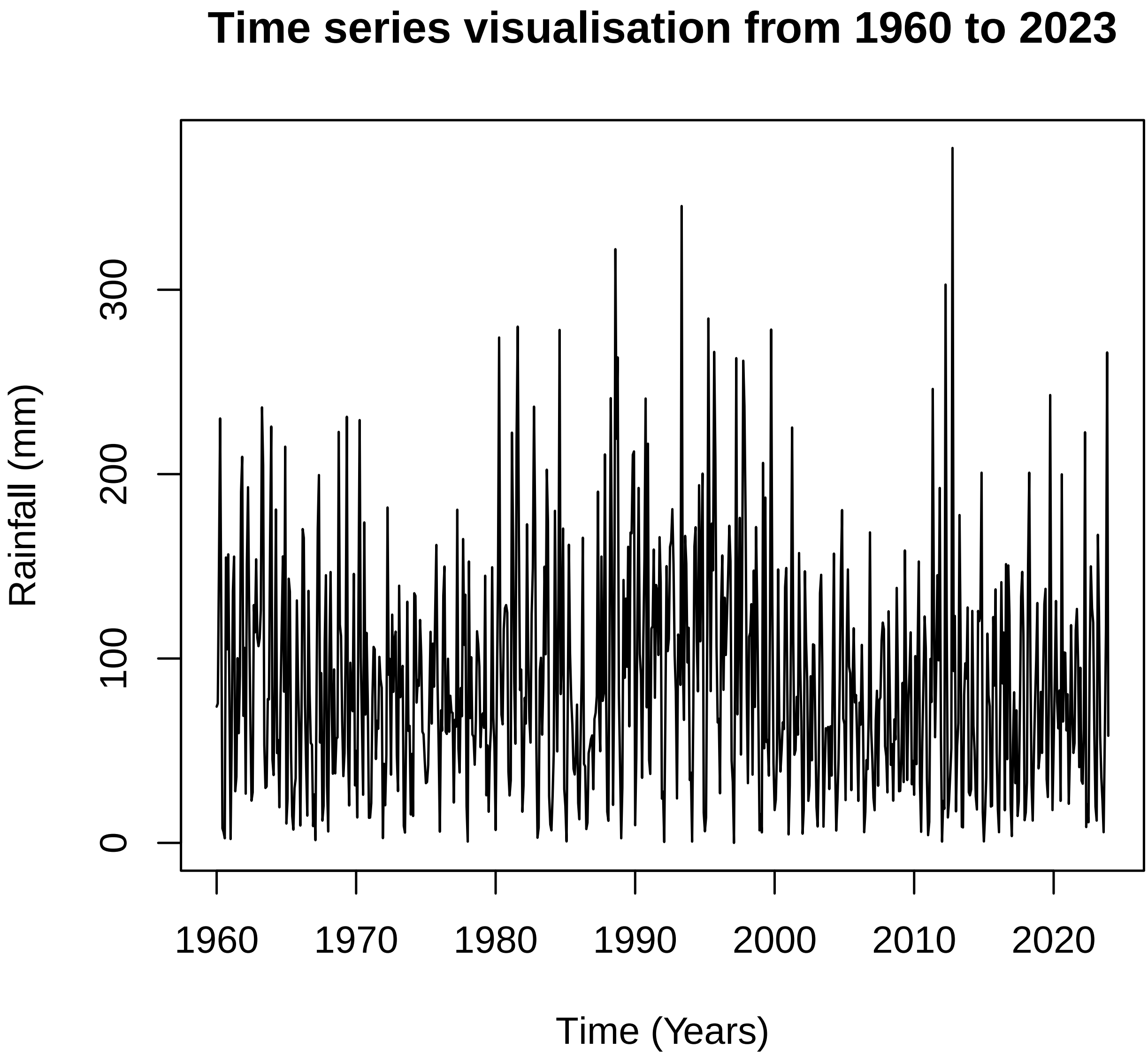


Figure 4.1: Time series plot for monthly rainfall values.

4.3 Objective (ii): Determining optimal solution of the model parameters

The second objective focused on identifying the best-fitting parameters for the forecasting model.

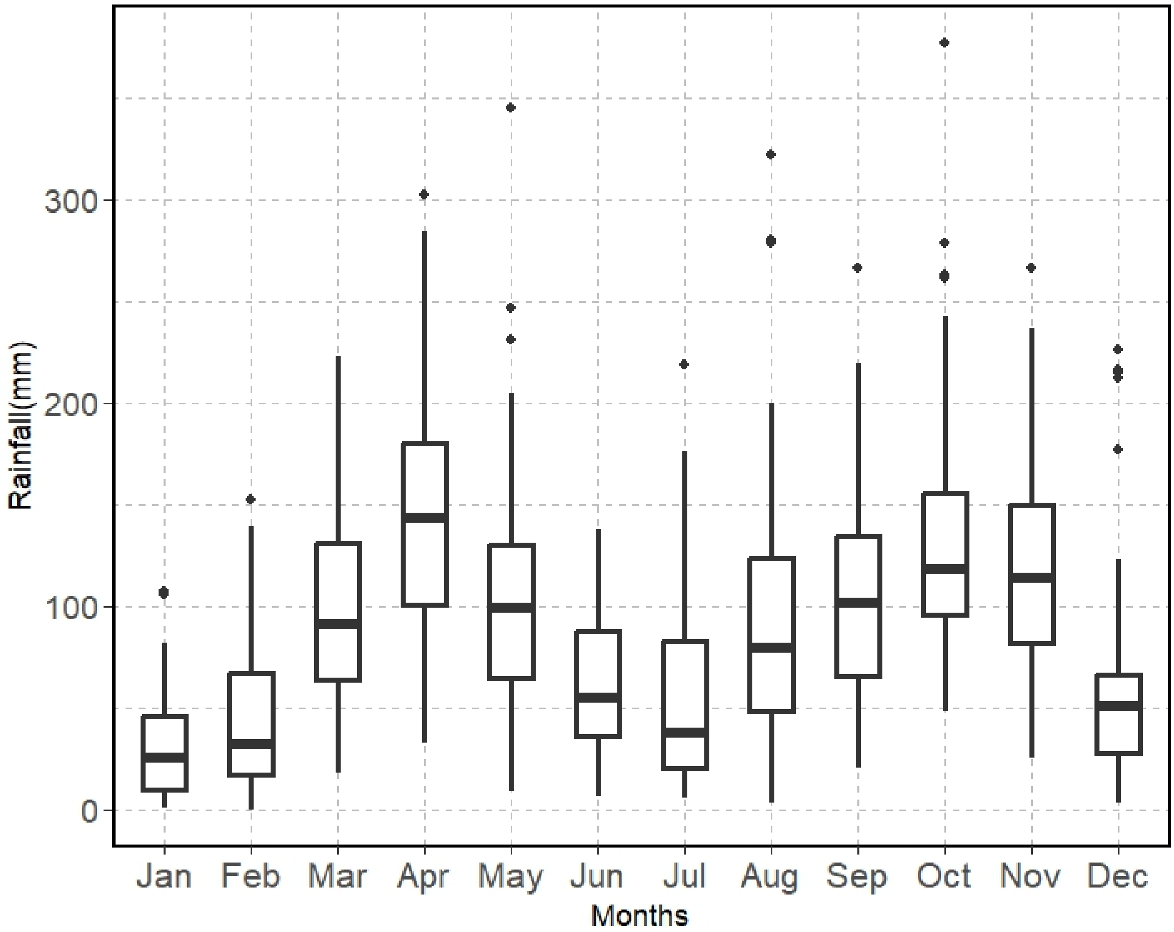


Figure 4.2: Box plot of the monthly rainfall values.

4.3.1 Preliminary data analysis using ACF and PACF plots

ACF and PACF plots (Figure 4.4) helped to understand some properties of the time series data, identify potential AR and MA components, indicating seasonality and autocorrelation patterns.

4.3.2 Stationarity test

A stationarity test was carried out using the ADF test to determine whether the time series data really exhibited some trends and/or seasonal patterns, to ensure reliability and accuracy of the results. From the analysis, the ADF test gave a value of -8.5557 and the p-value was 0.01.

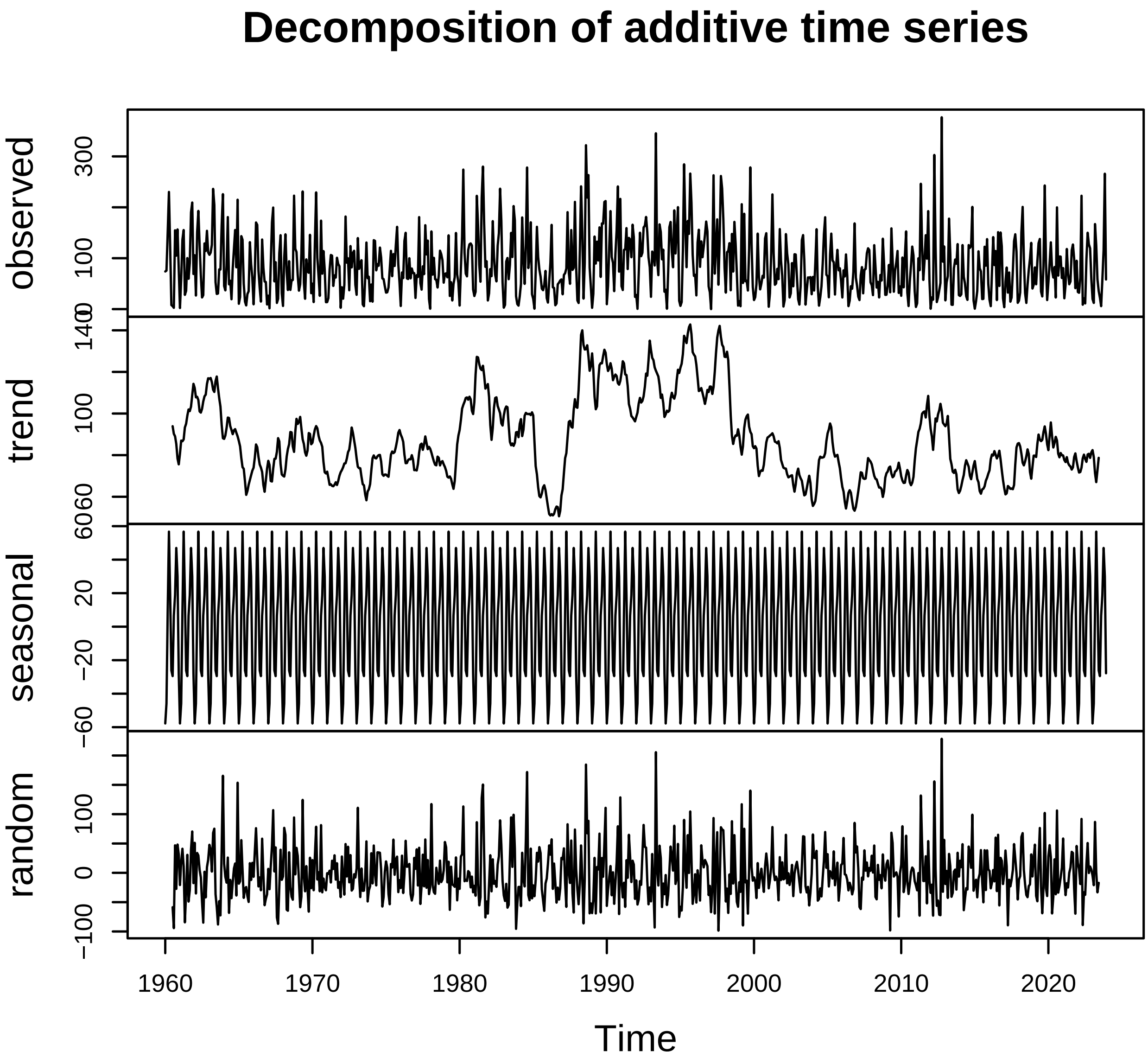


Figure 4.3: Decomposed time series plot for the monthly rainfall values.

4.3.3 Model identification and fitting

Using the *auto.arima* function in R, a $SARIMA(3,1,1)(1,0,0)[12]$ model was selected with mean of 87.53372, AIC of 7948.98 and BIC was 7976.55.

4.3.4 Model parameters and coefficients

Table 4.1 presented the estimated coefficients, showing all values within the typical range of -1 to 1, indicating model stability.

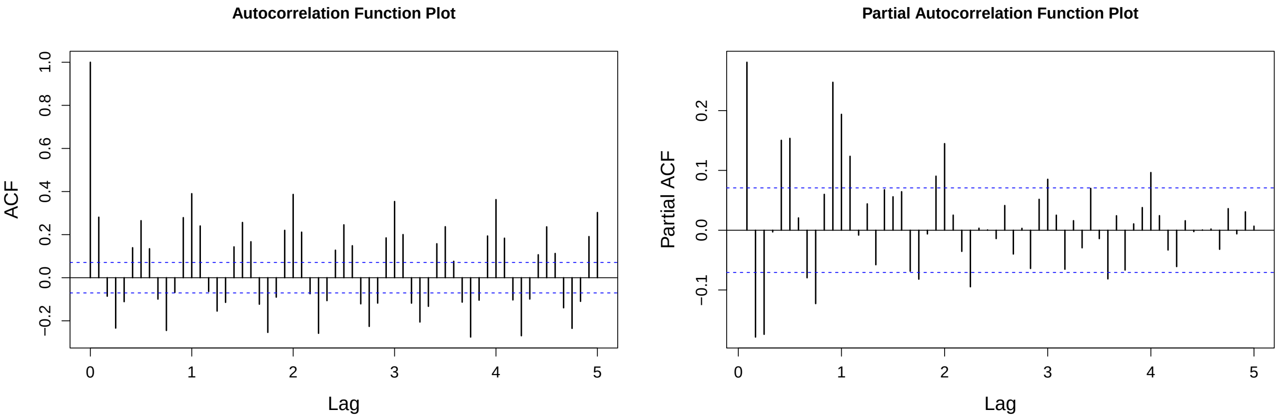


Figure 4.4: ACF and PACF plots for the monthly rainfall data.

Coefficients	ar1	ar2	ar3	ma1	sar1
Values	0.1754	-0.0982	-0.1689	-0.9709	0.2436
s.e	0.0395	0.0384	0.0395	0.0115	0.0428

Table 4.1: Coefficients and standard errors of the model

4.3.5 Fitted model

From Table 4.1 and Equation 2.1, the fitted model was:

$$(1-0.1754B^1+0.0982B^2+0.1689B^3)(1-0.2436B^{12})(1-B)Y_t = 87.53372+(1-0.9709B)\varepsilon_t \quad (4.1)$$

4.3.6 Model accuracy

To determine how close the predicted values match the actual observed values, the metrics; ME, RMSE, MAE, and MASE were computed. Table 4.2 shows the metrics computed to assess forecast accuracy.

Error measure	ME	RMSE	MAE	MASE	ACF1
Error values	-0.5610516	54.972	42.59623	0.8376039	-0.01006686

Table 4.2: Model accuracy metrics

4.4 Objective (iii): Forecasting monthly rainfall

The third objective involved forecasting monthly rainfall using the developed model. However, before forecasting was done, I first evaluated the model by checking residual diagnostics (e.g. ACF, normality, and independence), and assess its accuracy using error metrics as presented in the next sections of this chapter.

4.4.1 Forecast visualization

To validate the model, the testing dataset was used. The predicted values were compared with the actual values of the testing dataset. Figure 4.5 was used to display actual, fitted, and forecasted values with confidence intervals. The predicted values closely followed

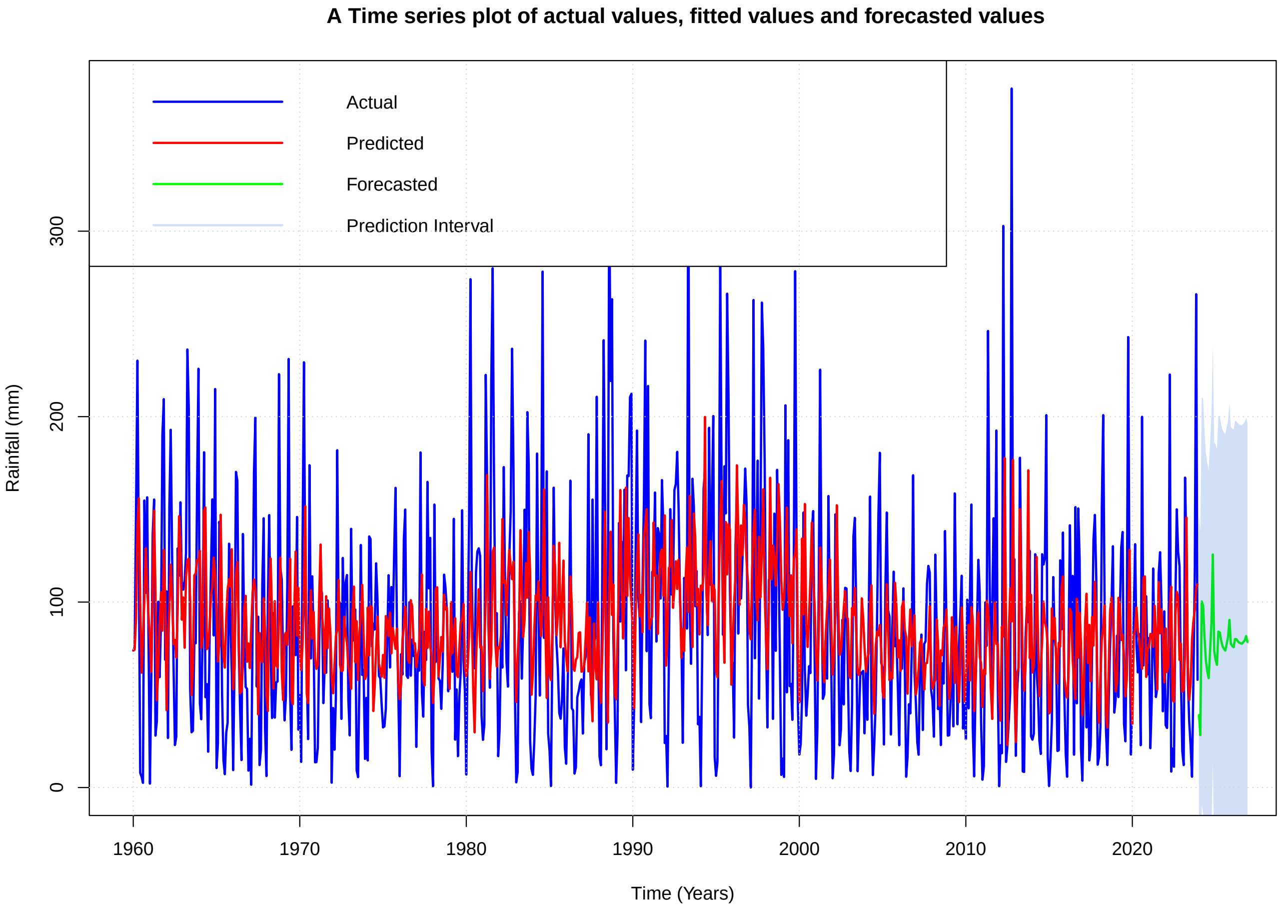


Figure 4.5: Time series plot of actual, fitted and predicted values

actual values, indicating good model performance.

4.4.2 Residual analysis

ACF and PACF plots of residuals

The acf plot of the residuals is shown in Figure (4.6). It displayed the autocorrelations of the residuals at different lags. The $y-axis$ represented the autocorrelation coefficients, and the $x-axis$ represented the lags (the time difference). The acf plot helps to check the adequacy of the model, whether the residuals are uncorrelated.

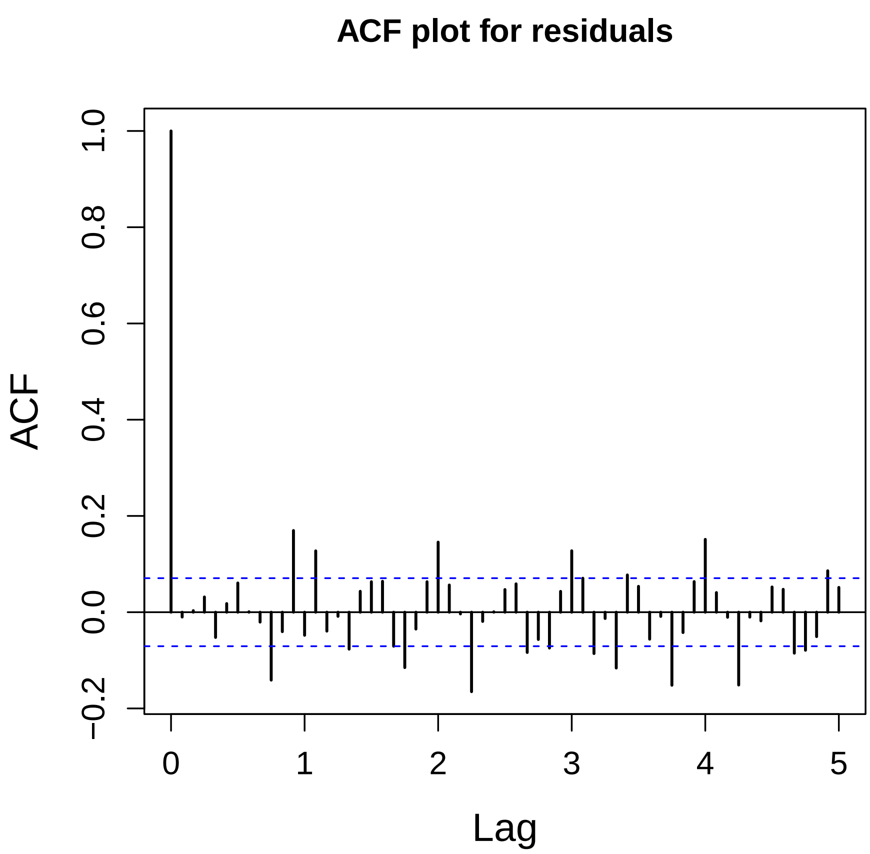


Figure 4.6: ACF plot of Residuals

Normality test of the residuals

To check for the normality of the residuals, Figure (4.7) was used. It represented the histogram plot of the residuals.

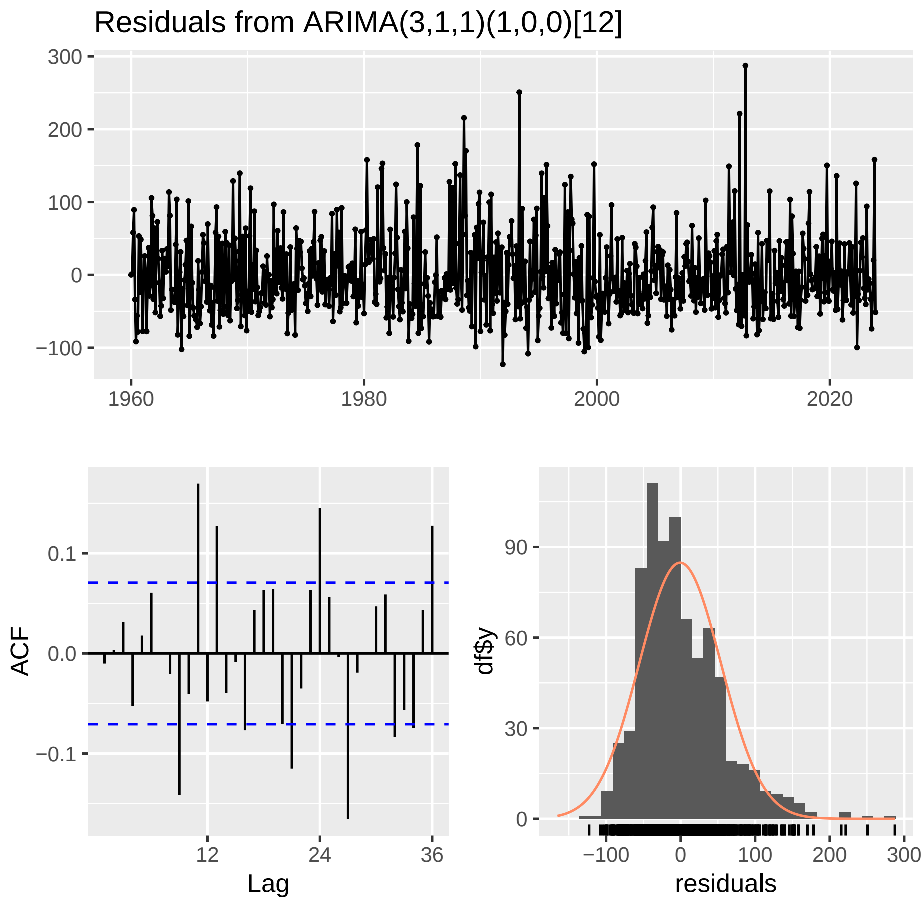


Figure 4.7: Residuals plot from SARIMA (3, 1, 1)(1, 0, 0)[12]

Q - Q plot of the residuals

Figure (4.8) shows the Q - Q plot of the residuals, which compared the quantiles of the residuals to the quantiles of a normal distribution. The residuals are normally distributed, if the points on this plot lie approximately along a straight line.

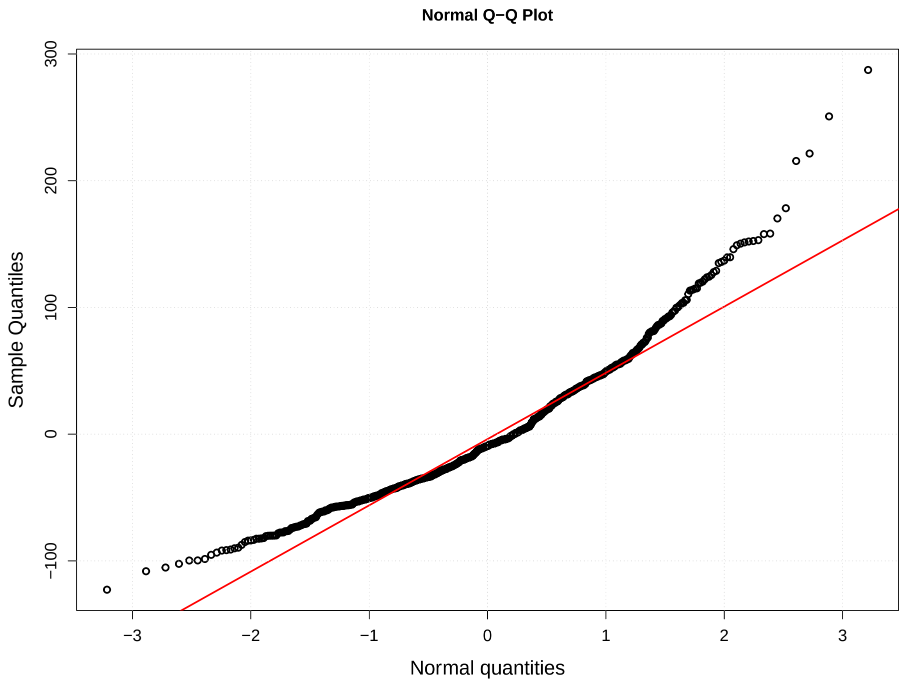


Figure 4.8: Q -Q plot of residuals

4.4.3 Forecasting

The model was used to forecast values for the next 3 years. The forecasted values (in green line) are presented in Figure 4.5. The 95% confidence interval is overlaid on the predicted values as shown in the grey region. The model was also used to predict values for the period from 1960 to 2023 and compare with the actual values. The predicted values are shown in red line and the actual values are shown in the blue line.

4.5 Discussion of Results

This section discussed how the results relate to the rainfall dynamics in Kasese district and the objectives of the study. The decomposition revealed strong seasonal patterns and trends. The SARIMA(3,1,1)(1,0,0)[12] model captured these dynamics well, achieving reasonable forecast accuracy. Residual diagnostics confirmed model adequacy, supporting the model's suitability for practical forecasting of monthly rainfall. The following

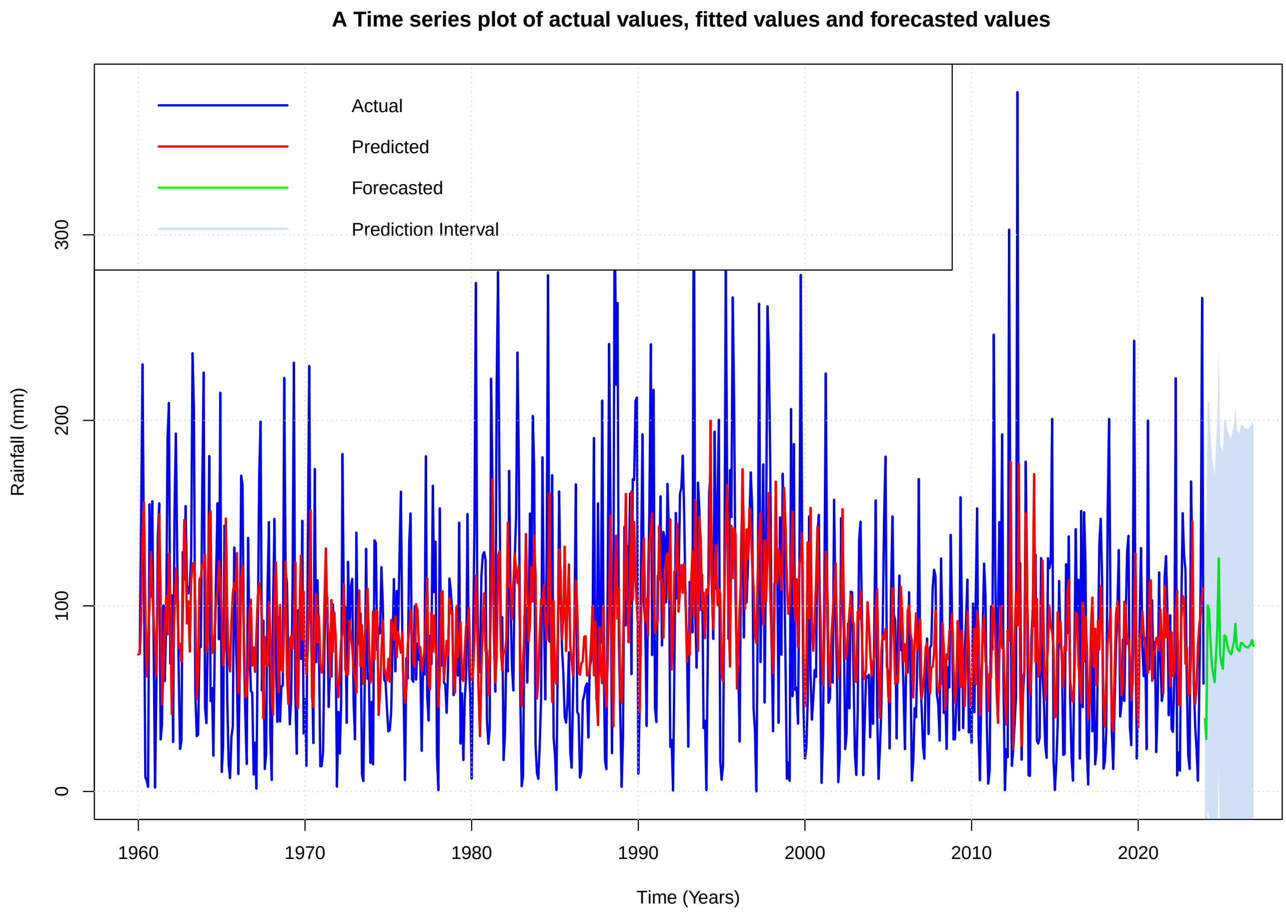


Figure 4.9: A Time series plot of the actual, fitted and predicted values

subsection further discuss the findings of the study as per specific objective:

4.5.1 Decomposition of data

From Figure (4.1), the up and down regular patterns suggested presence of seasonal variations.

Box plot

In Figure (4.2), each box represented the interquartile range (IQR) of rainfall for that particular month, with the line inside the box representing the median rainfall for that month over the period. The lines extending from the boxes are called whiskers, which represent the range of the data within 1.5 times the IQR from the lower (Q1) and upper (Q3) quartiles. The points outside the whiskers were considered outliers representing data points that are significantly different from the rest of the dataset values for each month. The box plot revealed the rainfall patterns, for example, the months of January and February have relatively a lower median rainfall and less rainfall variability, indicating that these months typically experience less rainfall. They are generally drier as compared to other months. However, some Januaries might have experienced unusual rainfall amounts slightly higher than the typical values for the month. The months of March to May and September to November have higher median rainfall and greater variability, indicating that these months generally experience more rainfall. This means that these months are generally wetter compared to others. Specifically, the mean monthly rainfall during these wettest seasons was approximately 118.03 mm, with a 95% confidence interval ranging from 111.32 mm to 124.74 mm. In contrast, the driest months—namely January, February, June, and July—had a mean monthly rainfall of about 47.53 mm, with a 95% confidence interval between 43.02 mm and 52.04 mm. These figures highlight the significant seasonal variation in rainfall amounts experienced in the district.

The taller boxes and whiskers suggested more variability of rainfall and the presence of outliers indicate extreme rainfall events especially in October with over 350mm of rainfall in 2013. By examining the medians across the months, guided the study in identifying the months which generally experience more or less rainfall. This helped in understanding seasonal patterns, which in turn would aid in proper planning for water resource management, agriculture, and other activities that depend on weather.

The trend component in Figure (4.3) gave the long-term changes in the rainfall over the years. For example: between the 1960s and 1970s, the trend was relatively stable with minor fluctuations in rainfall. The rainfall levels during this period did not change drastically. Between the mid- 1970s and mid- 1980s, there was a noticeable increase in the trend (an upward trend), indicating a period of increasing rainfall over these years. In the late 1980s and early 1990s, the trend reached a peak point and then showed a sharp decline. This suggested a significant decrease in rainfall during this period. This decline might have been due to climatic factors or changes during this period. After this sharp decline, between the mid-1980s and early 1990s, the trend increased. From the early 1990s and 2000s, the trend became more stable. This indicated a period of stable rainfall. From late 2000s upto 2023, the trend continued to show some fluctuations in rainfall, but there was a general decrease in the trend. This indicated a gradual decrease in rainfall levels in the recent years. The seasonal component indicated that the rainfall data had a consistent and repeating pattern for each year, which indicated a predictable annual cycle in rainfall. This component captures the seasonal fluctuations such as the dry and wet seasons that occur every year. The random component represented the residuals left after the trend and seasonal components were removed. It showed the irregular and random variations in rainfall that could not be attributed to the trend or seasonality.

4.5.2 Optimal solution of the model parameters

The Acf and Pacf plots

The acf plot revealed a sinusoidal pattern, which suggested presence of seasonality in the data. Thus, this was a clear sign that a seasonal $ARIMA$ model could be appropriate for the analysis. Since the acf pattern was sinusoidal, it was less useful in identifying the order of the MA component in the data. Similarly, the $pacf$ displayed many significant spikes, whereby the spikes showed a slow decay (the spikes did not cut off after a few lags). The presence of these significant spikes implied that the data had seasonal patterns, trends and cycles. To handle this effectively, we couldn't use the *AR* model, rather, a more advanced model like *SARIMA* was a good option to use in order to capture these patterns effectively.

Stationarity test

Since this p-value was less than 0.05 (at 5% level of significance), then it indicated that the data was stationary. However, the data portrayed some trends and seasons. Thus, to handle the seasonality and any non - stationarity in the data, a *SARIMA* model was automatically fit to capture potential issues the ADF test didn't fully capture.

Model identification and fitting

The non - seasonal part $(3,1,1)$, has an *AR* part of order 3, which means the model uses the past 3 values of the series to predict the current value. The differencing part is of order 1, meaning that the data is differenced once to remove any trends , thus making it stationary. The *MA* part has order of 1, implying that the model uses the past 1 error to correct the predictions. The seasonal part $(1,0,0)[12]$, has seasonal *AR* part of order 1, so the model considers the last 12 observations(which makes up one season) to predict the current value. There is no seasonal differencing applied, so $D = 0$, and also no seasonal *MA*. The model used a seasonal pattern which repeats every 12 periods i.e it was an annual seasonality based on monthly data.

Model coefficients

Firstly, all the coefficients are between the typical range of -1 to 1, which indicates that the model is likely stable and stationary, meaning it produces reliable and consistent forecasts without diverging or showing extreme variability.

4.5.3 Forecasting

Model accuracy and validation

The ME of -0.5610516 was slightly close to zero, meaning that, the model's predictions are, on average, fairly close to the actual values, with errors balancing out after some time. The negative means the model tends to predict lower rainfall than observed rainfall values. Thus, the model's predictions are quite close to the actual values, with only a minor tendency to predict lower rainfall than observed. The model gives lower predictions because it smooths out random fluctuations and outliers, focusing more on overall trends and patterns.

MAE was preferred than $RMSE$ to determine model accuracy, because $RMSE$ is more sensitive to outliers. It squares the errors before averaging them, giving more weight to large deviations. MAE on the other hand, treats all errors equally by averaging their absolute values, making it more robust to outliers. Therefore, MAE was preferred.

The MAE of 42.59623, means that, on average, the average monthly rainfall predictions deviate from the actual values by 42.46 mm which represents approximately 36% of the total variability of 116.89 in monthly rainfall, indicating that the model performs reasonably well in capturing rainfall the patterns.

Since $MASE$ of 0.8376039 was less than one, then the model's performance was better than a simple naive forecast. This indicated that the model's predictions were more accurate compared to simply using the previous period's value as the forecast.

The $ACF1$ value of -0.01006686, is very close to zero. This suggested that the model's errors were not autocorrelated, meaning the model captured most of the patterns in the data, which is good for accuracy.

For model validation, from Figure (4.5), the predicted values (in red line) showed a close relationship with the actual values (blue lines). The general trends of the forecasted and actual values appear to follow similar patterns, indicating that the model is capturing the broad movement of the rainfall data e.g., periods of rise and fall in rainfall. The forecasted line tends to flatten after some time because *SARIMA* models rely heavily on historical data to capture overall trends and patterns. So as the forecast horizon increases, the model runs out of recent data to base its predictions on, so it smooths out noise and random fluctuations, reverting to average values from the past, which causes the forecast to flatten.

From Figure (4.5), the $SARIMA(3,1,1)(1,0,0)[12]$ model fits well to the test data. The predicted values (in red line) tend to be close to the actual values (blue line).i.e there's a close relationship between them. Additionally, forecasted values lied within the 95% confidence interval.

Acf and Pacf plots of residuals

The acf plot (Figure 4.6) of the residuals shows a sharp decline after the first lag at zero. The sharp decline in the spikes after $lag0$ suggests that, beyond the immediate lag ($lag1$), the residuals do not exhibit significant autocorrelation. Though they might be some short-term correlation (as seen at lag 1), the residuals do not exhibit a strong, persistent autocorrelation at longer lags.

Normality test of the residuals

From Figure 4.7 for the histogram of the residuals, the Figure was bell shaped and symmetrical about the mean (zero), which is a good indicator of the residuals being normally distributed. This justified the proposed model.

Q - Q plot of the residuals

Figure (4.8) showed that the residuals were normally distributed because the points on the plot were approximately along a straight line.

Chapter 5: Summary, Conclusion and Recommendation

5.1 Introduction

This chapter presents the conclusions drawn from the research findings presented in chapter four, make recommendations and provide suggestions that can be considered for future research. It aims to breakdown the main findings from the study and offer guidance on how to address the research problem and inform relevant stakeholders.

5.2 Summary

This study developed a time series forecasting model for monthly rainfall in Kasese District, Uganda, using the SARIMA approach. Historical monthly rainfall data from 1960 to 2023 was analyzed through a structured methodology aligned with three specific objectives, as summarized below:

(i) **Decomposing the time series data into trend, seasonality, and residual components:** The data was first decomposed to reveal underlying structures. This decomposition highlighted clear seasonal patterns, particularly during March to May and September to November, which guided the subsequent modeling process.

(ii) **Determining the optimal solution of the model parameters for the rainfall forecasting model:** After performing stationarity tests and identifying autocorrelation patterns, the optimal SARIMA model was selected as SARIMA(3,1,1)(1,0,0)[12]. The choice was based on AIC/BIC values and diagnostic checks, confirming the model's adequacy and robustness.

(iii) **Forecasting monthly rainfall using the developed model:** The selected SARIMA model was used to generate forecasts, which showed good predictive accuracy. The model's reliability suggests its utility for decision-making in agricultural planning, disaster preparedness, and water resource management in Kasese District.

5.3 Conclusions

The region experience more rainfall in the months of March - April - May (MAM) and September - October - November (SON) and less rainfall in the months of January, February, June and July. The optimal parameters of the model were: $p = 3, d = 1, q = 1; P = 1, D = 0, Q = 0, s = 12$ and a $SARIMA(3,1,1)(1,0,0)[12]$ model with an AIC of 7948.98 was fit to the data and was identified as a good model for predicting the average monthly rainfall for the region. The model was used to forecast values for the next 3 years. These forecasted values lied within the 95% confidence interval as shown earlier. Further, it was observed that the $SARIMA(3,1,1)(1,0,0)[12]$ model fitted well to the test data. The predicted values (in red line) tend to be close to the actual values (blue line).i.e there's a close relationship between them. Thus, the model predicts well. The model can therefore be adopted by the UNMA and other relevant authorities like disaster preparedness, as a good tool to use in rainfall prediction and be in position to inform the public on expected rainfall amounts.

5.4 Recommendations

Based on the study conducted to develop a time series forecasting model for rainfall in the Kasese region, Uganda, the following recommendations were made in line with the study's specific objectives. First, given the clear seasonal and trend patterns revealed during the decomposition of the time series data, it is recommended that local authorities and planners incorporate these seasonal insights into their planning calendars for agriculture and water resource management. Second, the SARIMA(3,1,1)(1,0,0)[12] model, which was identified as the optimal model based on statistical performance and diagnostics, should be adopted for short-term rainfall forecasting to support timely and informed agricultural planning and disaster preparedness. Third, to ensure effective use of the model forecasts, it was recommended that workshops and training sessions be organized for local stakeholders—including farmers, community leaders, and policy-makers—to build their capacity in understanding, interpreting, and applying the model outputs in their decision-making processes.

5.5 Suggestions

While the SARIMA (3, 1, 1)(1, 0, 0)[12] model demonstrated strong performance in short-term forecasting, it exhibited limitations in long-term predictions due to evolving weather patterns and the assumption of stationarity. To address this, future research should explore alternative models, such as *SARIMAX*, which can incorporate exogenous variables, and hybrid models like *SARIMA* combined with machine learning techniques such as Artificial Neural Networks (*ANN*). These models may offer improved accuracy over extended time frames. Additionally, since climate conditions and rainfall patterns are subject to change any time, periodic updates with new data are also recommended, as they help maintain the models' accuracy, ensuring its continued relevance, as climatic conditions change. These steps are crucial for improving long-term forecasting capabilities and ensuring more robust predictions for the region.

References

Agyekum, T. P., Antwi-Agyei, P., & Dougill, A. J. (2022). The contribution of weather forecast information to agriculture, water, and energy sectors in East and West Africa: A systematic review. Frontiers in Environmental Science, 10, 935696.

Akaike, H. (1987). Factor analysis and AIC. *Psychometrika,* 52, 317-332.

Ali, A. A., Shire, A. S., Ayanle, A. T., Salaheldin, M., Aliyu, M. M., Atiku, F. A., & Abdulquyyoom, T. L. (2021). A Multi-Regression model based on monthly rainfall prognostication: Case study of Kasese district, in *East Africa.*

Alhassoun, R. (2011). Studies on factors affecting the infiltration capacity of agricultural soils. *Dissertationen aus dem Julius Kühn-Institut.*

Aliyu, A. S., Auwal, A. M., & Adenomon, M. O. (2021). Application of SARIMA models in modelling and forecasting monthly rainfall in Nigeria. Asian Journal of Probability and Statistics, 13(3), 30-43.

Biao, E. I., & Alamou, E. A. (2018). Stochastic modelling of daily rainfall for decision making in water management in Benin *(West Africa).*

Box, G. E., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). *Time series analysis: forecasting and control*. John Wiley & Sons.

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika, 52(3), 345-370.*

Brassington, G. (2017, April). Mean absolute error and root mean square error: which is the better metric for assessing model performance?. *In EGU General Assembly Conference Abstracts* (p. 3574).

Chatfield, C. (2000). *Time-Series Forecasting*. CRC Press.

Chawsheen, A. H., & Subhi Latif, I. (2006). Detection and treatment of outliers in data sets. *Iraqi journal of statistical sciences*, 6(1), 58-74.

Chattopadhyay, S., & Chattopadhyay, G. (2010). Univariate modelling of summer-monsoon rainfall time series: comparison between ARIMA and ARNN. Comptes Rendus Geoscience, 342(2), 100-107.

Chonge, M., Nyongesa, K., Mulati, O., Makokha, L., & Tireito, F. (2015). A time series model of rainfall pattern of Uasin Gishu County. *IOSR J Math,* 11(5), 77-84.

Dickey, D. A., & Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American statistical association,* 74(366a), 427-431.

Gnanasankaran, N., & Ramaraj, E. (2020). A multiple linear regression model to predict rainfall using Indian meteorological data. International journal of advanced science and technology, 29(8), 746-758.

Grevazi, V. M. (2024). Prediction of Rainfall by using Seasonal ARIMA Model in Namtumbo District, Tanzania. Journal of Statistics and Mathematical Engineering, 10(2), 38-44.

Kennedy, J., Blunden, J., Alvar-Beltrán, J., & Kappelle, M. (2021). State of the global climate 2020. World Meteorological Organization (WMO) Geneva, Switzerland.

Khan, F., Ali, S., & Pilz, J. (2018). Evaluation of statistical downscaling models using pattern and dependence structure in the monsoon-dominated region of Pakistan. Weather, 73(6), 193-203.

Kigobe, M., McIntyre, N., Wheater, H., & Chandler, R. (2011). Multi-site stochastic modelling of daily rainfall in Uganda. *Hydrological sciences journal, 56(1), 17-33*.

Kirchgssner, G., Wolters, J., & Hassler, U. (2012). *Introduction to modern time series analysis*. Springer Science & Business Media.

Liu, Z., Yang, Q., Shao, J., Wang, G., Liu, H., Tang, X., & Bai, L. (2022). Improving daily precipitation estimation in the data scarce area by merging rain gauge and TRMM data with a transfer learning framework. *Journal of Hydrology*, 613, 128455.

Mekouar, M. A. (2018). Food and agriculture organization of the united nations (FAO). Yearbook of International Environmental Law, 29, 448-468.

Moahmed Hassan, H., & Mahgoub Mohamed, T. (2014). Rainfall Drought Simulating Using Stochastic SARIMA Models for Gadaref Region, Sudan.

Momani, P. N. M. (2009). Time series analysis model for rainfall data in Jordan: case study for using time series analysis. *American Journal of Environmental Sciences,* 5(5), 599-604.

Montgomery, D. C., Jennings, C. L., & Kulahci, M. (2015). *Introduction to time series analysis and forecasting*. John Wiley & Sons.

Mugume, I., Waiswa, D., Mesquita, M. D. S., Reuder, J., Basalirwa, C., Bamutaze, Y., & Ayesiga, G. (2017). Assessing the performance of WRF model in simulating rainfall over western Uganda. *Journal of Climatology & Weather Forecasting,* 5(01).

Mushtaq, R. (2011). Augmented dickey fuller test.

Navid, M. A. I., & Niloy, N. H. (2018). Multiple linear regressions for predicting rainfall for Bangladesh. Communications, 6(1), 1-4.

Neath, A. A., & Cavanaugh, J. E. (2012). The Bayesian information criterion: background, derivation, and applications. *Wiley Interdisciplinary Reviews: Computational Statistics,* 4(2), 199-203.

Nirmala, M., & Sundaram, S. M. (2010). A seasonal ARIMA model for forecasting monthly rainfall in Tamilnadu. *National Journal on Advances in Building Sciences and Mechanics*, 1(2), 43-47.

Nsubuga, F. W., Botai, O. J., Olwoch, J. M., Rautenbach, C. D., Bevis, Y., & Adetunji, A. O. (2014). The nature of rainfall in the main drainage sub-basins of Uganda. *Hydrological Sciences Journal*, 59(2), 278-299.

Nsubuga, F. W., & Rautenbach, H. (2018). Climate change and variability: a review of what is known and ought to be known for Uganda. *International Journal of Climate Change Strategies and Management,* 10(5), 752-771.

Opio, R., Sabiiti, G., Nimusiima, A., Mugume, I., & Sansa-Otim, J. (2020). WRF Simulations of extreme rainfall over Uganda's Lake Victoria Basin: Sensitivity to parameterization, model resolution and domain size. *Journal of Geoscience and Environment Protection, 8(4), 18-31.*

Oriangi, G., Mukwaya, P. I., Luwa, J. K., Emmanuel, M., Maxwell, M. G., & Bamutaze, Y. (2024). Variability and Changes in Climate in Northern Uganda. *African Journal of Climate Change and Resource Sustainability*, 3(1), 81-97.

Petris, G., Petrone, S., & Campagnoli, P. (2009). *Dynamic linear models with R*. Springer Science & Business Media.

Pinsky, M., & Karlin, S. (2010). An introduction to stochastic modeling. *Academic press.*

Powers, J. G., Klemp, J. B., Skamarock, W. C., Davis, C. A., Dudhia, J., Gill, D. O., & Duda, M. G. (2017). The weather research and forecasting model: Overview, system efforts, and future directions. *Bulletin of the American Meteorological Society,* 98(8), 1717-1737.

Rankoana, S. A. (2020). Food security under unreliable rainfall: the case study of a rural community in Limpopo Province, South Africa. *Journal of Water and Climate Change*, 11(3), 677-684.

Ronald, O. (2019). Examining observed and simulated extreme rainfall events over

Lake Victoria Basin in Uganda *(Doctoral dissertation, Makerere University).*

Shumway, R. H., & Stoffer, D. S. (2017). *Time Series Analysis and Its Applications: With R Examples*. Springer.

Tibara, Y., Wasswa, H., & Semakula, H. M. (2022). Analysing drivers of community vulnerability to flood hazards in Kasese Municipality, *Uganda.*

Tuyizere, C., Mung'atu, J. K., & Ndanguza, D. (2017). Rainfall Forecasting in Gasabo District Using Markov Chain Properties. *International Journal of Scientific Engineering and Technology,* 6(4), 128-131.

UBOS. (2016). The national population and housing census 2014-main report. *Uganda Bureau of statistics Kampala.*

Wambura, F. J. (2020). Potential of rainfall data hybridization in a data-scarce region. *Scientific African*, 8, e00449.

Wangome, B. M. (2022). A Rainfall prediction model using long short-term neural networks for improved crop productivity: a case of maize planting in Machakos County *(Doctoral dissertation, Strathmore University).*

West, M., & Harrison, J. (2006). *Bayesian forecasting and dynamic models.* Springer Science & Business Media.

World Bank. (2021). The Climate Change Knowledge Portal: Precipitation data. https://climateknowledgeportal.worldbank.org/

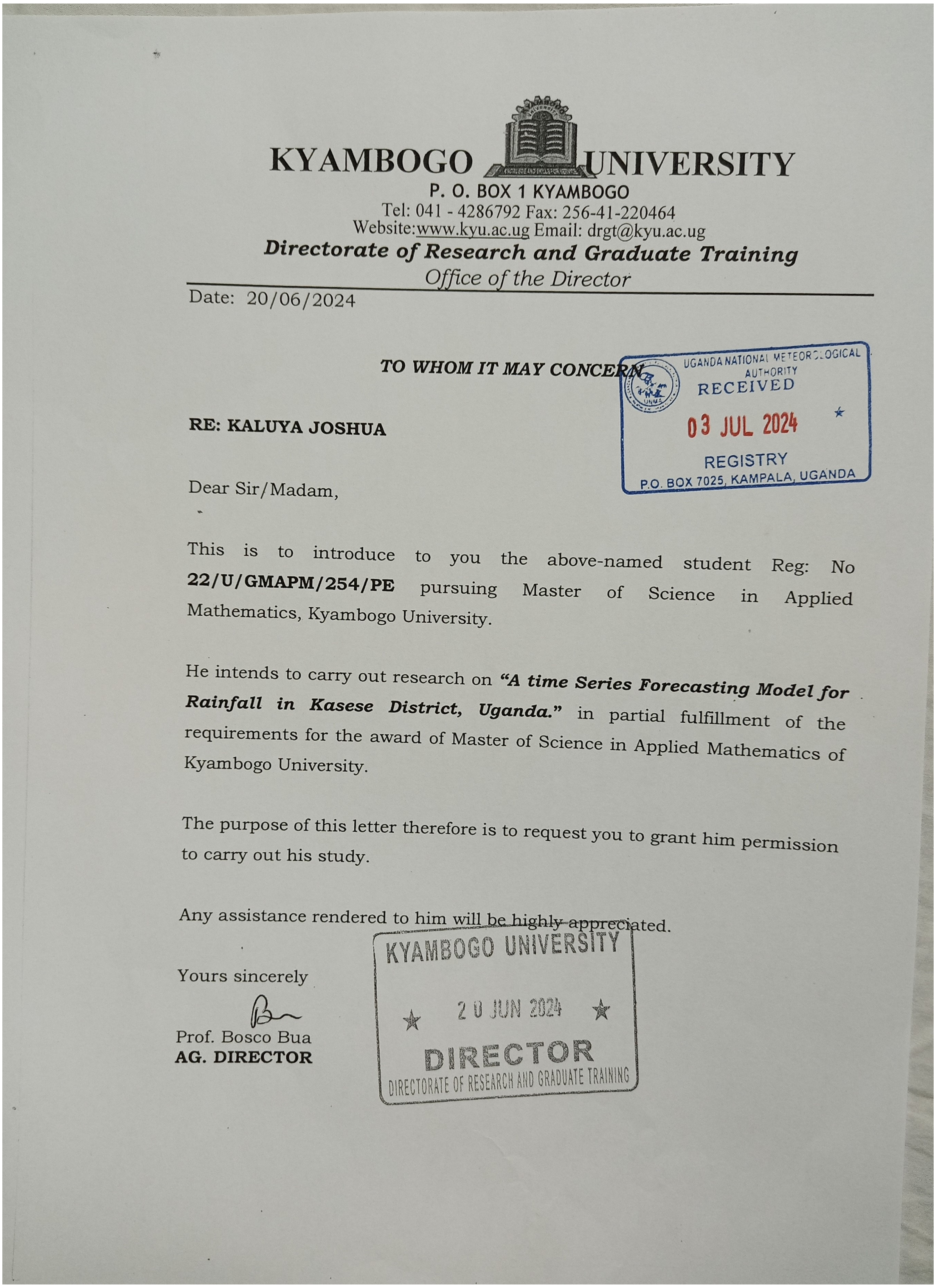
World Bank Group. (2020). Uganda: Climate Risk Country Profile (1901–2016 climatology).

Yeditha, P. K., Kasi, V., Rathinasamy, M., & Agarwal, A. (2020). Forecasting of extreme flood events using different satellite precipitation products and wavelet-based machine learning methods. *Chaos: An Interdisciplinary Journal of Nonlinear*

Science, 30(6).

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

Appendix



KYAMBOGO UNIVERSITY
P. O. BOX 1 KYAMBOGO
Tel: 041 - 4286792 Fax: 256-41-220464
Website:www.kyu.ac.ug Email: drgt@kyu.ac.ug
Directorate of Research and Graduate Training
Office of the Director

Date: 20/06/2024

TO WHOM IT MAY CONCERN

UGANDA NATIONAL METEOROLOGICAL AUTHORITY
RECEIVED
03 JUL 2024
REGISTRY
P.O. BOX 7025, KAMPALA, UGANDA

RE: KALUYA JOSHUA

Dear Sir/Madam,

This is to introduce to you the above-named student Reg: No **22/U/GMAPM/254/PE** pursuing Master of Science in Applied Mathematics, Kyambogo University.

He intends to carry out research on ***"A time Series Forecasting Model for Rainfall in Kasese District, Uganda."*** in partial fulfillment of the requirements for the award of Master of Science in Applied Mathematics of Kyambogo University.

The purpose of this letter therefore is to request you to grant him permission to carry out his study.

Any assistance rendered to him will be highly appreciated.

Yours sincerely

Prof. Bosco Bua
AG. DIRECTOR

KYAMBOGO UNIVERSITY
20 JUN 2024
DIRECTOR
DIRECTORATE OF RESEARCH AND GRADUATE TRAINING

Figure 5.1: Introduction letter