ARIMA AND VECTOR AUTOREGRESSIVE MODEL EVALUATION IN FORECASTING RAINFALL: A CASE OF KISUMU

BY

MAWORA THOMAS MWAKUDISA

A THESIS SUBMITTED IN FULFILMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY IN APPLIED STATISTICS

SCHOOL OF MATHEMATICS, STATISTICS AND ACTUARIAL SCIENCE

Maseno University

Declaration

This thesis is my own work and has not been presented for a degree award in any other institution.

MAWORA THOMAS MWAKUDISA

This thesis has been submitted for examination with our approval as the university supervisors.

DR. EDGAR OUKO OTUMBA

DR. JOYCE AKINYI OTIENO

Maseno University

2022

Acknowledgments

I would like to first thank my two supervisors, Drs. Edgar Otumba and Joyce Otieno for their tireless efforts in pushing me to get this thesis done. Prof. Roger Stern and Dr. David Stern were always very helpful to listen to my ideas, critique and give suggestions. Thank you my colleagues James Puti and Daniel Aoyi who helped me with getting the right LaTeX format for writng the thesis. I cannot forget my family for their moral support. They include my parents Pamphil and Virdiana Mwasheghwa, Henry and Teresia Itumbe and my lovely wife Judith Itumbe and children Judah and David Mawora. Finally, I thank the Lord Jesus Christ for this gift.

Dedication

To my wife and sons.

Abstract

Time Series Analysis has been used over the decades in data analysis and forecasting. Auto Regressive Integrated Moving Average (ARIMA) models have been fit on economic data and engineering data. The models have also been used in analysis of climate data. Previous studies have focussed on temperature data from National Meteorological Stations where summarized monthly values were used. In this study, we used daily rainfall data from Kenya Meteorological Services Station in Kisumu. The objectives included univariate time series modelling using ARIMA on long term rainfall data for daily, monthly, seasonal and annual data and forecasting rainfall for the different time periods. The other objective was to compare forecast from univariate ARIMA to Vector Autoregression (VAR) when rainfall, minimum and maximum temperature values are included in model. ARIMA models were fit on the KMS rainfall data, and VAR models were fit on temperature, minimum and maximum rainfall data from KMS. Finally, farmers' local rainfall data was compared to that of KMS for independence. Results showed that forecasts under VAR did not give a more precise forecast of future rainfall than ARIMA. Further, that there was not enough statistically significant evidence to suggest that rainfall data from KMS and farmers' locale were independent.

Contents

Declaration ii
Acknowledgements iii
Dedication iv
Abstract v
Table of Contents vi
List of Tables viii
List of Figures ix
Index of Notations x
Chapter 1: Introduction 1

1.1 Introduction and Background 1
1.2 Statement of the Problem 6
1.3 Objective of the Study 7
1.3.1 Main Objective 7
1.3.2 Specific Objectives 7
1.4 Significance of the Study 7
1.5 Overview of the chapters 8

Chapter 2: Literature Review 9

2.1 Introduction 9
2.2 Spatial Coverage of Climate Data 11
2.3 Analysis of Long-Term Rainfall Data 12
2.3.1 Modeling of Rainfall Data 12
2.3.2 Time Series Analysis of Rainfall Data 13
2.4 Forecasting Time Series Data 15
2.5 VAR(p) Models 16
2.6 Conclusion 17

Chapter 3: Methodology 18

3.1 Sampled Farmers 18
3.2 Analysis of Long-Term Rainfall 21
3.3 Uni-Variate Forecasting Models 21
3.3.1 ARIMA Model 21
3.3.2 Differencing in an ARIMA Model 22
3.3.3 ARIMA(0,1,0) Model 22

3.3.4 ARIMA(p,0,0) Model 23
3.3.5 ARIMA(0,0,q) Model 23
3.3.6 ARIMA(p,1,q) Model 23
3.3.7 Determining the Number of Differences 24
3.3.8 The SARIMA Model 25
3.3.9 Forecast Errors 26
3.4 Multi-Variate Forecasting Model 28
3.4.1 VAR(p) Model 28
3.4.2 Lag selection for the VAR models 28
3.4.3 Testing for Multi-Variate Stationarity 29
3.4.4 VECM Models for Farmers Data 30
3.5 An Overview of Analysis Applied on Different Datasets 31
Chapter 4: Results and Discussions **32**
4.1 Summary of the data 32
4.2 Forecasting Rainfall Data for KMS using ARIMA Models 33
4.2.1 Fitting ARIMA models to the rainfall data 34
4.2.2 ARIMA for the Daily and Annual Data 36
4.2.3 SARIMA for the Monthly and Seasonal Data 37
4.2.4 Forecasting Precipitation Data 39
4.3 Using VAR models with Rainfall, Maximum and Minimum Temperature Data to Forecast KMS Data 42
4.4 Comparing daily amount of rainfall between farmers location and Kisumu 50
4.5 Comparing the number of rainy days between farmers location and Kisumu 51
Chapter 5: Summary, Conclusion and Recommendetaions **53**
5.1 Summary 53
5.2 Conclusions 54
5.3 Recommendations 55
References **56**
Appendix **62**
A.1 Procedure for Sampling 62
A.2 Catalogue of Rainfall Recording Stations in Kisumu and Kericho 63

List of Tables

3.1 Farmer sampling according to the farmer groups 19

3.2 Analysis methods applied on available data 31

4.1 A summary of the long term rainfall in Kisumu 32

4.2 Fitting ARIMA model on KMS total annual rainfall data 37

4.3 Fitting ARIMA model on KMS daily rainfall data 37

4.4 Fitting ARIMA model on KMS total monthly rainfall data 38

4.5 Fitting ARIMA model on KMS total seasonal rainfall data 39

4.6 Forecast errors for forecasted daily, monthly and seasonal rainfall data for Kisumu 40

4.7 Forecasted daily rainfall for five days using KMS data 41

4.8 Forecasted monthly rainfall for five months using KMS data 42

4.9 Forecasted seasonal rainfall for five seasons using KMS data 42

4.10 VAR(5) model for predicting *rainfall* considering lagged rainfall, minimum and maximum temperatures 44

4.11 Measure of variability accounted for in VAR models for lags 1,2,3,4 and 5 46

4.12 VAR(5) model for predicting *maximum temperature* considering lagged rainfall, minimum and maximum temperatures 47

4.13 VAR(5) model for predicting *minimum temperature* considering lagged rainfall, minimum and maximum temperatures 48

4.14 Forecast results using VAR(5) model for daily rainfalland temperature data 49

4.15 Chi-Square test results for fifteen days comparing individual farmer's rainfall data to Kisumu 51

4.16 Chi-Square test results for fifteen days comparing number of rainy days between farmers and Kisumu 52

List of Figures

1.1 Geographic distribution of surface observing stations in Kenya as at 2021 (Source: Tageo website [44]) 4

3.1 Spatial distribution of farmers' farms in the Nyando region 20

4.1 Line graphs showing totals for daily, monthly, seasonal and annual rainfall for Kisumu (1961 - 2014) 35

4.2 ACF plots for the KMS Kisumu rainfall data 35

4.3 PACF plots for the KMS Kisumu rainfall data 36

Index of Notations

ARIMA Auto Regressive Integrated Moving Average 1
VAR Vector Autoregression 1
MTS Multi-Variate Time Series . . . 1
NMS National Meteorological Stations 1
WMO World Meteorological Organization 2
ENSO El Niño-Southern Oscillation . 2
IPCC Intergovernmental Panel for Climate Change 2
KMS Kenya Meteorological Services . 3
GCM General Circulation Models . . 5
CCAFS Climate Change Agriculture and Foods Security 5
TAHMO Trans African Hydro-Meteorological Observatory 5
WMO World Meteorological Organization 6

SDGs Sustainable Development Goals 9
KMS Kenya Meteorological Services . 10
CSA Climate Smart Agricultural . . 11
AR Autoregressive 13
MA Moving Average 13
USA United States of America 14
SST Sea Surface Temperatures 15
DSTs Decision Support Tools 15
FAO Food and Agriculture Organization 15
DSSAT Decision Support Tools for Agrotechnology Transfer 15
ARMA Autoregressive Moving Average 16
ANN Artificial Neural Networks . . . 16
SARIMA Seasonal Autoregressive Integrated Moving Average 16

ha Hectares 18
FOKODEP Friends of Katuk Odeyo Development Project 19
SAR Seasonal Auto-Regressive 25
SMA Seasonal Moving Average . . . 25
ME Mean Error 26
RMSE Root Mean Square Error . . . 26
MAE Mean Absolute Error 26
MAPE Mean Absolute Percentage Error 26
i.i.d Identical and Independently Distributed 28
AIC Akaike's Information Criterion . 28
HQ Hannan-Quinn Criterion 28
SC Schwarz Criterion 28
MLE Maximum Likelihood Estimator 28
FPE Final Prediction Error 29
VECM Vector Error Corrction Model 30
GIS Geographic Information System 31

BIC Bayesian Information Criterion . 35

VARMA Vector Auto Regressive Integrated Moving Average 55

Chapter 1

Introduction

1.1 Introduction and Background

Statistics has a symbiotic relationship with many disciplines since its methods are applied to help solve everyday challenges by offering a range of probable solutions. For instance, the Auto Regressive Integrated Moving Average (ARIMA) models have been used in forecasting one-variable data over time. One area where they have been used is in econometric analysis. Another time series model commonly used in econometrics is the Vector Autoregression (VAR). VAR is a Multi-Variate Time Series (MTS) Model which has additional variables alligned in equal time intervals with the variable of interest and are used to predict endogenously.

In weather and climate forecasting, numerical methods are used for both short and long term forecasts [54]. Weather forecasting is considered to be a "complex and challenging science", which depends on the efficient quantities of weather observations[54]. The data used include surface weather observation and upper air observation. The surface weather observations usually include measurements of atmospheric pressure, temperature, wind speed and direction, humidity and precipitation. The records are ideally measured at standard times across the world, by National Meteorological Stations (NMS). For better forecasts, a big network covering short distances is required. ARIMA models have been used to analyse and forecast weather elements, in most cases the monthly average temperature.

Dense climatic data is important in order to develop climate forecasting models. The

direction and strength of wind affects how far the rain stretches in a given area. The NMS and World Meteorological Organization (WMO) are able to obtain regional data hence produce forecasts for the globe and regions. This they do by including upper air observation like the El Niño-Southern Oscillation (ENSO) data which varies the atmospheric pressure hence predict the movement of wind across regions.

There are still gaps in the analysis of climatic data for various regions globally. The Intergovernmental Panel for Climate Change (IPCC) stated that "Uncertainty in modeling some modes of climate variability, and of the distribution of precipitation between heavy and light events, remains large. In many regions, projections of changes in mean precipitation also vary widely between models, even in the sign of the change. It is necessary to improve understanding of the sources of uncertainty"and again "Climate models remain limited by the spatial resolution and ensemble size that can be achieved with present computer resources, by the need to include some additional processes, and by large uncertainties in the modeling of certain feedbacks (e.g., from clouds and the carbon cycle)"[29].

One challenge when analyzing climatic data in Africa is the scarcity of both daily rainfall and temperature data. The presence of these gaps imply that the data is not dense enough to give the true local picture. For instance, Hulme *et al* [6] did not use local data in their study on African climate change. They used records of rainfall and temperature at regional level hence it would be difficult to translate their work to local users who may have a varying opinion. WMO has relied heavily on the data from NMS stations which is sparsely distributed, hence the forecasts are for large regions. It is not surprising that concerns are raised about how accurate the forecasting models used are on local context[3].

In his essay, Katz [31] identified the need to include modeling extreme events, specifically for the heat waves and the consideration of spatial dependencies. Rainfall is characterized by extreme occurrences from time-to-time and from point to point, unlike temperature. Rainfall is important since many small-scale farmers pay closer attention to it as opposed to temperature. Thus a lot of rainfall data recorded daily is required to further conduct such analyses.

In Kenya, the Kenya Meteorological Services (KMS) has been mandated to collect and store historical climatic data. The climatic data can be accessed from the main KMS through a written request to the Customer Service Office. In 2012, KMS had thirty surface climate observing stations countrywide (Figure 1.1), which was an average of one station per $19,377\ Km^2$. These stations collect thirteen climatic data including rainfall, air temperature, wind speed and direction, air pressure, soil temperature, solar radiation, sunshine duration, relative humidity, evaporation and cloud cover.

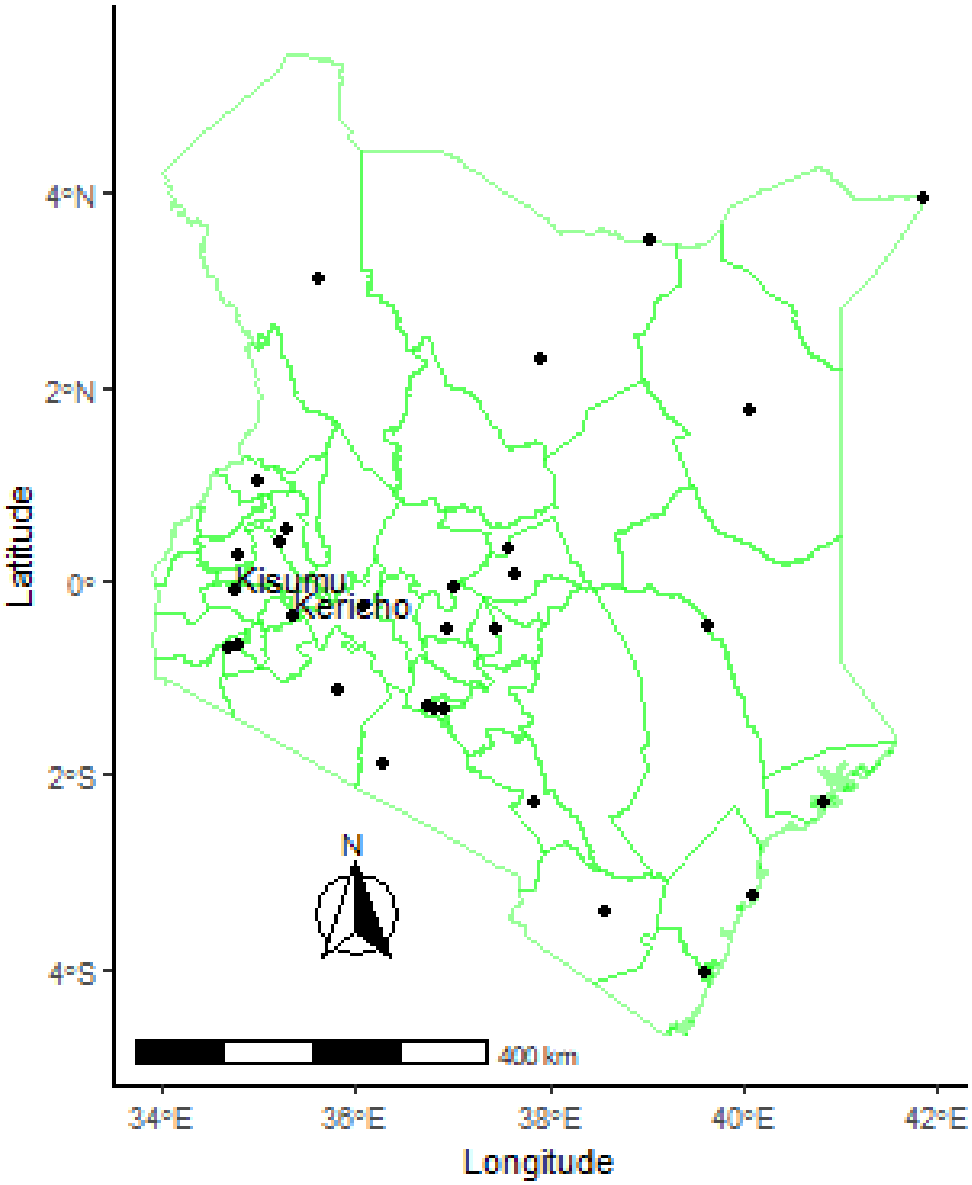

Figure 1.1: Geographic distribution of surface observing stations in Kenya as at 2021 (Source: Tageo website [44])

Thus, climatic data is sparsely distributed since one station serves an average of $19,377\ Km^2$. Because of this, micro-climatic differences are overlooked when forecasting weather for a given region in the country. One of the gaps identified in the working paper by Lindsey [40] and her team indicated the importance of having a better network of observations of climatic data. They identified a possible opportunity is working with the private sector. In addition, the General Circulation Models (GCM) used to estimate climate projections work at low resolutions, some of 100 Km^2, with better models representing 25-50 Km^2. There is unique opportunity to utilize farmers' daily collected rainfall data.

In November 2013, Climate Change Agriculture and Foods Security (CCAFS) empowered 100 farmers in Nyakach and Soin-Sigowett and provided them with rain-gauges to collect daily rainfall. This data is freely available to researchers. Furthermore, CCAFS partnered with Trans African Hydro-Meteorological Observatory (TAHMO) and installed two AWS in Nyakach and Soin that record hourly precipitation, minimum and maximum temperature, wind speed, solar radiation and atmospheric pressure. Challenges including inadequate technical capacity for operating and maintaining equipment resulted in gaps within the available data. Nyando was a good site which we used for this study given the availability of farmers recorded rainfall data, proximity to KMS Kisumu and Kericho and the presence of additional data from nearby volunteer stations.

Increasingly, there is need to use locally available data so as to solve local problems. The local problems may be hidden when one uses data from NMS alone which is sparse. In addition, when working on local contexts, the researchers are limited hence can't access upper air observation. Partnerships like the CCAFS can help expand the network of data that can be used for research and help inform members of the public. However, it is not feasible yet to find local data on temperature. Thus, there is a limitation on using rainfall data to understand and make informed decisions.

The rainfall data is time series data. Therefore, time series models such as ARIMA can be fit to it and used for forecasting. Most analysis so far use monthly summaries, however, short term forecasts including daily forecasts are important considerations when dealing with single variable. The use of ARIMA and VAR models can be adequately

explored and their effectiveness determined to enhance their utility for locally available data. It is noteworthy that, little has been done to fit VAR models on rainfall data. However, the VAR models, which use MTS data, can add value when applied to close range meteorological data. The data from different stations can be used endogenously hence help in the forecast for multiple points.

This study fits time series models for rainfall considering the spatial and temporal dependencies. The VAR model is fit on local farmer collected rainfall data from Nyando, Kenya, and used to forecast rainfall while considering the spatial dependencies between farmers' rainfall data. The models will be useful to predict local rainfall using spatially distributed local rainfall data.

1.2 Statement of the Problem

The World Meteorological Organization (WMO) use regression analysis to forecast regional climate. The local NMS thereafter downscale this using the analogous seasons, where they select a season that best resembles the expected forecast, and use recorded daily values to give the downscaled version. In addition, the forecasts provide information qualitatively using the scales like "above normal", "below normal" and "near normal". To produce the regional forecasts, WMO uses both aerial and surface values available at the NMS stations, hence there are concerns on how reliably the forecasts represent the local context. Increasingly, local volunteers and scientists are collecting climatic data yet they need to learn to interpret the quantitative rainfall values. In addition, there is need to apply statistical methods to help develop products that can help users understand and use the data they have collected. Autoregressive Integrated Moving Average (ARIMA) models can fit on long term univariate rainfall data and forecast for daily, monthly, seasonal and annual rainfall data. Although VAR models have potential to be used in short-term forecasting of MTS data, there is need to conduct a comparative analysis on how well they perform in comparison to univariate ARIMA models.

1.3 Objective of the Study

1.3.1 Main Objective

The main objective was to compare ARIMA and VAR models for forecasting rainfall data and to see how closely related the data from Kisumu KMS is to the farmers in Nyando rerion in Western Kenya.

1.3.2 Specific Objectives

The specific objectives were to:

1. Conduct univariate time-series analysis, modelling and forecasting for different time periods using rainfal data from Kenya Meteorological Serice Station (KMS) in Kisumu

2. Fit VAR models for Kisumu data using rainfall, minimum and maximum temperatures, and compare its forecast with that of the univariate forecast

3. Test how representative rainfall data from KMS is to the local Nyando region

1.4 Significance of the Study

In this study, there are three main beneficiaries. First, the research community who will be able to utilize and further critique the methods of using ARIMA and VAR models on rainfall data, and using it to forecast for short and long term periods. Secondly, this study is beneficial to the KMS to the extent that it provides a method of downscaling data by using short term local MTS rainfall data and get dependable short term forecasts. Finally, the study is important to users, including farmers, who can utilize the outputs of point short term forecasts to help them make local farm management decisions.

1.5 Overview of the chapters

This chapter gave a brief introduction to the study, cited the problem statement and gave the objectives. Chaper 2 will give a literature review of the works done that closely relate. Chapter 3 gives a brief on the methods applied in our work. Chapter 4 gives the results while discussing them. In the chapter, conduct long term univariate time series modelling and forecasting before using the VAR models. The last Chapter, 5, makes the concluding remarks and provide some recommendations.

Chapter 2

Literature Review

2.1 Introduction

Climate change has been an imprtant subject since towards end of the 20^{th} century. It is currently the thirteenth goal in the Sustainable Development Goals (SDGs). Climate consists of multiple elements, with most common being temperature and precipitation. In 2006, Hansen *et al* [56] found that the global surface tempreature had increased by ≈ 0.2°C per decade over a period of thirty years. The increase in temperature were projected to have higher impacts on the precipitation and increase the likelihoods of extreme rainfall. It has not been easy to find out the changes in rainfall globally due to several reasons. One of them is the lack of quality rainfall data.

Climatic data are useful to prepare climate users to mitigate the current variability and adopt for eventual climate change. This is driven by the fact that most Sub-Saharan Africa farmers practice rain fed agriculture. It is not clear if the change in rainfall will be positive (increase) or negative [41]. Cooper *et al* [41] proposed that academicians can use available data to study trends that will help farmers calculate risks involved with planting. In another study, 53 years of available climatic data was used to analyze risks that farmers face with planting different crops and varieties of crops [42]. However, an earlier study using daily data found that there are localized influences that affected the cessation of rainfall [43].

As a result of the scanty nature of historical climate data, data from satellite stations has been used to study past climate while GCM models have been used to simulate

projections for future climate trends. Hulme *et al* [6] studied available historical African climatic data and in their 2001 publication, they found that temperature had increased by a rate of 0.5°C per decade in the 20^{th} century. The GCM projected warming of between 0.2°C to more than 0.5°C per decade. Rainfall was found to be increasing in some areas of East Africa [6, 13]. The study projected drier months to have 5-10 % reduction in rainfall and the wetter months to have 5 - 20 % increase in precipitation. In addition, they projected increased occurrences of extreme weather conditions. This is not very suprising since East Africa is usually affected by the El Niño-Southern Oscillation (ENSO), even though it rarely experience extreme climate events. From historical data, September 1997 to March 1998 period recorded very high rainfall [6]. Other researchers have shown that there is too much variability in precipitation patterns hence no clear indication of increase or decrease in rainfall [10]. The idea of increased rainfall is further in contradiction to how farmers felt about their context.

The Kenya Meteorological Services (KMS) [35], is the main custodian of most weather information in Kenya. To be able to effectively monitor weather over the country, KMS has data collection stations which include among others thirty (30) synoptic surface weather observation stations, with more concentration in the wetter regions of the country. The synoptic weather stations collect daily data on rainfall, minimum and maximum temperatures, wind speed, wind direction, air pressure, soil temperature, solar radiation, sunshine duration, relative humidity, evaporation and cloud cover. There exists volunteer stations that collect data for personal use, hence increases spatial coverage. However, most volunteer stations collect rainfall data only.

Rainfall is an important weather element to farmers who are end users themselves of forecast information and any other products that result out of it. It is important to contextualize information to users since they eventaully make decisions using the output. For instance, this study we focused on the farmers from Kenya, and specifically in Nyakach and Soin-Sigowett, generally referred to here as Nyando.

2.2 Spatial Coverage of Climate Data

The scarcity of data presents a problem of spatial representation. Such data is important to aid in decision making for policy makers. Several tools have been developed to help researchers cover as much area as possible. One tool is the Precipitation Estimation from Remotely Sensed Information using Artificial Neural Networks-Climate Data Record (PERSIANN-CDR) that estimates daily rainfall for an area of 0.25°[27]. This is equivalent to 756.25 Km^2 at the equator. There also exists the Australian dataset for quality climate [20] and for North America [27] and a more recent German National Forest Inventory [33]. To assess quality, a spatial data can be used to look at questions like; Are there significant regional differences in accuracy among data sets? How accurate are their mean values compared with extremes? Does their accuracy depend on spatial resolution. Benkhe *et al* in 2016 considered downscaled datasets from different stations and found considerable differences for the rainfall, but not so much with temperature [17]. Bengtsson and Shukla [55] in 1988 cited limitations of climate data where they saw a need of multi-vatiate data to help with the forecasts. Inclusion of dense rainfall data can be used to overcome the limitations.

Farmer-recorded rainfall is another useful resource and can be used for research such as to understand spatial differences in rainfall magnitude and trends. This is particularly important since rain-fed agriculture is the main source of livelihood to a vast proportion of the population in Western kenya. To enhance food production in rain-fed agriculture, long-term rainfall records need to be recorded and analyzed to help inform farmers on the best Climate Smart Agricultural (CSA) practices to either mitigate or adapt against rainfall extremities.

For the farmers in Soin-Sigowett, cummulatively refereed to here as Nyando, the main observatory sites are in Kisumu and Kericho. These are more cosmopolitan areas and the people who receive this actual rainfall are not the main users. The users are several kilometers away from the KMS stations. Nyakach is more than 20 kilometers by road from the Kisumu Observatory while Soin-Sigowett is around 30 kilometers away. Due to distances from the main observatory centers, farmers believed that the rainfall recorded in the main observatory centers is not representative of the rainfall experienced locally. For

this reason, we compared the rainfall data from individual farmers with the overlapping KMS data. This would confirm whether or not the data is truly significantly different from KMS data.

2.3 Analysis of Long-Term Rainfall Data

2.3.1 Modeling of Rainfall Data

From literature, work has been done to understand the rainfall variability, model the rainfall occurrence and amount, and also model extreme rainfall. In this section we look at sample works done for each.

Researchers have analysed the inter and intra-seasonal rainfall variability using Zambia as a case study [9]. For this, the researchers based their discussions and analysis on events that play a significant role in growth and development of maize. They analyzed monthly summaries for possible trends by using polynomial trend up to cubic terms, and by using non-parametric spline functions. For analysis purposes, a clear definition for "rainy day" and for "start of season" had to be defined. Rainy day was defined to be a day with a minimum of 0.85 mm of recorded rain. Start of season was defined as a day after, either 1^{st} March or 1^{st} September, which received more than 20 mm rain in a span of three days. Coe and Stern [9] concluded by stating the need for skillful seasonal forecasts that would help farmers make informed decisions on their cropping pattern for the next day. This raised the issue of considering spatial variability and not just using the KMS recorded data.

Modelling has also been done to be able to understand the chance of rainfall occurring on a single day and possibly the amount. For this, Stern and Coe [2] first discussed on the distribution of rainfall amount on a given day of the year[2]. Stern and Coe [2] used daily rainfall data to fit non-stationary Markov chains models to rainfall occurrence. Analysis outputs showed that the daily rainfall amount followed a Gamma distribution, however, the scale and location parameters varied according to the time of year [2].

Models developed in the studies mentioned in [2, 12] were useful in simulation. This could still be improved by using daily data.

2.3.2 Time Series Analysis of Rainfall Data

The Ordinary Linear Models (OLM) and the Generalized Linear Models (MGLM) are the classical regression models used to predict values within specified ranges of independent variables. However, when dealing with time series data, one is interested to learn more about the period exceeding the last independent variable. The time series methods apply the idea of linear models, but uses a lag with the lagged values being the explanatory variables. The theoretical developments in time series analysis started early with stochastic processes, however, first actual application of Autoregressive (AR) models to data can be found back to the work of George Udny Yule through his three papers in the 1920's "On the time-correlation problem", "an investigation of a form of spurious correlation" and "On a Method of Investigating Periodicities in Disturbed Series, with Special Reference to Wolfer's Sunspot Numbers" in the 1920s. Over time ARMA Models were developed through Herman Wold starting with his PhD thesis titled "A Study in the analysis of stationary time series". The thesis resulted in the Wold decomposition which implies that any stationary discrete-time stochastic process can be decomposed into a pair of uncorrelated processes, one deterministic, and the other being a Moving Average (MA) process.

Time series analysis and forecasting methods, in particular ARIMA Models, have been used in many areas with a majority being in econometrics [49, 50, 51, 52] and engineering [53]. Climate data is time series data as many of its elements are measured daily. However, researchers have applied ARIMA models using monthly [23] and annual [24] summaries while focussing on evapotranspiration [19, 18]. It is important to generate models that utilize the daily values and see how well they do in comparison to the seasonal and monthly values.

Time series analysis is useful both to understand the data and for prediction. There are many studies across disciplines that have used the time series to understand trends and forecast events. Time series has been applied in the study of phenology of North

America using 21 years of data [38]. Reed *et al*[38] found there was a change in seasonality, which was an integrated response to climate change and changes in land use practices. For climate studies, long term climatic data is encouraged. For instance, long term climatic data was used to show that Seuss Wiggles were present 11,000 years ago in the Holocene [36]. In Africa, work has been done to understand the effect of climate change on the resurgence of Malaria in East African Highlands [14]. Hay *et al* [14] conducted both spatial and aerial analyses and concluded that there was not enough change in temperature and rainfall for climate to be responsible for Malaria cases in East African Highlands. This might have been as a result of sparse data on climate. Thus, work has been done to establish the availability of climate data and their quality for regions in Western Kenya. The data has been used to study risks involved with the farming of essential crops in Western Kenya [42]. In all the cases, historical data was used, but no forecasts were made.

Trend analysis has been applied to different climatic data sets across the globe. The Mann-Kendall analysis was conducted in most of the work done [30, 32], with some utilizing the Mann-Kendall despite correlations being present in data [16]. Lettenmaier *et al* [21] applied trend analysis using data from 1036 KMS and 1009 stream flow stations from (USA) using Mann-Kendall tests. The trend analysis has also been done using data from Japan, focussing on rainfall data [22]. The data used was annual and monthly summaries and the study found significantly negative trends. In Canada, annual total precipitation was used and the results showed increasing trends in annual precipitation totals [32]. The researchers used the methods by Von Storch and Navara [8] to remove effects of serial correlation before applying the Mann-Kendall test.

In Africa, trend analysis was applied to precipitation data from 96 KMS stations across Turkey [37]. Turkey coveres a land area of $783,562Km^2$. One station, on average, was used to cover a landmass of $8,162Km^2$. The researchers considered station specific trend analysis and region-based trend analysis. They used serial correlation, Mann-Kendall and sequential Mann-Kendall tests for the station-based trend analysis. They further used Sen's T test and Sen's estimator for the station based analysis.

2.4 Forecasting Time Series Data

In Kenya, weather forecasting is the reserve of KMS [35]. They monitor the oceanic wind movements, the Indian Ocean Dipoles, and predict the weather for seasons, months and weeks. For long range forecasts, the KMS use empirical statistical regression of (SST) , SST gradients and expected evolution of global SST patterns together with upper air circulations patterns. However, their focus is on wide area forecasts, and they use qualitative methods for presenting forecasted rainfall, cloud cover and temperature. They use the terms like ‶normal″, ‶above normal″and ‶below normal″to represent the intensity of rainfall in comparison to the long-term average. Kenya Meteorological Services (KMS) further selects an analogue year, which is a year in recent or distant past that exhibits similar characteristics as the forecast year, to help disseminate downscaled information. This is particularly useful for the small-scale farmers who depends on rain fed agriculture. In addition to weather forecast, researchers use mathematical Decision Support Tools (DSTs) to help farmers make informed decisions on risk involved. Such tools, most of which are crop models, require detailed information which are not provided in the KMS forecasts.

The DSTs are used by researchers across disciplines to help make informed decisions. They can be calculations on papers, apps or software. There are many of them and they cover all disciplines. For instance, in 2016, a study was done that catalogued 395 DSTs specific to Agriculture [15]. The models can be simplistic, for example Food and Agriculture Organization (FAO) CROPWAT [1]. This is a computer program for irrigation planning and management that utilizes the crop satisfaction index using ten-day (dekadal) data. Others like Decision Support Tools for Agrotechnology Transfer (DSSAT) [34]and APSIM [7] are more complex and require inputs like daily weather data and farm management options. The DSTs help researchers and farmers make future decisions using actual quantitative data. Most of the DSTs require a minimum of daily rainfall data in order to calculate water satisfaction indices for crops. Thus, consideration of forecast methods that can provide quantitative data is important. Time series is a main methodology used for forecasting quantitative data.

Time series analysis and forecasting utilizes the time lag to forecast the next few

events in similar time gaps. The ARIMA model integrates Auto Regression and the Moving Averages to generate a model of best fit. The ARIMA models have been used for rainfall forecasting, for instance, it was applied in Hyderabad region, India, where 93 years of annual data was train and 10 years as the test data [24]. The ARIMA models were preferred to the Autoregressive Moving Average (ARMA) when forecasting rainfall data over Thailand using data from 31 stations [28]. In addition, ARIMA and the Artificial Neural Networks (ANN) have been used to forecast weekly evapotranspiration for Northern Spain [19]. In Africa, the seasonal component in the Seasonal Autoregressive Integrated Moving Average (SARIMA) model was applied on data for Ashanti Region, Ghana, for the period 1974 to 2010, and used forecast monthly rainfall [23]. Psilovikos [18] found ARIMA (2,1,2)1,1,2)[6] to be the best model to forecast daily evapotranspiration over Nile Delta Region, Egypt. Psilovikos [18] focused on evapotranspiration, as opposed to rainfall. None of the studies modeled using daily rainfall data, which is an important component for use in DSTs.

Comparison of model performances are not new area of study. For instance, Adamowski [26] compared linear models, the ARIMA model and ANN to forecast peak daily water demands for the summer months residents of Ottawa, Canada. His models showed that ANN was better than the other two, however, he focused on rainfall occurrence compared to amount.

2.5 VAR(p) Models

The ARIMA models are univariate and focus on time lags for observed values, the moving average terms. In case data is not staionary, differencing is done to make it stationary. In this study we advanced the use of VAR and applied it to long term climate data for KMS Kisumu, and short term rainfall data from farmers. The long term climate data had rainfall, maximum and minimum temperatures as the endogenous variables.

The VAR models have been used in macroeconomic studies to provide very useful forecasts. Examples of VAR in use include the study of Australian Economy where 11 variables were used to study the economy for 19 years starting 1980 [5]. The co-integrated

VAR models have also been applied to find the direct effects of oil price shocks in the output and price for the G-7 countries [25]. The VAR model was used in Semarang-Central Java Indonesia where the rainfall, humidity and temperature were used to forecast with the results showing that it was better than ARIMA [39]. In this study, we used the VAR model to predict the next five events of rainfall and temperature data for KMS Kisumu. The events were in time gaps of days, month and seasons.

2.6 Conclusion

The literature shows that a lot of work has been done to incorporate time series models across disciplines. Studies also show that a lot of the focus has been on the use of summarized data, not daily. One area that is still not yet clear is applying the time series models to capture both space and time. The VAR models can be used to bridge this gap when data from multiple stations that are not far apart are used as MTS.

The VAR has been used extensively in the econometric models, but not applied a lot in the analysis of rainfall data and forecasting. What is also not clear is how well it can represent the forecast values when compared to the ARIMA models which are univariate.

Chapter 3

Methodology

The methods used can be classified into methods that checked on the data quality, methods used for univariate data analysis and forecasting, and Multivariate Time Series (MTS) model fitting and forecasting.

3.1 Sampled Farmers

All farmer recorded data used in this study were secondary data obtained from farmers under a CCAFS project in Nyando region and from the KMS. Nyando region (0.2833°S, 35.1167°E) was an opportune site to extend studies related to rainfall and tailor them for farmers' use. Most farmers are small-holder with average farm sizes of less than one hectares (ha). They have high levels of poverty, their environment has been subjected to widespread soil erosion, declined soil fertility and deforestation over decades of use. Despite a consistent decline in maize yields over the years, rain fed agriculture is predominant. Many farmers blame low erratic rainfall and climate shocks like floods, droughts and temperature stresses for this decline in yield.

For sampling, the Yamane [11] formula was used:

$$n = \frac{N}{(1 + N \times e^2)} \tag{3.1}$$

By substituting N with 1174 (Table 3.1) and setting e=0.1, the sample size was

calculated as:

$$n = \frac{1174}{(1 + 1174 \times 0.1^2)} = 92.15 \approx 93 \tag{3.2}$$

The project had sampled 100 (> 93) farmers from three farmer groups to collect rainfall data using simple raingauges. All 100 farmers came from the (10×10) km^2 region where CCAFS East Africa works in Nyando region, and its environs. The spatial representation is provided in the map on Figure 3.1. Figure 3.1 gives the spatial position for the farmers, with the top pointing to north. Each dot represents the physical location of the farmer. The farmers are all within the (10×10) km^2. The shapes and coler have been used to distinguish the farmers from different farmer groups. Friends of Katuk Odeyo Development Project (FOKODEP) had most farmers (circles) while Kapsokale had more spread among farmers (triangles). The square shapes represent the spatial distribution of farmers in NECODEP. Both FOKODEP and NECODEP farmers resided in Kisumu County while KAPSOKALE were from Kericho County. Farmers in Kericho had relatively bigger portions of land.

The project had selected the farmers randomly but used purposely allocated the proportion for each group. The three farmer groups were allocated the given slots. Despite having a many farmers, FOKODEP (circles) was allocated 40 slots since the farmers were concentrated in a very small area compared to NECODEP and KAPSOKALE farmers (Figure 3.1). The third column in Table 3.1 gives the ideal proportion had stratified sampling bee applied on the sample size of 93. Purposive allocation was applicable since farmers from Kapsokale (triangles) had a lot more spread, compared to those from FOKODEP.

Table 3.1: Farmer sampling according to the farmer groups

Farmer group	Number of farmers in group	Number of Sampled farmers	
		Calculated	Actual
FOKODEP	790	63	40
NECODEP	240	19	30
KAPSOKALE	144	11	30
Total	1174	93	100

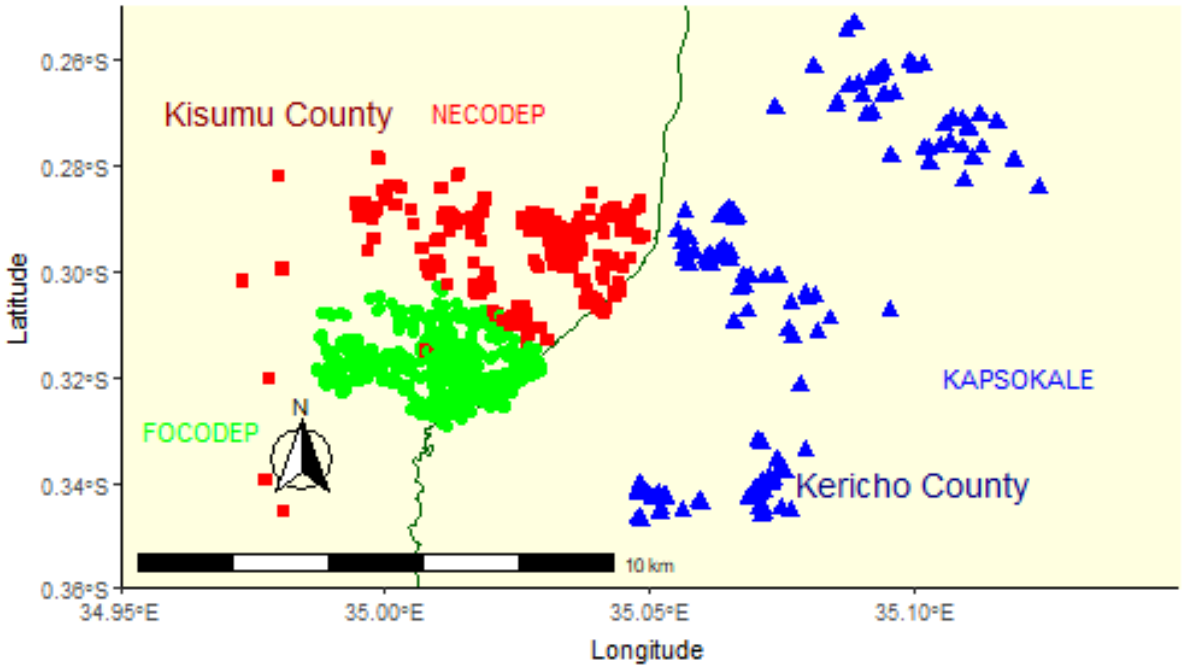

Figure 3.1: Spatial distribution of farmers' farms in the Nyando region

Four sets of rainfall data were used in this study. They included long term daily rainfall, for period 1961-2015, provided by KMS for Kisumu and Kericho Meteorological Stations, daily rainfall data from thirty (30) volunteer stations in the area, farmer-recorded daily rainfall data for the period 2014-2015 and qualitative data on farmers' recollection of previous rainfall events. The rainfall data from the volunteer stations varied in length for different stations but were all within the period 1961-2015. Since farmers were not compelled to carry out the survey, there was selective attrition for some due to various reasons which included theft and general depreciation of the rain gauges. Eventually, data was collected from forty-two (42) farmers who had recorded daily rainfall data for at least seventy-five percent (75%) of the period 2014-2015.

3.2 Analysis of Long-Term Rainfall

Spatial analysis essentially incorporates the exact positioning of the data collection. A common method is to use pictorial representations using the GIS software, with color codes used to help visualize the event of interest across different points. In this study, the spatial analysis included the discussion on how different data from farmers was compared to that from KMS situated 20 km away. Data from volunteer stations were used as control since they were closer to farmers hence more representative and they catered for different lenths of time hence gave a better picture of the region in comparison wwith the KMS data.

Farmers felt the KMS data was not representative of their locale. Hence, their recollection from past experiences was plotted on charts through focus group discussions. For the first spatial comparison, a descriptive analysis of long term KMS rainfall data was done using line plots and then compared to the qualitative plots by farmers.

3.3 Uni-Variate Forecasting Models

The univariate time series models were fit on daily, monthly, seasonal and annual rainfall values from KMS Kisumu. The choice time series model was the ARIMA model. An ARIMA was fit for each of the time periods given and used to forecast daily, monthly, seasonal and annual data.

3.3.1 ARIMA Model

Time series analysis is used to analyze data that are collected in equal time periods. They can be single variable, or multiple variable data. In this context, rainfall has been collected daily, hence time series methodology is appropriate. There are basic time series models, but the most commonly used model is the ARIMA model.

The ARIMA model is a time series model that constitutes Auto-Regressive (AR) of order p (AR(p)) and Moving Average (MA) of order q (MA(q)) components and a

differencing of order d. Stationary time series data does not have trend. In addition, it's mean and variance do not change significantly over time. The AR(p) implies that p lags of the data are used in the prediction. MA(q implies that q lags of the forecast errors are included in the model. In case no differencing is done, then ARIMA becomes ARMA.

3.3.2 Differencing in an ARIMA Model

To difference is to find the difference between the current term and the preceding term. If we let Y_t denote the original model, y_t the differenced model, and d be the differencing conducted, then

if d=0: $y_t = Y_t$ no difference

if d=1: $y_t = Y_t - Y_{t-1}$ first difference

if d=2: $y_t = (Y_t - Y_{t-1}) - (Y_{t-1} - Y_{t-2}) = Y_t - 2Y_{t-1} + Y_{t-2}$ second difference

if d=p: $y_t = (Y_t - Y_{t-1}) - (Y_{t-1} - Y_{t-2}) - \cdots - (Y_{t-p-1} - Y_{t-p})$ p-difference

When $d = 2$, the second difference is the first-difference of the first-difference $(Y_{t-1} - Y_{t-2})$, not difference from two previous periods $(Y_t - Y_{t-2})$.

3.3.3 ARIMA(0,1,0) Model

In a case with no AR or MA terms, but with one difference, then the model is a random walk. Say we want to predict the value at time t, $\hat{Y}_t$ and that μ is the average period-to-period change in Y, then the ARIMA(0,1,0) model can be written as:

$$\hat{Y}_t - Y_{t-1} = \mu$$
$$\Rightarrow \hat{Y}_t = \mu + Y_{t-1}$$

3.3.4 ARIMA(p,0,0) Model

The ARIMA(p,0,0) model is the equivalent of an AR(p) model. This is so since the data was not difference (d=0) and there are no MA terms (q=0). This model depends on the preceding p values in the forecast. It can be written mathematically as:

$$\hat{Y}_t = \mu + \sum_{i=1}^{p} \phi_i Y_{t-i}$$

The AR(1) model is different from the random walk (ARIMA(0,1,0)). The coefficient ϕ is not in the random walk, thus it is weakly dependent on the preceding term.

3.3.5 ARIMA(0,0,q) Model

The ARIMA(0,0,q) model is the equivalent of an MA(q) model. No difference has occurred and AR terms (p=0). Thus the model would depend on the preceding q forecast error terms. It can be written mathematically as:

$$\hat{Y}_t = \mu + \sum_{j=1}^{q} \theta_j e_{t-j}$$

In the equation $e_{t-j} = Y_{t-j} - \hat{Y}_{t-j}$, which is the forecast error at time $t - j$.

3.3.6 ARIMA(p,1,q) Model

A model with AR(p) and MA(q) terms but no differencing is referred to as an ARMA(p,q) model. It can be considered to be an ARIMA(p,0,q) model since no differencing has been done. The prediction equation can be represented mathematically as:

$$\hat{Y}_t = \mu + \sum_{i=1}^{p} \phi_i Y_{t-i} - \left(\sum_{j=1}^{q} \theta_j e_{t-j} \right)$$

When differencing occurs, then we have the ARIMA models. The forecast model for

ARIMA(1,1,1) is represented mathematically as:

$$\hat{Y}_t - Y_{t-1} = \mu + (\phi_1(Y_{t-1} - Y_{t-2})) - (\theta_1(e_{t-1} - e_{t-2}))$$

The forecast model for ARIMA(p,1,q) is represented mathematically as:

$$\hat{Y}_t - Y_{t-1} = \mu + \left(\sum_{i=2}^{p} \phi_i(Y_{t-i} - Y_{t-i-1})\right) - \left(\sum_{j=2}^{q} \theta_j(e_{t-j} - e_{t-j-1})\right)$$

and the forecast model for ARIMA(p,d,q) is represented mathematically as:

$$\begin{aligned}\hat{Y}_t - 2Y_{t-1} + Y_{t-d} = \mu &+ \left(\sum_{i=d+1}^{p} \phi_i(Y_{t-i} - 2Y_{t-i-1} + Y_{t-i-d})\right) \\ &- \left(\sum_{j=d+1}^{q} \theta_j(e_{t-j} - 2e_{t-j-1} + e_{t-j-d})\right)\end{aligned}$$

3.3.7 Determining the Number of Differences

Consider the following:

$$\begin{aligned}\text{Mean} &= E(Y_{t_1}) = \mu_t = \mu \\ \text{Variance} &= V(Y_{t_1}) = \sigma_t^2 = \sigma^2 \qquad \textit{and} \\ \text{Auto Covariance} &= V(Y_{t_1}, Y_{t_2}) = \gamma_{t_1,t_2} = 0 \qquad \text{where } t_1 \neq t_2\end{aligned}$$

For sample time data, $y_1, y_2, \ldots, y_n$, we calculate the mean, variance and auto covari-

ance using the formulae below

$$\text{Sample mean} = \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

$$\text{Sample variance} = s_0 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 \qquad \textit{and}$$

$$\text{Sample autocovariance} = s_k = \frac{1}{n}\sum_{i=1}^{n-k}(y_i - \bar{y})(y_{i+k} - \bar{y})$$

$$= \frac{1}{n}\sum_{i=1+k}^{n}(y_i - \bar{y})(y_{i-k} - \bar{y})$$

The auto-correlation is the correlation between observations at times t and $t+k$, where $k = 1, 2, 3, \ldots$ is the lag. Given that the recorded rainfall amount at time t is y_t, then the auto-correlation between y_t and y_{t-k} is given by

$$\rho_k = \frac{cov(y_t, y_{t-k})}{var(y_t)} = \frac{\sigma_k}{\sigma_0}$$

3.3.8 The SARIMA Model

A time series data may contain seasonality, that in most cases are regular. For instance, data may be affected by the month of recording. In such a case, seasonal difference of Y at time t is $Y_t - Y_{t-12}$. This removes the gross seasonality and trend in the data. Thus an ARIMA(0,0,0)(0,1,0) model has no non-seasonal difference (d = 0) while the seasonal difference is 1 (D=1). Considering the season to be months, then the following would hold.

if $d = 0$ and $D = 0$:$y_t = Y_t$	$ARIMA(0,0,0) \times (0,0,0)$
if $d = 0$ and $D = 1$:$y_t = Y_t - Y_{t-12}$	$ARIMA(0,0,0) \times (0,1,0)$
if $d = 1$ and $D = 1$:$y_t = (Y_t - Y_{t-12}) - (Y_{t-1} - Y_{t-13})$	$ARIMA(0,1,0) \times (0,1,0)$

The Seasonal Auto Regressive Integrated Moving Average (SARIMA) models assume the $ARIMA(p,d,q) \times (P,D,Q)$ nature where P is the number of Seasonal Auto-Regressive (SAR), Q is the number of Seasonal Moving Average (SMA) and D is the num-

ber of Seasonal Differences. Normally, only one seasonal difference is used in SARIMA models. The best practice is to use one order for seasonal differencing and one order for non-seasonal differencing in a SARIMA model.

SARIMA model was used incase where there was a seasonal contribution in the univariate time-series model.

3.3.9 Forecast Errors

In order to test the forecast errors, the Mean Error (ME), Root Mean Square Error (RMSE), Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE) are used to measure for both the train and test datasets. The measures are calculated as shown. Given that

- the training data for daily rainfall is given by $\{y_1^d, y_2^d, \ldots, y_{T^d}^d\}$ and the test data is given by $\{y_{T^d+1}^d, y_{T^d+2}^d, \ldots\}$
- the training data for monthly rainfall is given by $\{y_1^m, y_2^m, \ldots, y_{T^m}^m\}$ and the test data is given by $\{y_{T^m+1}^m, y_{T^m+2}^m, \ldots\}$, and
- the training data for seasonal rainfall is given by $\{y_1^s, y_2^s, \ldots, y_{T^s}^s\}$ and the test data is given by $\{y_{T^s+1}^s, y_{T^s+2}^s, \ldots\}$

ME is a forecast error, that is, the deviation of forecast value from the observed value in the test dataset. Since we used multiple time ranges, we calculated the ME for daily, monthly and seasonal data as given below.

$$ME_{Daily} = e_{T^d+h}^d = y_{T^d+h}^d - \hat{y^d}_{T^d+h|T^d} \tag{3.3}$$

$$ME_{Monthly} = e_{T^m+h}^m = y_{T^m+h}^m - \hat{y^m}_{T^m+h|T^m} \tag{3.4}$$

$$ME_{Seasonal} = e_{T^s+h}^s = y_{T^s+h}^s - \hat{y^s}_{T^s+h|T^s} \tag{3.5}$$

In addition, we calculated the scale dependent errors MAE and RMSE as shown in the equations below. They were calculated for daily, monthly and seasonal data using the formulae:

$$MAE_{Daily} = mean(|e^d_{t^d}|) \tag{3.6}$$

$$MAE_{Monthly} = mean(|e^m_{t^m}|) \tag{3.7}$$

$$MAE_{Seasonal} = mean(|e^m_{t^m}|) \tag{3.8}$$

We calculate the RMSE for each instance as:

$$RMSE_{Daily} = \sqrt{mean((e^d_{t^d})^2)} \tag{3.9}$$

$$RMSE_{Monthly} = \sqrt{mean((e^m_{t^m})^2)} \tag{3.10}$$

$$RMSE_{Seasonal} = \sqrt{mean((e^s_{t^s})^2)} \tag{3.11}$$

Apart from the scale dependent errors, we calculated the percentage errors, MAPE. The percentage error is all the time frames were same would be given as

$$p_t = 100 \times e_t / y_t \tag{3.12}$$

The MAPE is thereafter the mean of the percentage errors.

$$MAPE_{Daily} = mean(|p^d_{t^d}|) \tag{3.13}$$

$$MAPE_{Monthly} = mean(|p^m_{t^m}|) \tag{3.14}$$

$$MAPE_{Seasonal} = mean(|p^s_{t^s}|) \tag{3.15}$$

3.4 Multi-Variate Forecasting Model

For multi-variate time series analyis and forecasting, the VAR models were used. It was fit on the KMS rainfall and temperature data, which was long term data.

3.4.1 VAR(p) Model

The VAR(p) model is given by the equation:

$$\mathbf{Y_{t^d}^d} = \mathbf{a} + \mathbf{A_1 Y_{t^d-1}^d} + \mathbf{A_2 Y_{t^d-2}^d} + \cdots + \mathbf{A_p Y_{t^d-p}^d} + \epsilon_{\mathbf{t^d}} \tag{3.16}$$

where:

- $Y_{t^d}^d = (y_{1,t^d}^d, y_{2,t^d}^d, \ldots, y_{n,t^d}^d)'$ is an $(n \times 1)$ series of time series variables with n endogenous variables.
- $\mathfrak{a}$ is an $(n \times 1)$ vector of the intercepts for each of the n endogenous variables.
- $A_i; i = 1, 2, \ldots p$ are $(n \times n)$ coefficient matrices, for each of the p lags, and
- ϵ_{t^d} is $(n \times 1)$ vector of unobserved (i.i.d) zero mean error term

3.4.2 Lag selection for the VAR models

Four methods were applied in determining the lag selection for the VAR models. They were the Akaike's Information Criterion (AIC), , Hannan-Quinn Criterion (HQ) and Schwarz Criterion (SC) tests to determine the best lag. The AIC is defined as:

$$AIC(n) = \ln |\tilde{\Sigma}_u(n)| + \frac{2}{T} nK^2 \tag{3.17}$$

where T denotes the sample size, K is the dimension of time series, n is the order of VAR fit to data and Σ_u is the white noise covariance matrix. Further, $\tilde{\Sigma}_u(n)$ is the Maximum Likelihood Estimator (MLE) of Σ_u obtained by fitting the VAR(n) model to the data, nK^2 is the number of freely estimated parameters in the VAR model and $\frac{2nK^2}{T}$ is a

penalty term which converges to zero for $T \to \infty$. The best model is one for which the AIC is minimized.

The HQ, SC criterion and Final Prediction Error (FPE) are defined as

$$HQ(n) = \ln|\tilde{\Sigma}_u(n)| + \frac{2\ln(\ln(T))}{T} nK^2 \tag{3.18}$$

$$SC(n) = \ln|\tilde{\Sigma}_u(n)| + \frac{\ln(T)}{T} nK^2 \tag{3.19}$$

$$FPE(n) = \left(\frac{T + nK + 1}{T - nK - 1}\right)^K |\tilde{\Sigma}_u(n)| \tag{3.20}$$

The best model is one for which the HQ, SC and FPE are minimized.

3.4.3 Testing for Multi-Variate Stationarity

For an Multivariate Time Series (MTS) to be stationary, it means that its correlation information does not change over time. Given a data of $\mathbf{y}_{t,i}$ for $1 \leq t \leq n$ and $i = 2, 3, \ldots, k$, we can get two matrices $\mathbf{y}_{(t-1),i}$ and $\mathbf{y}_{(t-2),i}$. Then, if $\mathbf{y}_{t,i}$ is stationary, the correlation matrices $Corr(\mathbf{y}_{(t-1),i})$ and $Corr(\mathbf{y}_{(t-2),i})$ would not be statistically different. The two correlation matrices would be different in case of non-stationarity.

For $i = 2, 3, \ldots, k$, we do not expect all the time series data in the MTS to be stationary, hence co-integration. A univariate time series y_t is said to be integrated of order d, written $I(d)$, if it needs to be differenced d times to make it stationary. If two series$\mathbf{y}_{(t-1),i}$ and $\mathbf{y}_{(t-2),i}$ are both $I(d)$, then any linear combination of the two series will usually be $I(d)$ as well. However, if a linear combination exists for which the order of integration is less than d, say $I(d-b)$, then the two series are said to be co-integrated of order (d, b), written $CI(d, b)$. If this linear combination can be written in the form $\alpha^T \mathbf{y}_{t,i}$, where $\mathbf{y}_{t,i}^T = (\mathbf{y}_{(t-1),i}, \mathbf{y}_{(t-2),i})$, where α is a co-integrating vector.

The most common method to test for stationarity is by considering the unit tests. However, formal statistical tests lay assumptions and conduct a test of hypothesis based on assumptions. Some common formal tests include the Augmented Dickey-Fuller [46] and Phillips-Perron Tests [47], both of which focus on Univariate Time Series and are

based on the Dickey-Fuller test. The co-integrated Augmented Dickey Fuller Test and Phillips-Ouliaris tests [48] test for evidence of co-integration among the residuals between two time series. In this study, we settled on using Johansen's test since it could be applied on MTS with greater than two variables, it is based on accepted likelihood ration principle and it performs better than other methods [45].

The Johansen test checks for situation of no co-integration, that is, $\mathbf{A} = 0$. It does this by conducting an eigen value decomposition of $\mathbf{A}$. Johansen test tests sequentially whether the rank of $\mathbf{A}$, $r = 0, 1, \ldots, n-1$, where n is the number of time series under the test. The null hypothesis for the tests is that $H_0 : r \leq 0$ against the alternative that $H_1 : r > 0$. The *urca* package for R statistical software was used to conduct the Johansen's test. The package calculated critical values for up to 11 variables, but not more.

3.4.4 VECM Models for Farmers Data

Consider a VAR $\mathbf{y}_t = \mu + \mathbf{A}_1\mathbf{y}_{t-1} + \cdots + \mathbf{A}_p\mathbf{y}_{t-p} + \mathbf{w}_t$ where μ is the vector-valued mean of the series, $\mathbf{A}_i$ are the coefficient matrices for each lag and $\mathbf{w}_t$ a multivariate Gaussian noise term with mean zero. We can form a Vector Error Corrction Model (VECM) by differencing the time series data as:

$$\Delta\mathbf{y}_t = \mu + \mathbf{A}\mathbf{y}_{t-1} + \Gamma_1\Delta\mathbf{y}_{t-1} + \cdots + \Gamma_p\Delta\mathbf{y}_{t-p}\mathbf{w}_t \tag{3.21}$$

where $\Delta\mathbf{y}_t = \mathbf{y}_t - \mathbf{y}_{t-1}$ is the differencing operator, $\mathbf{A}_i$ is the coefficient matrix for the first lag and Γ_i are the matrices for each differenced lag.

VECM models helps us decide on the number of exogenous variables that can be used in a MTS. The best model is selected using AIC values with the low AIC values implying a good model.

3.5 An Overview of Analysis Applied on Different Datasets

The analysis conducted on different datasets are summarized in Table 3.2. The factors considered when deciding the length of data and how the data related with other datasets. For instance, there were Geographic Information System (GIS) data for farmers, but not for volunteer stations. Thus spatial autocorrelation could be calculated for farmers' data only. KMS had daily climate data for 54 years for the elements rainfall, minimum and maximum temperature. This data was analysed using univariate time series models (ARIMA and SARIMA) for daily data and summarised monthly, seasonal and annual rainfall totals. Presence of temperature data opened a window for multivariate time series models (VAR) which was later used with individual farmers models. There were no temperature data from farmers.

Table 3.2: Analysis methods applied on available data

Analysis	Data source				Section
	KMS	Volunteer Stations	Farmers Quant.	Farmers Qual.	
Descriptive Analysis	TRUE			TRUE	4.1
ARIMA Modelling	TRUE				4.2.2
SARIMA Modelling	TRUE				4.2.3
VAR Modeling	TRUE	TRUE	TRUE		4.3

In the next chapter, we use the methodology discussed to get output when applied on data. In addition, we discuss the output in relation to the objectives. The sequence of the analysis is as provided in the Table 3.2.The specific objectives are responded to starting from Section 4.2.

Chapter 4

Results and Discussions

4.1 Summary of the data

This study considers data from two sources: Daily rainfall and temperature data from the KMS Kisumu (1961-2015) and daily rainfall data recorded by farmers in Nyakach and Soin-Sigowett (16^{th} July 2014 and 31^{st} July 2015) but with some slight farmer variability. Some farmers stopped collecting earlier in the event that their rain gauges were destroyed or stolen.

The Kisumu KMS observatory is located at the Kisumu airport (0.1°S, 34.75°E). There were no missing data for the Kisumu station for the period 1961-2014. Table 4.1 gives the numerical summary of the overall daily rainfall pattern. There were days that had recorded rainfall of up to 128 mm which were floods.

Table 4.1: A summary of the long term rainfall in Kisumu

Event	Min	1st Qu.	Median	Mean	3rd Qu.	Max
Rain overall	0.000	0.000	0.000	3.717	2.200	128.600
Rain on rainy days only	0.90	2.40	5.90	11.24	14.70	128.60

A rainy day was defined as a day in which the amount of rainfall recorded was more than 0.85 mm [10, 2]. An extreme event was any value that exceed the third quartile by

1.5 times of the inter-quartile range. From Table 4.1, the lower boundaries of the extreme rainy day can be calculated as:

$$\text{Outlier lower limit for Kisumu} = 14.70 + 1.5 \times (14.70 - 2.40) = 33.15mm$$

Thus, rainy days that exceeded 33.15 mm were considered extreme rainfall events in Kisumu.

4.2 Forecasting Rainfall Data for KMS using ARIMA Models

First we focused on the long term univariate time series analysis and forecasting. The analysis used data from the KMS Kisumu, for the period 1961-2015. Daily data was used to generate the monthly, seasonal and annual summaries. All analyses were conducted in R. The models were fitted using a train dataset and a test dataset of a five-period time gap used for testing the accuracy of forecasts from the models.

Analysis using ARIMA was conducted for the daily, monthly, seasonal and annual values. An ARIMA(p,d,q) model constitutes AR(p), I(d) and MA(q). The Integrated (I) part considers the number of differencing used to make the data stationary. The AR component of the model considers the influence of the values of previous p terms on the current observation. The MA gives the influence of the previous q value of error terms on the current. The VAR model is a multivariate time series analysis model that applies the Auto Regression on a vector, depending on the specified lag. VAR(p) considers the effect of the last p terms in the vector on the events in current time.

In this study, the time series analysis was categorized into three. First, raw data was explored using the basic techniques of plotting and differencing. For the second step, the ARIMA and SARIMA models were fitted to the data and prediction conducted. The data was divided into the test and train data for forecasting purposes. Both ARIMA and SARIMA models assume data to be stationary, that is, has constant mean and variance over time.

4.2.1 Fitting ARIMA models to the rainfall data

The data was further plotted in order to check for trend, seasonality and cyclicity. A line plot was used (Figure 4.1) to explore it. From the figure, the daily and annual total rainfall exhibited white noise. However, monthly total rainfall had seasonality in them. In order to well differentiate noise from the signals, the Autocorrelation functions (ACFs) were plotted.

Figures 4.2 and 4.3 are best used together. In Figure 4.2, the ACF for daily rainfall data decreases exponentially. However, many of the ACF values at different lags have crossed the significance line. The ACF value is below the blue line at lag 9. ACF lines crossed the significance line after the first lag for monthly and before any lag for seasonal data. None of them crossed the blue line for the annual data.

For an AR model, the theoretical are equal to 0 beyond the order of the model. In this case we consider the values below the significance line. For an MA model, the theoretical PACF tapers toward 0 in some manner, however, the ACF will have non-zero autocorrelations only at lags involved in the model. In Figures 4.2 and 4.3, one would expect there to be no AR and MA terms for the annual data and AR(3) for the seasonal rainfall totals. There is no distinct pattern for the daily and monthly values. The daily, monthly and seasonal totals had significant ACF values at lags greater than 1. This justified the use of ARIMA models for the daily, monthly and seasonal data since at least one lag will be required.

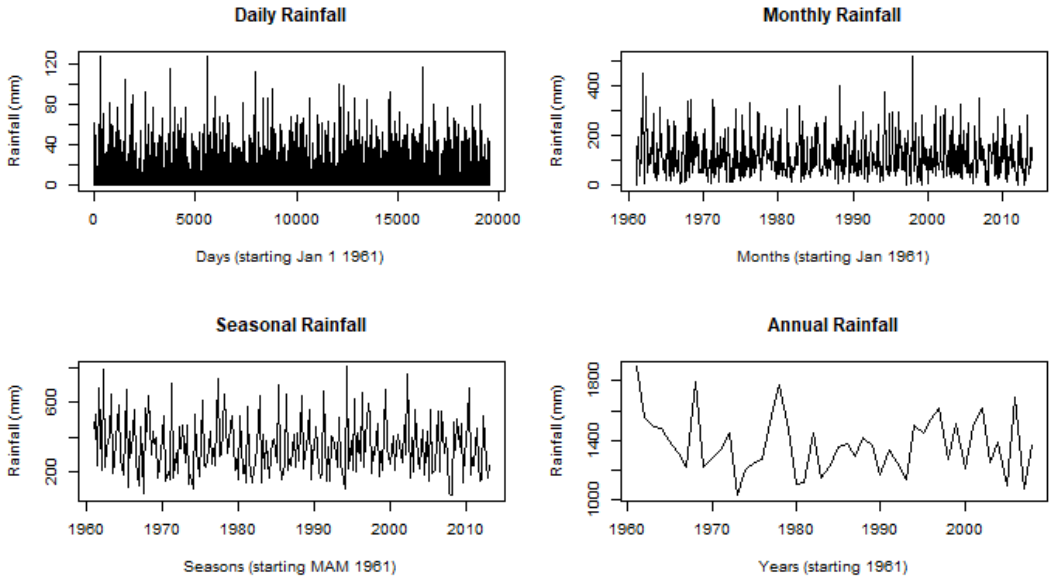

Figure 4.1: Line graphs showing totals for daily, monthly, seasonal and annual rainfall for Kisumu (1961 - 2014)

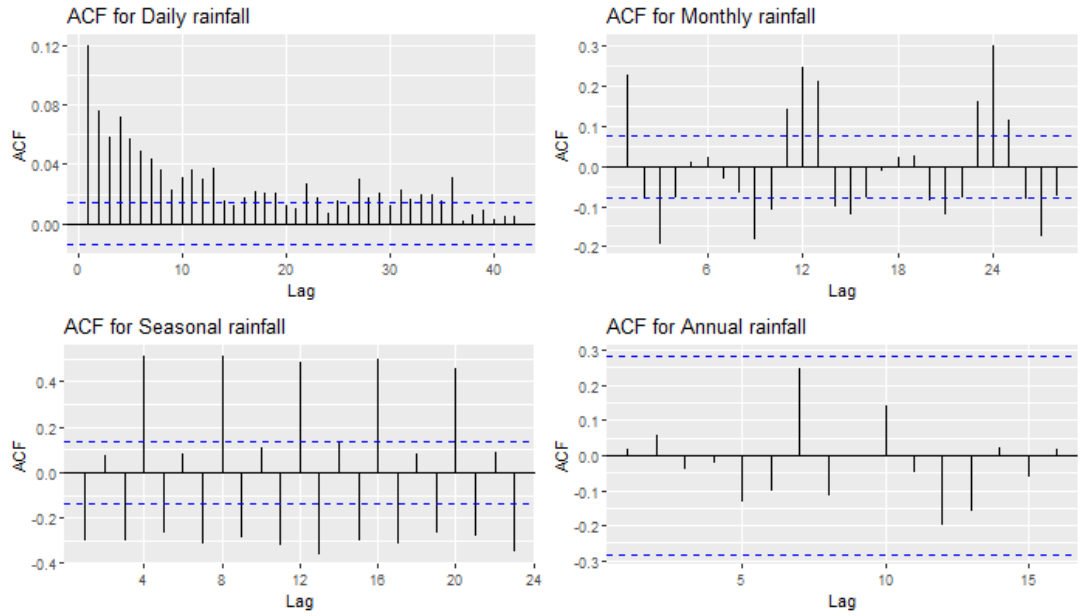

Figure 4.2: ACF plots for the KMS Kisumu rainfall data

The models were generated using the "auto.arima" function of the forecast library in R 3.6.0. The function used the Bayesian Information Criterion (BIC) and the AIC to select the best model. The model whose BIC and the AIC values were lowest were selected. As expected, the more the parameters, the lesser the AIC value. Thus, BIC was particularly

useful since it added a penalty term for every coefficient included in the model.

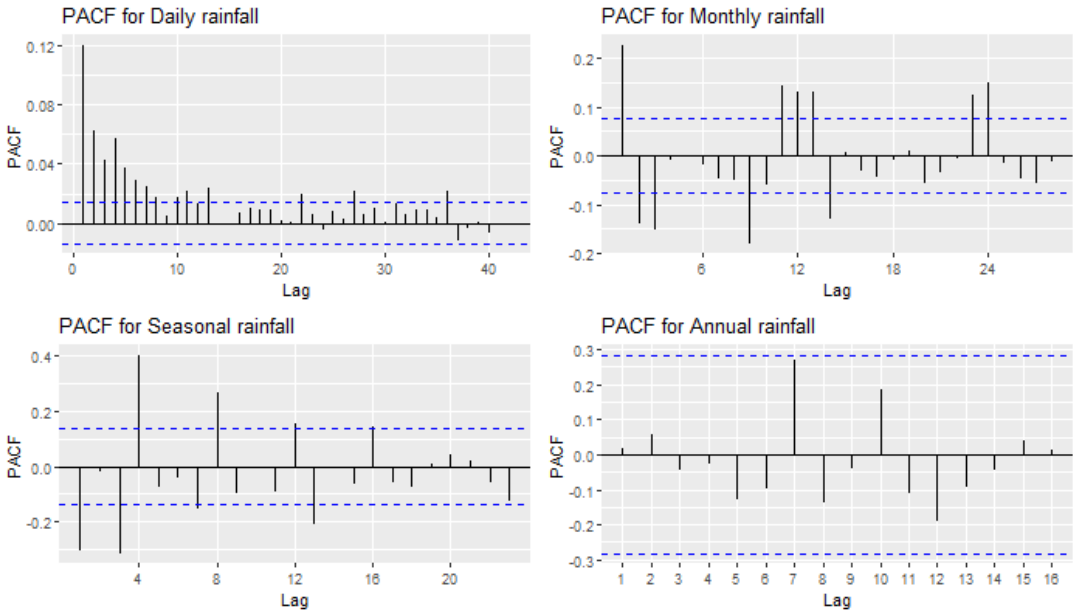

Figure 4.3: PACF plots for the KMS Kisumu rainfall data

The models were generated for the daily, monthly, seasonal and annual data. The suitable models selected were ARIMA(1,0,2) with non-zero mean, ARIMA(0,0,4)(2,0,0)[12] with non-zero mean, ARIMA(0,0,0)(2,1,0)[4] with drift and ARIMA(0,0,0) with non-zero mean for daily, total Monthly, total seasonal and total annual rainfall. There was seasonal differencing conducted on the seasonal data while annual data was purely white noise since the bars do not cross the blue horizontal lines(Table 4.5).

The time gap and the unit summed differed for daily, monthly, seasonal and annual data. We let $Y_{t^d}^d$ be the amount of rainfall experienced on day t^d, $Y_{t^m}^m$ be the total monthly rainfall for month t^m and $Y_{t^s}^S$ be the total seasonal rainfall for season t^s.

4.2.2 ARIMA for the Daily and Annual Data

The total annual rainfall was completely random and the best model selected was ARIMA(0,0,0) with non-zero mean. Because of this, annual data was not used for forecasting. The average total annual rainfall experience at KMS Kisumu was 1372.075 mm with a 95% Confidence interval of (1317.31, 1426.84) mm (Table 4.2).

Table 4.2: Fitting ARIMA model on KMS total annual rainfall data

parameter	Estimate	se	z	p	LB	UB
intercept	1372.075	27.942	49.104	$< 2.2e-16$	1317.31	1426.84

The best model selected for the daily rainfall was ARIMA(1,0,2) with non-zero mean, Table 4.3. This ARIMA model had statistically significant coefficients ($p-value < 0.05$) for the AR(1) and the MA(2) terms. The intercept was 3.7171 mm of rainfall. The model can be represented as below.

$$
\begin{aligned}
Y_{t^d}^d &= C + \phi_1^d Y_{(t^d-1)}^d + \theta_1^d w_{(t^d-1)}^d + \theta_2^d w_{(t^d-2)}^d + e_{t^d} \\
Y_{t^d}^d &= 3.717 + 0.918 * Y_{(t^d-1)}^d - 0.817 * w_{(t^d-1)}^d - 0.040 * w_{(t^d-2)}^d + e_{t^d}
\end{aligned} \tag{4.1}
$$

The model shows that the amount of rainfall experienced on the previous day contributed positively to the amount recorded for the next day. However, the error terms experienced on the preceding two days had reducing effect on the amount of rainfall experienced on the day of interest.

Table 4.3: Fitting ARIMA model on KMS daily rainfall data

parameter	Estimate	se	z	p	LB	UB
ar1	0.918	0.014	66.804	$< 2.2e-16$	0.891	0.945
ma1	-0.817	0.016	-52.138	$< 2.2e-16$	-0.848	-0.786
ma2	-0.040	0.008	-4.769	$< 2.2e-16$	-0.057	-0.024
intercept	3.717	0.116	32.016	$< 2.2e-16$	3.489	3.944

4.2.3 SARIMA for the Monthly and Seasonal Data

The monthly data and the seasonal data both exhibited seasonality of frequencies twelve (12) and four (4) respectively. The model selected for the monthly data was ARIMA(0,0,4)(2,0,0)[12] with non-zero mean (Table 4.4), while for the seasonal data was ARIMA(0,0,0)(2,1,0)[4] with drift (Table 4.5).

Table 4.4: Fitting ARIMA model on KMS total monthly rainfall data

parameter	Estimate	se	z	p	LB	UB
ma1	0.160	0.041	3.882	$< 2e-16$	0.079	0.240
ma2	-0.016	0.041	-0.382	0.703	-0.095	0.064
ma3	-0.120	0.042	-2.832	0.005	-0.203	-0.037
ma4	-0.049	0.040	-1.232	0.218	-0.128	0.029
sar1	0.137	0.041	3.348	0.001	0.057	0.217
sar2	0.242	0.041	5.968	$< 2e-16$	0.163	0.322
intercept	113.782	4.325	26.310	$< 2e-16$	105.306	122.258

The ARIMA(0,0,4)(2,0,0)[12] for the monthly total rainfall indicated that the amount of rainfall experienced in the same month for two preceding years contributed significantly to the amount of cumulative rainfall for the month of interest. The model is represented as:

$$\begin{aligned} y^m_{t^m} &= C + w_{t^m} + \theta^m_1 w^m_{(t^m-1)} + \theta^m_2 w^m_{(t^m-2)} \\ &\quad + \theta^m_3 w^m_{(t^m-3)} + \theta^m_4 w^m_{(t^m-4)} + \phi^m_2 Y^m_{(t^m-12)} + \phi^m_3 Y^m_{(t^m-24)} \end{aligned} \tag{4.2}$$

And after substituting the modeled coefficients

$$\begin{aligned} y^m_{t^m} &= 113.782 + 0.160 * w^m_{(t^m-1)} - 0.016 * w^m_{(t^m-2)} \\ &\quad - 0.120 * w^m_{(t^m-3)} - 0.049 * w^m_{(t^m-4)} + 0.137 * Y^m_{(t^m-12)} + 0.242 * Y^m_{(t^m-24)} \end{aligned} \tag{4.3}$$

In the model in Equation 4.3, the seasonal AR terms, the non-zero mean, and first and third MA terms had statistical significance at $\alpha = 0.05$. The second and fourth MA terms had p-values greater than 0.05. Though this was the best model for total monthly rainfall, using it for forecasting might result in some slight eviation from the actual data.

The ARIMA(0,0,0)(2,1,0)[4] for the seasonal total rainfall showed that the amount of rainfall experienced in the same season in the previous year contributed significantly

to the amount of cumulative rainfall for the current. Only the seasonal AR terms had statistical significance at $\alpha = 0.05$. The equation for the model may be represented as below:

$$Y^s_{t^s} - Y^s_{(t^s-4)} = C + \phi^s_1 \left(Y^s_{(t^s-1)} - Y^s_{(t^s-2)}\right) + \phi^s_2 \left(Y^s_{(t^s-1)} - Y^s_{(t^s-5)}\right) \tag{4.4}$$

Table 4.5: Fitting ARIMA model on KMS total seasonal rainfall data

parameter	Estimate	se	z	p	LB	UB
sar1	-0.696	0.067	-10.354	$< 2.2e-16$	-0.828	-0.564
sar2	-0.329	0.067	-4.899	$< 2.2e-16$	-0.461	-0.198
ARIMA(0,0,0) with non-zero mean for the total annual rainfall						
intercept	1372.075	27.942	49.104	$< 2.2e-16$	1317.31	1426.84

When the coefficients were substituted, the model becomes:

$$\begin{aligned} Y^s_{t^s} - Y^s_{(t^s-4)} &= -0.696 * \left(Y^s_{(t^s-1)} - Y^s_{(t^s-2)}\right) - 0.329 * \left(Y^s_{(t^s-1)} - Y^s_{(t^s-5)}\right) \\ Y^s_{t^s} &= Y^s_{(t^s-4)} - 0.696 * \left(Y^s_{(t^s-1)} - Y^s_{(t^s-2)}\right) - 0.329 * \left(Y^s_{(t^s-1)} - Y^s_{(t^s-5)}\right) \end{aligned} \tag{4.5}$$

4.2.4 Forecasting Precipitation Data

For the above analysis, the train data was used to fit models appropriate for the region. A test dataset was left out for forecasting purpose and to validate the models. Test data of 5 values were used for models for daily, monthly, seasonal and annual data. Although this figure was selected arbitrarily, a lot of importance is attached to the next day, month or season for general forecasts. The annual data was excluded from the forecasts.

Table 4.6 has several measures of forecast accuracy, they include the ME, RMSE, MAE, MAPE and MASE. The formulas for calculating the MEs in Table 4.6 are provided by Equations 3.3, 3.4 and 3.5. The MAEs are calculated as given in Equations 3.6, 3.7 and 3.8. The RMSEs are calculated as given in Equations 3.9, 3.10 and 3.11. The MAPEs are calculated as given in Equations 3.13, 3.14 and 3.15.

When evaluating the forecast errors, the smaller the measurements are to zero the better the forecast. In our case, the daily values were closest to the actual observed values in the test data. However, the errors were greater in test data when forecasting for monthly and seasonal totals. This shows that the model is more accurate for the daily data. However, since the denominator for $p^d_{t^d} = 100 \times e^d_{t^d}/y^d_{t^d}$ has the value zero, the MAPE recorded inf. Thus MAPE can only be used in cases when $y^d_{t^d} > 0$.

Forecast values are provided in Table 4.9 for the next five time periods, that is, days, months and seasons. The monthly and seasonal forecasts did not deviate very much from the long-term averages, except for the fourth quarter. The test data for days ran for the period 1st July 2014 to 5th July 2014. All the forecast values exceeded above 2 mm of rain. When you compare forecast values to the actual value in the last column of Table 4.9, there was some disparity.

Table 4.6: Forecast errors for forecasted daily, monthly and seasonal rainfall data for Kisumu

Period	Dataset	ME	RMSE	MAE	MAPE
Daily	Training	0.00	9.29	5.18	Inf
	Test set	-2.64	2.65	2.64	Inf
Monthly	Training	-0.460	70.661	53.812	Inf
	Test set	17.203	60.555	55.549	1938.448
Seasonal	Training	-6.609	127.521	97.762	37.787
	Test set	85.946	288.462	243.699	61.325

The forecast values for daily was ever positive. However the actual observed values for the test data were 0.0, 0.5, 0.0, 0.0 and 0.0 respectively. In this study, we wanted forecasts that present the local values as much as possible. One method could include the use of multiple vectors to help improve the forecast. The time series Model, VAR, has been used for this especially in econometric data. In our case, additional daily data included the maximum and minimum temperatures for KMS Kisumu. We therefore fit data to get the VAR model of best fit, and thereafter forecasted using it.

Table 4.7: Forecasted daily rainfall for five days using KMS data

Date	Forecast	Lo80	Hi80	Lo95	Hi95	Observed average
1-Jul-2014	2.479	-9.420	14.378	-15.719	20.676	0
2-Jul-2014	2.682	-9.278	14.642	-15.609	20.973	0
3-Jul-2014	2.767	-9.210	14.743	-15.550	21.083	10
4-Jul-2014	2.844	-9.146	14.835	-15.493	21.182	0
5-Jul-2014	2.916	-9.086	14.918	-15.440	21.271	0

We used values in Table 4.7 and calculated the mean absolute error and the root mean square error for the daily ARIMA forecast using Equations 4.6 and 3.9 to get:

$$\begin{aligned} MAE_{ARIMA} &= \frac{\sum_{i=1}^{n} |e_i^d|}{n} = \frac{\sum_{i=1}^{5} |e_i^d|}{5} \\ &= \frac{|(0-2.479)| + |(0-2.682)| + |(10-2.767)| + |(0-2.844)| + |(0-2.916)|}{5} \\ &= 3.6308 \end{aligned} \tag{4.6}$$

$$\begin{aligned} RMSE_{ARIMA} &= \sqrt{mean((e_{td}^d)^2)} \\ &= \sqrt{\frac{(0-2.479)^2 + (0-2.682)^2 + (10-2.767)^2 + (0-2.844)^2 + (0-2.916)^2}{5}} \\ &= \sqrt{16.44925} = 4.055767 \approxeq 5.06 \end{aligned} \tag{4.7}$$

The mean absolute error for the daily rainfall forecast under the model was 3.6308 mm while the root mean square error for the forecast was 5.06 mm.

Table 4.8: Forecasted monthly rainfall for five months using KMS data

Date	Forecast	Lo80	Hi80	Lo95	Hi95	Long term average
Jan 2014	96.865	6.309	187.422	-41.628	235.359	81.268
Feb 2014	102.798	11.096	194.501	-37.448	243.045	80.432
Mar 2014	140.143	48.430	231.856	-0.120	280.406	164.936
Apr 2014	125.086	32.732	217.439	-16.157	266.328	217.183
May 2014	134.493	42.031	226.955	-6.915	275.901	162.379

Table 4.9: Forecasted seasonal rainfall for five seasons using KMS data

Date	Forecast	Lo80	Hi80	Lo95	Hi95	Long term average
2013 Q2	546.586	381.574	711.597	294.222	798.949	544.498
2013 Q3	227.363	62.352	392.375	-25.000	479.727	228.443
2013 Q4	195.506	30.494	360.517	-56.858	447.869	323.608
2014 Q1	205.119	40.107	370.130	-47.245	457.482	264.217
2014 Q2	505.598	333.124	678.071	241.822	769.373	544.498

4.3 Using VAR models with Rainfall, Maximum and Minimum Temperature Data to Forecast KMS Data

The AR models regress the observation at time t on the lagged time intervals $t - i$, $i = 1, 2, \ldots$. The vector in the VAR adds other variables in AR model. We first used the KMS data, which had long term time series data for rainfall, minimum and maximum temperatures. The VAR model is a model that treats all variables as dependent variables. The model coefficients will be used for forecasting the values at time t^d for all the elements included in the model. In our case we had temperature, minimum and maximum temperatures. The resultant model could be used to forecast the three elements, but the

focus on this study was the rainfall data.

The VAR(p) model in this case was represented by Equation 3.16. Applying it on the rainfall, maximum and minimum temperature data, the coefficients were adjusted to be:

- $Y_{t^d}^d = (y_{1,t^d}^d, y_{2,t^d}^d, y_{3,t^d}^d)'$ is an (3×1) series of time series variables rainfall, minimum temperaure and maximum temperature respectively .

- a is a (3×1) vector of the intercepts for rainfall, minimum and maximum temperature respectively.

- $A_i; i = 1, 2, \ldots p$ are (3×3) coefficient matrices, and

- ϵ_{t^d} is (3×1) vector of unobserved i.i.d zero mean error term

The VAR model was fit with for rainfall, maximum and minimum tempeature values using the "VAR" function in R software. We fit VAR(P) for p= 1,2,3,4,5. When run, the output provides separate coefficients for rainfall, minimum and maximum temperatures. From Table 4.11, VAR(5) had higher R^2 for rainfall, maximum and minimum temperature, than VAR(4), VAR(3), VAR(2) and VAR(1). However all the VAR models had statistical significance. The VAR(5) model coefficients with rainfall as the dependent variableare provided in Table 4.10. In the VAR(5) model, the five lags of rainfall contributed significantly to rainfall at time t.

From the Table 4.10, the VAR(5) model overall had statistical significance (p-value $< 2.2e - 16$. However, the model explained only 2.975% ($R^2 = 0.02975$) of the total variation in the data. This might be occassioned by the high variability in rainfall as compared to the temperature data. The coefficients for rainfall had statistical significance at all lags (p-value < 0.05). On the other hand, the coefficients for minimum temperature had statistical significance for lags 1, 2 and 3 only. Only the coefficient for maximum temperature at the first lag had statistical significance. The model is represented in Equation 4.10.

Table 4.10: VAR(5) model for predicting *rainfall* considering lagged rainfall, minimum and maximum temperatures

	Estimate	Std. Error	t value	$Pr(>\|t\|)$	
Rain.l1	0.099	0.007	13.839	$< 2e-16$	***
MaxT.l1	-0.109	0.050	-2.189	0.02864	*
MinT.l1	0.233	0.061	3.827	0.00013	***
Rain.l2	0.046	0.007	6.426	1.34e-10	***
MaxT.l2	-0.112	0.057	-1.950	0.05120	.
MinT.l2	0.133	0.067	1.981	0.04758	*
Rain.l3	0.028	0.007	3.910	9.25e-05	***
MaxT.l3	0.066	0.057	1.158	0.24675	
MinT.l3	0.193	0.067	2.856	0.00430	**
Rain.l4	0.049	0.007	6.763	1.39e-11	***
MaxT.l4	-0.049	0.057	-0.851	0.39479	
MinT.l4	-0.041	0.067	-0.614	0.53901	
Rain.l5	0.031	0.007	4.347	1.39e-05	***
MaxT.l5	-0.002	0.050	-0.042	0.96651	
MinT.l5	0.101	0.060	1.682	0.09258	.
const	-1.847	1.492	-1.238	0.21575	
Residual standard error: 9.278 on 19518 degrees of freedom					
Multiple R-Squared: 0.02975, Adjusted R-squared: 0.029					
F-statistic: 39.9 on 15 and 19518 DF, p-value: $< 2.2e-16$					

The residual covariance matrix was found to be:

$$\Sigma = \begin{bmatrix} 86.083 & -0.699 & 0.407 \\ -0.699 & 1.772 & 0.097 \\ 0.407 & 0.097 & 1.196 \end{bmatrix} \tag{4.8}$$

And the residual correlation matrix

$$\rho = \begin{bmatrix} 1.000 & -0.057 & 0.040 \\ -0.057 & 1.000 & 0.067 \\ 0.040 & 0.067 & 1.000 \end{bmatrix} \tag{4.9}$$

Using $y^d_{1t^d}$, $y^d_{2t^d}$ and $y^d_{3t^d}$ as the rainfall, minimum and maximum temperatures respectively, then model can be represented as:

$$\begin{bmatrix} y^d_{1t^d} \\ y^d_{2t^d} \\ y^d_{3t^d} \end{bmatrix} = \begin{bmatrix} -1.847 \\ 6.492 \\ 3.320 \end{bmatrix} + \begin{bmatrix} 0.099 & -0.109 & 0.233 \\ -1.310e-2 & 0.547 & 1.683e-2 \\ -3.48e-2 & 0.125 & 0.474 \end{bmatrix} \begin{bmatrix} y^d_{1t^d-1} \\ y^d_{2t^d-1} \\ y^d_{3t^d-1} \end{bmatrix} + \begin{bmatrix} 0.046 & -0.112 & 0.133 \\ -1.297e-03 & 9.243e-02 & -2.366e-2 \\ 5.87e-3 & -1.054e-2 & 8.634e-2 \end{bmatrix} \begin{bmatrix} y^d_{1t^d-2} \\ y^d_{2t^d-2} \\ y^d_{3t^d-2} \end{bmatrix} + \begin{bmatrix} 0.028 & 0.066 & 0.193 \\ 3.560e-4 & 5.522e-2 & 1.730e-4 \\ 3.299e-3 & -5.075e-3 & 3.008e-2 \end{bmatrix} \begin{bmatrix} y^d_{1t^d-3} \\ y^d_{2t^d-3} \\ y^d_{3t^d-3} \end{bmatrix} + \begin{bmatrix} 0.049 & -0.049 & -0.041 \\ -2.240e-5 & 4.467e-2 & -2.829e-2 \\ 2.152e-3 & -8.246e-3 & 3.208e-2 \end{bmatrix} \begin{bmatrix} y^d_{1t^d-4} \\ y^d_{2t^d-4} \\ y^d_{3t^d-4} \end{bmatrix} + \begin{bmatrix} 0.031 & 0.101 & -1.847 \\ -6.271e-7 & 6.253e-2 & 4.843e-3 \\ 3.183e-3 & -1.787e-2 & 4.14e-2 \end{bmatrix} \begin{bmatrix} y^d_{1t^d-5} \\ y^d_{2t^d-5} \\ y^d_{3t^d-5} \end{bmatrix} \tag{4.10}$$

The output generated contained a set of coefficients for the VAR(5) model when the main dependent variable is rainfall (Table 4.10), maximum temperature (Table 4.12) and minimum temperature (Table 4.13). In the three tables, it can be seen that data from 19518 days were used to fit the VAR(5). Overall, each model had statistical significance (p-value: $< 2.2e-16$ for each fit model). Despite the model having statistical significance,

not all the coefficient estimates were statistically significant. Further, the three tables show that the VAR(5) model for rainfall explained 3% (Table 4.10) of total variabation in data. The model for maximum temperature explained 51% (Table 4.12) and the VAR(5) model for minimum temperature explained 38% of the variability (Table 4.13). The low value for the R^2 for the VAR(5) for rainfall can be explained by the high variability in rainfall values, which is not a case with temperatures. In this case, the VAR(5) is suitable for forecasting the temperature values.

The VAR(5) model was used to forecast for the same five days as used in Table 4.9. The forecast results are provided in the last three columns of Table 4.14. The observed values were used as the five lags used in the model. The model resulted in accurate prediction of the temperature values. However, the forecast errors were bigger compared to forecast errors when the ARIMA(1,0,2) was used (Table 4.9).

Table 4.11: Measure of variability accounted for in VAR models for lags 1, 2, 3, 4 and 5

p	R^2			p-value		
	Rainfall	Max_T	Min_T	Rainfall	Max_T	Min_T
1	0.020	0.49	0.36	$< 2.2e-16$	$< 2.2e-16$	$< 2.2e-16$
2	0.024	0.50	0.38	$< 2.2e-16$	$< 2.2e-16$	$< 2.2e-16$
3	0.026	0.51	0.38	$< 2.2e-16$	$< 2.2e-16$	$< 2.2e-16$
4	0.028	0.51	0.38	$< 2.2e-16$	$< 2.2e-16$	$< 2.2e-16$
5	0.030	0.51	0.38	$< 2.2e-16$	$< 2.2e-16$	$< 2.2e-16$

For demonstration, we used VAR(5) model since we had used similar lag used with the univariate data (Section 4.2.4). We had a test data of five time periods (days, months and seasons). We fit the VAR(5) model on the daily, monthly and was constructed for the monthly and seasonal data.

Table 4.12: VAR(5) model for predicting *maximum temperature* considering lagged rainfall, minimum and maximum temperatures

	Estimate	Std. Error	t value	$Pr(>\|t\|)$	
Rain.l1	-1.310e-02	1.029e-03	-12.734	$< 2e-16$	***
MaxT.l1	5.472e-01	7.172e-03	76.297	$< 2e-16$	***
MinT.l1	1.683e-02	8.729e-03	1.929	0.05380	.
Rain.l2	-1.297e-03	1.037e-03	-1.251	0.21110	
MaxT.l2	9.243e-02	8.205e-03	11.265	$< 2e-16$	***
MinT.l2	-2.366e-02	9.654e-03	-2.451	0.01427	*
Rain.l3	3.560e-04	1.039e-03	0.343	0.73176	
MaxT.l3	5.522e-02	8.222e-03	6.716	1.92e-11	***
MinT.l3	1.730e-04	9.682e-03	0.018	0.98575	
Rain.l4	-2.240e-05	1.038e-03	-0.022	0.98278	
MaxT.l4	4.467e-02	8.204e-03	5.445	5.24e-08	***
MinT.l4	-2.829e-02	9.651e-03	-2.932	0.00337	**
Rain.l5	-6.271e-07	1.035e-03	-0.001	0.99952	
MaxT.l5	6.253e-02	7.231e-03	8.647	$< 2e-16$	***
MinT.l5	4.843e-03	8.615e-03	0.562	0.57402	
const	6.492e+00	2.141e-01	30.321	$< 2e-16$	***
Residual standard error: 1.331 on 19518 degrees of freedom					
Multiple R-Squared: 0.5128, Adjusted R-squared: 0.5125					
F-statistic: 1370 on 15 and 19518 DF, p-value: $< 2.2e-16$					

Table 4.13: VAR(5) model for predicting *minimum temperature* considering lagged rainfall, minimum and maximum temperatures

	Estimate	Std. Error	t value	$Pr(>\|t\|)$	
Rain.l1	-0.0034870	0.0008454	-4.125	3.73e-05	***
MaxT.l1	0.1250813	0.0058913	21.231	$< 2e-16$	***
MinT.l1	0.4735711	0.0071709	66.041	$< 2e-16$	***
Rain.l2	0.0058796	0.0008519	6.902	5.30e-12	***
MaxT.l2	-0.0105425	0.0067403	-1.564	0.117810	
MinT.l2	0.0863381	0.0079303	10.887	$< 2e-16$	***
Rain.l3	0.0032985	0.0008532	3.866	0.000111	***
MaxT.l3	-0.0050750	0.0067544	-0.751	0.452445	
MinT.l3	0.0300881	0.0079531	3.783	0.000155	***
Rain.l4	0.0021523	0.0008528	2.524	0.011621	*
MaxT.l4	-0.0082467	0.0067396	-1.224	0.221112	
MinT.l4	0.0320817	0.0079279	4.047	5.21e-05	***
Rain.l5	0.0031826	0.0008499	3.745	0.000181	***
MaxT.l5	-0.0178675	0.0059403	-3.008	0.002635	**
MinT.l5	0.0413994	0.0070769	5.850	5.00e-09	***
const	3.3201300	0.1758745	18.878	$< 2e-16$	***
Residual standard error: 1.094 on 19518 degrees of freedom					
Multiple R-Squared: 0.3839, Adjusted R-squared: 0.3834					
F-statistic: 810.7 on 15 and 19518 DF, p-value: $< 2.2e-16$					

Table 4.14: Forecast results using VAR(5) model for daily rainfall and temperature data

Date	Observed			Forecast		
	Rainfall	Max_T	Min_T	Rainfall	Max_T	Min_T
21-Jun-2014	0.6	26.3	20.3			
22-Jun-2014	0.0	28.5	19.5			
23-Jun-2014	0.0	29.5	18.2			
24-Jun-2014	0.0	29.0	18.3			
25-Jun-2014	0.0	29.5	16.8			
26-Jun-2014	0.0	28.2	17.1	6.285	29.261	17.401
27-Jun-2014	0.5	30.2	17.5	5.756	28.824	17.127
28-Jun-2014	0.0	29.3	16.3	5.869	29.850	17.493
29-Jun-2014	0.0	30.0	18.8	5.582	29.481	16.806
30-Jun-2014	0.0	28.8	15.6	5.708	29.920	17.934

We calculate the mean absolute error and the root mean square error for the daily VAR forecast using Equations 4.11 and 3.9 to get:

$$\begin{aligned} MAE_{VAR} &= \frac{\sum_{i=1}^{n} |e_i^d|}{n} = \frac{\sum_{i=1}^{5} |e_i^d|}{5} \\ &= \frac{|(0-6.285)| + |(0.5-5.756)| + |(0-5.869)| + |(0-5.582)| + |(0-5.708)|}{5} \\ &= 5.74 \end{aligned} \tag{4.11}$$

$$\begin{aligned} RMSE_{VAR} &= \sqrt{mean((e_{td}^d)^2)} \\ &= \sqrt{\frac{(0-6.285)^2 + (0.5-5.756)^2 + (0-5.869)^2 + (0-5.582)^2 + (0-5.708)^2}{5}} \\ &= \sqrt{]33.06238} = 5.74999 \approxeq 5.75 \end{aligned} \tag{4.12}$$

The mean absolute error for the daily rainfall forecast under the VAR model was 5.74

mm (Equation 4.11). This is greater than the mean absolute error for the daily rainfall forecast under the ARIMA model which was 3.6308 mm (Equation 4.6). In addition, the root mean square error for the forecast under the VAR model was 5.75 mm (Equation 4.12) which was greater than the root mean square error for the forecast under the ARIMA model (Equation 4.7).

Both the MAE and the RMSE under the univariate ARIMA forecast (Equations 4.11 and 4.7) are smaller than the same under VAR (Equations 4.12 and 4.7 respectively). The univariate ARIMA forecasts are closer to the actual value and less diverse in this case. However, none of the two models was able to accurately forecast the case of zero (0) rainfall in the five days.

4.4 Comparing daily amount of rainfall between farmers location and Kisumu

The amount of rainfall in Kisumu and 60 farmers who had complete data was compared. This was done using the chi-square test and done on the days the data overlapped. This was between 16th June 2014 to 30th June 2014. The test was run for each day and the values were compared with each other. The Kisumu daily rainfall data was treated as the expected rainfall while the farmer daily rainfall data was treated as the observed rainfall. The results are provided in Table 4.15.

Daily data from 60 farmers' sites were compared to KMS data, however, only 22 farmers' data met the threshold of comparison of having at least one rainy day. There were four farmers' sites whose comparison with the KMS data yielded statistically significant output with p-value < 0.05. They are F09, F16, K05 and K08 (Table 4.15). For the four, we conclude that their rainfall patterns were completely independent from the KMS data.

Table 4.15: Chi-Square test results for fifteen days comparing individual farmer's rainfall data to Kisumu

Station	df	$\chi^2 - value$	p-value	Station	df	$\chi^2 - value$	p-value
F02	21	16.875	0.718655154	F22	30	32	0.36752736
F03	15	16	0.382051662	F23	6	0.576923077	0.996772673
F09	21	36.875	0.017398818	F24	33	36.875	0.294288265
F12	24	31.07142857	0.151733801	F25	27	36.875	0.097424373
F14	21	31.07142857	0.072477382	K01	24	18.33333333	0.786545547
F15	27	18.33333333	0.893001691	K02	12	16.875	0.154360204
F16	18	30.78125	0.030509052	K04	24	24.6875	0.422898113
F17	24	18.33333333	0.786545547	K05	15	30.78125	0.009396124
F18	24	18.33333333	0.786545547	K06	24	30	0.184751799
F19	12	1.363636364	0.999921932	K07	15	16.875	0.326391697
F20	15	16	0.382051662	K08	21	45	0.001731974

4.5 Comparing the number of rainy days between farmers location and Kisumu

The null hypothesis for the tests was that the number of rainy days at an individual farmer's locale is independent to the number of rainy days at KMS. In the table, the farmers' locale that did not experience rainfall in the time period has been excluded since the chi-square test requires at least a count of one for each of the two categories. In the case of farmers in Nyakach, all the p-values were greater than 0.05. Hence we reject the null hypothesis and conclude that there was not enough evidence to indicate independence in number of rainy days between farmers' locale and the KMS. The results are provided in Table 4.16.

From the two chi-square tests, it is clear that the rainfall experienced in Kisumu is not independent from that experienced by farmers in their locale. This is with the exception of the four when we consider the daily rainfall amount shown in in Table 4.15.

Therefore, the data from Kisumu KMS can be used to represent that of Nyando region (Soin Sigowett and Nyakach) in Kenya.

Table 4.16: Chi-Square test results for fifteen days comparing number of rainy days between farmers and Kisumu

Station	df	$\chi^2 - value$	p-value	Station	df	$\chi^2 - value$	p-value
F02	1	3.54E-33	1	F22	1	4.82E-31	1
F03	1	4.82E-31	1	F23	1	9.41E-31	1
F09	1	0.004783163	0.94486193	F24	1	1.246565934	0.26420938
F12	1	3.54E-33	1	F25	1	2.00E-30	1
F14	1	3.54E-33	1	K01	1	0.044642857	0.832662106
F15	1	0.044642857	0.832662106	K02	1	3.54E-33	1
F16	1	3.54E-33	1	K04	1	9.41E-31	1
F17	1	3.54E-33	1	K05	1	3.54E-33	1
F18	1	0.044642857	0.832662106	K06	1	1.05E-31	1
F19	1	1.05E-31	1	K07	1	3.54E-33	1
F20	1	3.44E-30	1	K08	1	3.30E-32	1

Chapter 5

Summary, Conclusion and Recommendetaions

5.1 Summary

In this study, we conducted a quality analysis for the farmer's data. This included the analyzing completeness of their data, and an analysis of its representativeness of the local setting. This was done using a standard dataset from the KMS data, with climatic data from nearby volunteer stations being used as control for local experience. The data showed that half of the farmers had good quality data, and well representative of the local setting.

We used the data from 60 farmers who had relatively good datasets and conducted spatial analysis by calculating the Morans Index for farmers' average rainfall and number of rainy days. The results showed that several distant farmers experienced similar rainfalls compared to close farmers. This can be attributable to other factors not collected in the study, like wind direction etc.

In the study, we conducted an analysis of farmer perception with respect to observed rainfall patterns. Historical rainfall data from KMS was used as the quantitative data with the famers perception scored on charts through participatory approaches being qualitatively alligned. Line graphs and descriptive summaries were used to present the long term KMS seasonal rainfall totals and seasonal number of rainy days. The results showed

that farmer perception is not sufficient to accurately inform the actual historical extreme events.

In the study, we fit ARIMA models on daily, monthly, seasonal and annual rainfall data from KMS. The best model selected for annual data was ARIMA(0,0,0) with non-zero mean while for daily data was ARIMA(1,0,2) with non-zero mean. Monthly data had a seasonal lag with the best model being ARIMA(0,0,4)(2,0,0)[12] while the seasonal total rainfall followed ARIMA(0,0,0)(2,1,0)[4]. The models were used to forecast with the daily forecasts having lower forecast errors for the test dataset than training daset Table 4.6.

VAR(5) was fit on the KMS data with rainfall, maximum and minimum temperatures as the endogenous variables. VAR(5) was a good model since it had the highest R^2 as can be seen in Table 4.11 in the models. The models were statistically significant.

The RMSE of the forecast under univariate ARIMA model (Equation 4.7) was smaller than the RMSE for forecast values under the multivariate VAR model (Equation 4.7). The univariate ARIMA ARIMA(0,0,4)(2,0,0) model was more accurate in the forecast than the VAR(5) model.

5.2 Conclusions

1. In this study, we compared long-term rainfall data from KMS to farmer perceptions. This helped us establish that the farmer perceptions from historical recollection is not very acurate. The farmers' information and KMS rainfa data on seasons with extreme events did not tally half of the time. The study further found out that though only few farmers experienced a statistically significantly different rainfall from KMS data, it is important to use the local data since it best represents them.

2. The first objective was to conduct univariate time series modelling using ARIMA on long term rainfall data for Kisumu KMS daily, monthly, seasonal and annual data and forecast rainfall for the different time periods. The forecast showed that forecasting daily data was resulted in less forecast errors than for seasonal and

monthly totals.

3. The second objective was to fit VAR(p) Model on long term daily climate data for Kisumu using rainfall, maximum and minimum temperatures and forecast, and determine whether it was suitable in comparison to the ARIMA models. The VAR models were better for the mximum and minimum temperatures as opposed to rainfall. The study established that for the KMS data, univariate ARIMA models were better for forecasting daily rainfall data in comparison to VAR(1), VAR(2), VAR(3), VAR(4) and VAR(5) models with minimum and maximum temperatures as endogenous variables.

4. There was not enough evidence to indicate independence in number of rainy days between farmers' locale and the KMS.

5.3 Recommendations

1. The study used VAR models which were very accurate. However, the VAR model lack the error term in the model, but conducts an Ordinary Least Squares method on the lagged values. We recommend the utilization of Vector Auto Regressive Integrated Moving Average (VARIMA) models in forecasting the daily farmer rainfall data.

2. The VAR might perform differently when the spatial coverage increases. We recommend the increase of spatial and temporal coverage, and using a different locality for future studies.

References

[1] M. Smith, 1992, A computer program for irrigation Planning and Management, *FAO Irrigation and Drainage*

[2] R. D. Stern and R. Coe, 1984, A model fitting analysis of daily rainfall data, *Journal of the Royal Statistical Society. Series A (General)*, 147, Part l,pp. 1-34

[3] N. MacKellar, M. New and C. Jack, 2014, Observed and modelled trends in rainfall and temperature for South Africa: 1960-2010, *South African Journal of Science*, Vol 110, No. 7, August

[4] S.S. Jones, R.S. Evans, T.L. Allen, A. Thomas, P.J. Haug, S.J. Welch and G.L. Snow, 2009, A multivariate time series approach to modeling and forecasting demand in the emergency department, *Journal of Biomedical informatics*, Volume 42, Issue 1, February 2009, Pages 123-139

[5] M. Dungey and A. Pagan, 2000, A structural VAR model of the Australian economy, *Economic record*, Vol. 76, No. 235, December . 321-342

[6] M. Hulme, R. Doherty, T. Ngara, M. New and D. Lister, 2001, African climate change: 1900-2100, *Climate research*, Vol. 17, No. 2.

[7] B. A. Keating, P. S. Carberry, G. L. Hammer, M. E. Probert, M. J. Robertson, D.Holzworth, D N. I. Huth, J. N. G. Hargreaves, H. Meinke and Z. Hochman, 2003, An overview of APSIM, a model designed for farming systems simulation, *European journal of agronomy*, Volume 18, Issues 3–4, January, Pages 267-288

[8] H. Von Storch, and A. Navarra, 2013, Analysis of climate variability: Applications of statistical techniques proceedings of an autumn school organized by the Commission of the European Community on Elba from October 30 to November 6, 1993, *Springer Science & Business Media*

[9] R. Coe and R. D. Stern, 2011, Assessing and addressing climate-induced risk in sub-Saharan rainfed agriculture: Lessons learned, *Experimental Agriculture*, Volume 47, April, pp. 395-410

[10] R. D. Stern and P. J. M. Cooper, 2011, Assessing climate risk and climate change using rainfall data–a case study from Zambia, *Experimental Agriculture*, Volume 47, April, pp. 241-266

[11] T. Yamane, 1967, Statistics: An Introductory Analysis, 2nd Ed., *New York: Harper and Row*

[12] E. Smith, 2005, Bayesian modelling of extreme rainfall data, *Thesis*

[13] P. Collier, G. Conway and T. Venables, 2008, Climate change and Africa, *Oxford Review of Economic Policy*, Volume 24, Issue 2, Summer, Pages 337–353

[14] S. I. Hay, J. Cox, D. J. Rogers, S. E. Randolph, D. I. Stern, G. D. Shanks, M. F. Myers and R. W. Snow, 2002, Climate change and the resurgence of malaria in the East African highlands, *Nature*, Volume 415, February , Pages 905–909

[15] D. C. Rose, W. J. Sutherland, C. Parker, M. Lobley, M. Winter, C. Morris, S. Twining, C. Ffoulkes, T. Amano and L. V. Dicks, 2016, Decision support tools for agriculture: Towards effective design and delivery, *Agricultural systems*, Volume 149, November, Pages 165-174

[16] D. H Burn and M. A. H. Elnur, 2002, Detection of hydrologic trends and variability, *Journal of hydrology*, Volume 255, Issues 1–4, 2 January, Pages 107-122

[17] R. Behnke, S. Vavrus, A. Allstadt, T. Albright, W. E. Thogmartin and V. C. Radeloff, 2016, Evaluation of downscaled, gridded climate data for the conterminous United States, *Ecological Applications*, Volume 26, Issue 5, July , Pages 1338-1351

[18] A. Psilovikos and M. Elhag, 2013, Forecasting of remotely sensed daily evapotranspiration data over Nile Delta region, Egypt, *Water Resources Management*, Volume 27, pages 4115–4130

[19] G. Landeras, A. Ortiz-Barredo and J. J. López, 2009, Forecasting weekly evapotranspiration with ARIMA and artificial neural network models, *Journal of Irrigation and Drainage Engineering*, Volume 135, Issue 3, June

[20] D. A. Jones and W. F. Wang, 2009, High-quality spatial climate data-sets for Australia, *Australian Meteorological and Oceanographic Journal*, Volume 58, Pages 233-248

[21] D. P. Lettenmaier, E. F. Wood and J. R. Wallis, 1994, Hydro-climatological trends in the continental United States, 1948-88, *Journal of Climate*, Volume 7, Issue 4, Pages 586–607

[22] S. Yue and M. Hashino,2003, Long term trends of annual and monthly precipitation in Japan, *Journal of the American Water Resources Association*, Volume 39, Issue 3, June, Pages 587 - 596

[23] A.R. Abdul-Aziz, M. Anokye, A. Kwame, L. Munyakazi and N. Nsowah, 2013, Modeling and forecasting rainfall pattern in Ghana as a seasonal ARIMA process: The case of Ashanti region, *International Journal of Humanities and Social Science*, Vol. 3 No. 3, February, Pages 224-233

[24] V. K. Somvanshi, O. P. Pandey, P. K. Agrawal, N. V. Kalanker, M. R. Prakash and R. Chand, 2006, Modeling and prediction of rainfall using artificial neural network and ARIMA techniques, *J. Ind. Geophys. Union*, Vol.10, No.2, April pp.141-151

[25] A. Cologni and M. Manera, 2008, Oil prices, inflation and interest rates in a structural cointegrated VAR model for the G-7 countries, *Energy economics*, Volume 30, Issue 3, May 2008, Pages 856-888

[26] J. F. Adamowski, 2008, Peak daily water demand forecast modeling using artificial neural networks, *Journal of Water Resources Planning and Management*, Volume 134, Issue 2, March

[27] H. Ashouri, K. L. Hsu, S. Sorooshian, D. K. Braithwaite, K. R. Knapp, L. D. Cecil and O. P. Prat, 2015, PERSIANN-CDR: Daily precipitation climate data record

from multisatellite observations for hydrological and climate studies, *Bulletin of the American Meteorological Society*, Volume 96, Issue 1, January, Page 69–83

[28] U. Weesakul and S. Lowanichchai, 2005, Rainfall forecast for agricultural water allocation planning in Thailand, *Science & Technology Asia*, Vol. 10, No. 3, July-September, Pages 18-27

[29] A. Salami, A. B. Kamara and Z. Brixiova, 2010, Smallholder agriculture in East Africa: Trends, constraints and opportunities, *African Development Bank Tunis*, No 105, April

[30] T. Murat, 1996, Spatial and temporal analysis of annual rainfall variations in Turkey, *International journal of Climatology*, Volume 16, Issue 9, September, Pages 1057-1076

[31] R. W. Katz, 2010, Statistics of extremes in climate change, *Climatic Change*, Vol 100, pages 71–76

[32] X. Zhang, L. A. Vincent, W. D. Hogg and A. Niitsoo, 2000, Temperature and precipitation trends in Canada during the 20th century, *Atmosphere - ocean*, Volume 38, 2000 - Issue 3, November, Pages 395-429

[33] H. Dietrich, T. Wolf, T. Kawohl, J. Wehberg, G. Kandler, T. Mette and J. Bohner, 2019, Temporal and spatial high-resolution climate data from 1961 to 2100 for the German National Forest Inventory (NFI), *Annals of Forest Science*, Volume 76, Article number 6, January

[34] J. W. Jones, G. Hoogenboom, C. H. Porter, K. J. Boote, W. D. Batchelor, L. A. Hunt, P. W. Wilkens, U. Singh, A. J. Gijsman and J. T. Ritchie, 2003, The DSSAT cropping system model, *European Journal of Agronomy*, Volume 18, Issues 3–4, January 2003, Pages 235-265

[35] Z. W. Shilenje and B. A. Ogwang, 2015, The role of Kenya meteorological service in weather early warning in Kenya, *International Journal of Atmospheric Sciences*

[36] D. J. Thomson, 1990, Time series analysis of Holocene climate data, *Phil. Trans. R. Soc. Lond. A*, Volume 330, Issue 1615, April

[37] T. Partal and E. Kahya, 2006, Trend analysis in Turkish precipitation data, *Hydrological processes*, Volume 20, Issue 9, June, Pages 2011-2026

[38] B. C. Reed, 2006, Trend analysis of time-series phenology of North America derived from satellite data, *GIScience & Remote Sensing*, Volume 43, Issue 1, Pages 24-38

[39] A. Nugroho, S. Hartati and K. Mustofa, 2014, Vector Autoregression (Var) Model for Rainfall Forecast and Isohyet Mapping in Semarang–Central Java–Indonesia, *Editorial Preface*, Vol. 5, No. 11, Pages 57-62

[40] L. Jones, E. Carabine, A. Hickman and L. Langston2014, Exploring the Role of Climate Science in Supporting Long Term Adaptation and Decision Making in Sub Saharan Africa, *Climate and Development Knowledge Network (CDKN)*

[41] P. J. M. Cooper, R. D. Stern, M. Noguer and J. M. Gathenya, 2013, Climate change adaptation strategies in Sub-Saharan Africa: foundations for the future, *Climate change—realities, impacts over ice cap, sea level and risks*, Pages 327-356

[42] J. Musyoka and E. Otumba, 2016, Assessing Agricultural Risks related to the Start, End and Length of the March–May Season in Nyando, *International Journal of scientific research and management (IJSRM)*

[43] E. M. Mugalavai, E. C. Kipkorir, D. Raes and M. S. Rao, 2008, Analysis of rainfall onset, cessation and length of growing season for western Kenya, *Agricultural and forest meteorology*, Volume 148, Issues 6–7, 30 June, Pages 1123-1135

[44] Tageo Website, 2021, *https://www.tageo.com/index-e-ke-weather-ke.htm*

[45] S. Johansen, 1995, Likelihood-based inference in Co-Integrated Vector AutoRegressive Models *Oxford University Press*

[46] R.Mushtaq, 2011, Augmented dickey fuller test, *SSRN*, Aug

[47] K. Kim and P. Schmidt, 1990, Some evidence on the accuracy of Phillips-Perron tests using alternative estimates of nuisance parameters, *Economics Letters*, Volume 34, Issue 4, December, Pages 345-350

[48] H. E. Reimers, 1992, Comparisons of tests for multivariate cointegration, *Statistical papers*33, December, pages335–359

[49] Contreras J., Espinola, R., Nogales, F.J., Conejo, A.J., 2003, ARIMA Models to Predict Nextday Electricity Prices, *IEEE Transactions on Power Systems*, Vol. 18, No. 3, pp. 1014- 1020.

[50] Aidan M., Geoff K. and Terry Q., 1998, Forecasting Irish inflation using ARIMA models, *Central Bank and Financial Services Authority of Ireland Technical Paper Series*, Vol. No. 3/RT/98, pp. 1-48.

[51] H. H. M. Hatta, F. M. Daud and N. Mohamad, 2018, An Application of Time Series ARIMA Forecasting Model for Predicting the Ringgit Malaysia-Dollar Exchange Rate, *Journal of Data Analysis*, Vol.1, No.1, Month , p. 42-48 42.

[52] K. Manoj and A. Madhu, 2014, An Application Of Time Series Arima Forecasting Model For Predicting Sugarcane Production In India, *Studies in Business and Economics*, Lucian Blaga University of Sibiu, Faculty of Economic Sciences, vol. 9(1), pages 81-94, April

[53] S.L.Ho and M.Xieb, 1998. The use of ARIMA models for reliability forecasting and analysis. *Computers & Industrial Engineering*. Volume 35, Issues 1–2, October, Pages 213-216

[54] A. J. Iseh and T. Y. Woma, 2013, Weather Forecasting Models, Methods and Applications *International Journal of Engineering Research & Technology*. Volume 2, Issue 12, December

[55] L. Bengtsson and J. Shukla, 1988, Integration of Space and In Situ Observations to Study Global Climate Change, *American Meteorological Society*. Volume 69, Issue 10, October, pp 1130–1143

[56] J. Hansen, M. Sato, R. Ruedy, K. Lo, D. W. Lea, and M. Medina-Elizade, 2006, Global temperature change, *The National Academy of Sciences of the USA*. Volume 103, September, pp 14288-14293

APPENDIX

A.1 Procedure for Sampling

```
> ## Allocatng numbers to farmers
> fokodep.poplation<-seq(1,790,1)
> necodep.poplation<-seq(1,240,1)
> kapsokale.poplation<-seq(1,144,1)
>
> ## Samping farmers from the three farmer groups
> set.seed(1000)
> fokodep.sample<-sample(fokodep.poplation,40);fokodep.sample
 [1] 260 599  90 544 406  54 580 457 169 201 273 589 247 673 593  57 382 491
[19]  66 451 126 620 223  56 421 518 346 629  70  94 432  38 426 732 386 530
[37]  16 591 440 239
> necodep.sample<-sample(necodep.poplation,30);necodep.sample
 [1] 240 127  27 151 189 168 165 132  61 118 235 157   9 231 143 155 160 123
[19] 198 107  92   6 166 208  52 195  69 222 228 218
> kapsokale.sample<-sample(kapsokale.poplation,30);kapsokale.sample
 [1]  83 102  51  60 120  21  32  27  13  22  41  24 122   9  75  18  62  35
[19]  53  55 116 143  73   2  66 130  10  89  20  47
```

A.2 Catalogue of Rainfall Recording Stations in Kisumu and Kericho

Station Name	Station Number	Latitude	Longitude	Start year	End Year
Kisumu P.C'S Office	9034004	-0.1	34.75	1903	-
Kisumu Meteorological Station	9034025	-0.1	34.75	1938	-
Kisumu New Prison	9034060	-0.07	34.72	1951	-
Kibos Kisumu Water Supply	9034069	0	34.82	1955	-
Kisumu Municipal Council	9034085	-0.1	34.75	1962	-
Kisumu K.U.R.(Marine Supt)	9034003	-0.1	34.75	1904	1940
Nyalunya, Kisumu	9034035	-0.32	34.92	1941	1947
Siaya Kisumu	8934083	0.07	34.3	1951	1952
Nyakach Dispensary, Kisumu	9034020	-0.38	34.93	1939	1954
Kisumu Ahero Health Centre	9034019	-0.15	34.92	1937	1962
Kiboswa,Kisumu	9034077	-0.02	34.78	1953	1971
Kericho District Office	9035003	-0.38	35.28	1904	
Jamji Estate (Kericho)	9035001	-0.48	35.18	1923	
Kericho Kabianga H. School	9035044	-0.43	35.13	1932	
Litein Mission,Kericho	9035059	-0.58	35.18	1935	
Kaisugu House,Kericho	9035075	-0.32	35.37	1939	
Kericho Chagaik Estate	9035235	-0.33	35.33	1954	
Laliat Farm Ainabkoi Kericho	9035200	-0.27	35.25	1959	
Ainamoi Chiefs Camp Kericho	9035199	-0.3	35.27	1960	
Kericho Timbilil	9035244	-0.35	35.35	1963	
Kericho Ngoina Estate	9035261	-0.55	35.05	1965	
Koiwa Estate,Kericho	9035260	-0.62	35.32	1965	

Station Name	Station Number	Latitude	Longitude	Start year	End Year
Kericho Manaret Settlement Schme	9035268	-0.7	35.07	1968	
Kipsitet Chief'S Office Kericho	9035269	-0.22	35.17	1968	
Hail Research Station Kericho	9035279	-0.37	35.27	1973	
Soitit Market,Kericho	9035322	-0.58	35.38	1983	
Karabwet Kericho	9035004	-0.38	35.33	1913	1937
Cheburget , Kericho	9035105	-0.92	35.12	1946	1952
Kaisugu Forest,Kericho	9035076	-0.33	35.37	1939	1956
Kerr Kericho	9035193	-1.65	36.65	1940	1961
Upper Chepsir Kericho	9035203	-0.28	35.42	1956	1962
Bujenge School Kericho	9035213	-0.27	35.33	1958	1962
Kericho Tea Researcg Institute	9035145	-0.35	35.33	1951	1964
Chebigen Kericho	9035186	-0.3	35.35	1939	1965