ADDIS ABABA UNIVERSITY

OFFICE OF GRADUATE STUDIES

DEPARTMENT OF STATISTICS

MODELING AND FORECASTING MONTHLY RAINFALL IN TIGRAY REGION: A CASE STUDY BASED ON MEKELE STATION

By Amaha Gebretsadikan

October 2010

ADDIS ABABA UNVERISITY

OFFICE OF GRADUATE STUDIES

DEPARTMENT OF STATISTICS

MODELING AND FORECASTING MONTHLY RAINFALL IN TIGRAY REGION: A CASE STUDY BASED ON MEKELE STATION

By Amaha Gebretsadikan

A THESIS SUBMITTED TO THE OFFICE OF GRADUATE STUDIES OF ADDIS ABABA UNIVERSITY IN PARTIAL FULFILLMENT OF THE REQUIREMENT FOR THE DEGREE OF MASTER OF SCIENCE IN STATISTICS

October 2010

ADDIS ABABA UNVERISITY

OFFICE OF GRADUATE STUDIES

DEPARTMENT OF STATISTICS

MODELING AND FORECASTING MONTHLY RAINFALL IN TIGRAY REGION: A CASE STUDY BASED ON MEKELE STATION

By Amaha Gebretsadikan

APPROVED BY BOARD OF EXAMINERS

.....	.....
Department Chairman	*Signature*
Emmanuel G. yohannes (Dr.)	.....
Examiner	*Signature*
Prof. Getu Hagos	.....
Examiner	*Signature*

ACKNOWLEDGMENT

Foremost the glory goes to the almighty god; through him all things are possible. In him,

I put my trust for protection and guidance. It is my greatest pleasure to take this opportunity to express my gratitude and thank you to all the people involved either direct or indirectly in my effort to successfully finish this thesis work.

It is my great pleasure to be grateful to my advisor Dr. M.K. Sharma, Associate Professor of Statistics at Addis Ababa University for his inspiration, valuable suggestions and providing helpful materials that contributed to the completion of the thesis work.

I would like to thank the National Meteorological Agency of Ethiopia for providing me the monthly rainfall data.

At last but not least I would like to extend my thanks to my friend Goitom Telele, for providing his laptop that helped me to accomplish the task conveniently.

Table of Contents **Pages**

LIST OF TABLES **Pages**

LIST OF FIGURES **Pages**

Acronyms

ARIMA	Autoregressive Integrated Moving Average
ARMA	Autoregressive Moving Average
ADF	Augmented Dickey Fuller
AR	Autoregressive
ACF	Autocorrelation Function
AIC	Akakie's Information Criterion
DF	Dickey Fuller
NMA	National Meteorological Agency
MA	Moving Average
MSE	Mean Square Error
MAPE	Mean Absolute Percentage Error
MAE	Mean Absolute Error
PACF	Partial Autocorrelation Function
RACF	Residual Autocorrelation Function
RPACF	Residual Partial Autocorrelation Function
SARIMA	Seasonal Autoregressive Integrated Moving Average
SAS	Statistical Analysis System
SPSS	Statistical Package for Social Science
SAR	Seasonal Autoregressive
SMA	Seasonal Moving Average
SSE	Sum of Squared Error
SST	Sum of Squared Total

ABSTRACT

In this study an attempt is made to explore the practical procedures in time domain univariate Box-Jenkins methodology for modeling and forecasting monthly rainfall in Tigray region. In particular, the study employs Seasonal Autoregressive Integrated Moving Average (SARIMA) model for monthly rainfall data collected by the National Meteorology Agency at Mekele station for the period from January 1975 to December 2009. Through the various model identification, estimation and diagnostics methods, we developed models that can adequately fit to the data. Residual analysis which is the important tool for diagnostic checks shows that there was no violation of assumptions in relation to model adequacy. Further model selection was performed at the forecasting stage using forecasting accuracy methods based on the validation period. The point forecast results showed a very close match with the pattern of the actual data and better forecasting accuracy in the validation period. Accordingly, the more parsimonious SARIMA $(0, 0, 1) \times (1, 1, 4)_{12}$ model is found appropriate to describe the observed data. Therefore, the results of the study indicate that SARIMA model of Box-Jenkins methodology allows in capturing more complex description of the seasonality, autocorrelation structure and non–stationary of the series and appears to be reasonably good in forecasting the monthly rainfall series. Future forecast results of the model show that there seems to be no trend of increasing or decreasing pattern over the period from January 2010 to September 2011.

Key words: Monthly rainfall, SARIMA, Box-Jenkins, Forecasting.

CHAPTER ONE

INTRODUCTION

1.1 Background of the Study

Rainfall is one of the natural climatic phenomena whose modeling and forecasting is challenging and demanding. Its forecast is of particular relevance to agriculture sector, which contributes significantly to the economy of the nation. In worldwide scale, numerous attempts have been made to predict behavioral pattern of rainfall using various techniques (Yevjevich, 1972; Dulleur and Kavas, 1978; Tsakiris, 1998). With an aim to investigate the temporal dynamics of rainfall and its distribution in Ethiopia, a study by Wing *et al.* (2008) has been undertaken using different time series methods with which the variability and trend in seasonal and annual rainfall were analyzed at water shed scale, at station, regional and national levels. Thus, it is important to obtain accurate rainfall forecast at regional, station or/and national levels in order to help decision makers improve their decisions by taking into consideration the available and future water resources. Study and development of models for climate variables, through forecasts, helps plan and program strategistists for better management and control, soil conservation, irrigation and drainage management, proper time for farm practices as well to make forecast for growth of crops, yield and post harvest management.

Many methods and approaches for formulating time series forecasting models are available in the literature. Development of forecasting models using time series analysis is important for series recorded over time. A recent study by Nail and Momani (2009) has used Box-Jenkins approach and revealed that this method posses many appealing features. Because, this model allows the researcher who has data from past period, rainfall as an example, to forecast future rainfall without having to search for other related time series data. As a result, Box-Jenkins modeling has been successfully applied in various hydrology, environmental management and industrial applications. The following are examples where time series analysis and forecasting are effective (Nail and Momani, 2009).

Water resources: Time-series analysis has become a major tool in hydrology. It is used for building mathematical models to generate synthetic hydrologic records, to forecast hydrologic events, to detect trends and shifts in hydrologic records and to fill in missing data and extend records.

Process control: Forecasting can also be an important part of a process control system through monitoring key processes. It may be possible to determine the optimal time and extent of control action; for example, a chemical processing unit may become less efficient as the duration of continuous operation increases. Forecasting the performance of the unit will be useful in planning the shutdown time and overhaul schedule.

Montgomery and Johnson (1976) considered Box and Jenkins models as probably the most accurate models for forecasting phenomena occurring in time series. Principally, according to Caldwell (2006), the Box-Jenkins method is particularly suited for development of models of processes exhibiting strong seasonal behavior. The author stated that in practice, methods for exponential smoothing to time series are often used. Although results are reasonably good, there are forecast techniques exploring the reliance among observations yielding better results; most of those forecast techniques are based on recent advances in time series analysis consolidated and developed by Box and Jenkins (1976) and further discussed in some other resources such as Chatfield (1996), Harvey (1993), Brockwell and Davis (1996), Caldwell (2006). Three basic stages are required to build models for time series: model identification, estimation of the parameters or fitting the identified model and diagnostic checks. The Box-Jenkins approch allows one to decide whether to go back to the identification stage or not, according to the fitting level that the model presents. The original Box-Jenkins modeling procedure involved an iterative three-stage process of model selection, parameter estimation and model checking as mention above. But, recent explanations of the process by Makridakis et al.(1998) often add a preliminary stage of data preparation and a final stage of model application (or forecasting including calculation and evaluation of the forecast).

i. Data preparation involves transformations and differencing. Transformations of the data (such as square roots or logarithms) can help stabilize the variance in a series where the variation changes with the level. The data are differenced until there are no obvious

patterns such as trend or seasonality left in the data. "Differencing" means taking the difference between consecutive observations or between observations a year apart.

ii. Model selection in the Box-Jenkins framework uses various graphs based on the transformed and differenced data to try to identify potential ARIMA processes which might provide a good fit to the data. Later developments have led to other additional model selection tools such as Akaike's Information Criterion.

iii. Parameter estimation means finding the values of the model coefficients which provide the best fit to the data. There are sophisticated computational algorithms designed to do this.

iv. Model checking involves testing the assumptions of the model to identify any area where the model is inadequate. If the model is found to be inadequate, it is necessary to go back to Step (ii) and try to identify a better model.

v. Forecasting is what the whole procedure is designed to accomplish once the model has been selected, estimated and checked.

Although originally designed for modeling time series with ARIMA processes, the underlying strategy of Box and Jenkins is applicable to a wide variety of statistical modeling situations. It provides a convenient framework which allows an analyst to think about the data, and to find an appropriate statistical model which can be used to help answer relevant questions about the data.

Due to well documentation about the Box-Jenkins method in vast literatures, the use of Box-Jenkins models for modeling and forecasting monthly rainfall will be explored in this study.

1.1 Statement of the Problem

Rainfall modeling and forecasting has for a long time been the subject of study by several researchers from various disciplines including climatology, meteorology and hydrology for different objectives. However, in perspective to the vital role in agriculture and rainwater resources management reliable rainfall forecasting at station, regional and national levels has become important in particular, in a country where agriculture is highly dependent on seasonal rainfall, like Ethiopia. In this regard, in the course of time, high motivation in developing time series models for modeling and rainfall forecasting has also become increasingly important.

So far, there is no forecasting model developed for monthly rainfall in Tigray region that aims to help decision makers for better preparations. This study is actually helpful to fill some of the gaps and could be used to alleviate the problem in getting appropriate and reliable monthly rainfall forecasting model. In addition, this study responds to the question of how one can use historical time series data to present a formal way of forecasting monthly rainfall using Box-Jenkins methodology (based on time domain approach).

1.3 The Case Study Area

Ethiopia with an approximate area 1,127,127 sq kms, constitutes ten regional states, of which, 7444 sq km is water. It has a tropical monsoon climate with a wide climatic variation according to wide varying topography and its location is approximately between 34^0 and 47^0 longitude and 4^0 and 14^0 latitude. Diverse rainfall and temperature patterns are largely the result of Ethiopia's location in Africa's tropical zone and the country's varied topography. Altitude-induced climatic conditions form the basis for three climatic zones: cool, temperate, and hot (US Library of congress, 2005). Variations in precipitation throughout the country are the result of differences in elevation and seasonal changes in the atmospheric pressure systems that control the prevailing winds. In January the high pressure system that produces monsoons in Asia crosses the Red Sea. Although these northeast trade winds bring rain to the coastal plains and the eastern escarpment in Eritrea, they are essentially cool and dry latitude and provide little moisture to the country's interior. Their effect on the coastal region, however, is to create a Mediterranean-like climate.

Winds that originate over the Atlantic Ocean and blow across Equatorial Africa have a marked seasonal effect on much of Ethiopia. The resulting weather pattern provides the highlands with most of its rainfall during a period that generally lasts from mid-June to mid-September (Tamiru, 2009).

Tigray is one of the regional states in Ethiopia located at the northern tip of the country, which is bounded by 13^0 and 14^0latitude and 36^0 and 42^0 longitude, with an area of some 10,000 square kilometers. The region shares common borders with Eritrea in the north, Afar in the east, Amhara in the south, and Sudan in the west. The rainy season of the region is between June and mid-September while the slack period is from December to April. Average rainfall of the region ranges between 500 and 900 mm per year, with a uni-modal pattern, except in the southern part of the study area where a second (smaller) rainy season locally allows growing two successive crops within one year (Nyssen *et.al*,2005). The climate of the region is characterized as "Kola" (semi arid) 39%, "Woina dega" (warm temperate) 49%, and "Dega" (temperate) 12%. The region is divided into north western and south lowlands (700-1500) meters above sea level and central high lands (1500-3000) meters above sea level. According to the new administrative set up, the region constitutes six zones. These are Western, North western, Central, Eastern, and Southern and Mekele . Out of these zones Mekele which is located in the center of the study area is the capital city of Tigray. The meteorological station Mekele is located at13^0, 30' latitude, 39^0, 29'longitude, 1905 altitude and has mean annual rainfall of 562 mm (Conway, 2000). This station is used as the source of data for this thesis work.

1.4 Objective of the Study

The main objective of this research is to develop a time series model to forecast monthly rainfall in Tigray region based on historical data recorded for the period from January1975 to December 2009 at Mekele station.

The specific objectives of the study include:

- To construct a model of rainfall by using Box and Jenkins time domain method.
- To explore the method based on the historical data.
- To asses future rainfall pattern in the study area.

1.5 Significance of the Study

The results of this research paper are hoped to be utilized:

- In decision making about rain water resources management. In particular, the forecasting model to be developed will be a valuable instrument for agriculture. Moreover, this model can provide guideline to develop other resources.
- As a basis for further study.

1.6 Scope of the Study

In view of the large number of aspects of rainfall modeling, it would be an impossible task to investigate rainfall modeling as a whole in one research work. In addition, it may be too tedious to assess rainfall modeling through different models of time series analysis. One aspect of rainfall modeling with which researches have been concerned over the last years is prediction of future values at different periods of time (i.e. daily, monthly or yearly) for different objectives. As a consequence, this study focuses on developing a Box-Jenkins time domain forecasting model for monthly rainfall at station level in which the specified study site is the target. So as to maintain the forecasting ability of the method, the forecasting horizon is short–term basis.

1.7 Limitation of the study

The limitation of this study, like other similar studies, is that the study deals with rainfall forecasting in a developing country where the available information is not adequate enough to address the forecasting problem. Years of civil war have limited historical data from the region. For example, Mekele station has data missing for years from 1990 and 1991 because of civil conflicts. In this study, monthly rainfall data of 35 years long is used to construct forecasting model employing time domain Box-Jenkins method. Mainly, based on frequency domain, investigating the cyclical pattern, and drought occurrence cycles by estimating a spectral density function might be important. However, the annual amount of data required (a minimum of 100-200) Chatfield (1996), for this technique of analysis has not been recorded.

CHAPTER TWO

LITERATURE REVIEW

2.1 Introduction

Univaraite time series analysis and forecasting has become a major tool in different applications in hydrology, environmental management, and climatic fields. There are time series methods used for modeling and forecasting rainfall data in numerous literatures. According to Pankratiz (1983) the Box and Jenkins method is the most general way of approaching to forecast unlike other models, there is no need to assume initially a fixed and specified pattern. The univariate Box and Jenkins models are useful for analysis of single time series. But, there are many situations in which modeling and better forecasting performance are possible. Transfer function methods, for example, are a special case of general regression approach with ARIMA models. It can cope with two or more independent variables or inputs to the target variable where the effect of these inputs is only unidirectional effect.

2.2 Rainfall Characteristics

Rainfall is an end product of a number of complex atmospheric processes, both in space and time. Rainfall (with wind, temperature, and clouds) is the most striking element of climate. When water vapor in the atmosphere condenses sufficiently, rain water returns to the ground to complete the water cycle. Water in all its forms and in all its various activities plays a crucial role in sustaining both the climate and life. In particular, the atmosphere self-cleanses naturally, through which precipitation removes pollutants from the air. In addition, rain water is the most significant climatic factor affecting all aspects of the economy of tropical Africa (Balek, 1977). Rainfall varies with latitude, elevation, topography, seasons, distances from the sea, and sea-surface temperature. Rainfall amounts are heavy at low latitudes, there is minimum rainfall about the tropics, and a second maximum occurs around 50^0(Linacre and Greets, 1997). To fully

appreciate the significance of a rainfall anomaly, it is necessary to know its relative size and expected frequency of occurrence as well. One of the approaches which could solve this problem is through intentional approach. Intentional approach is an anticipation effort through development of reliable seasonal forecasting technique and method also through various models and data application. In any attempt of modeling, However, the characteristics of rainfall in the past are the only accessible guide for advanced decision making regarding water resources development and management, particularly for a fragile agrarian economy like that of Ethiopia (Osman, and Sauerborn, 2002).

2.3 Rainfall Seasonality and Modeling

Many time series data exhibit fluctuations which is periodic within a year or a fraction of thereof. Due to the rotation of the earth around the sun the monthly rainfall time series exhibit a yearly periodicity in addition to the other components (Dulleur and Kavas, 1978). The proper treatment of such seasonality, whether stochastic or deterministic is the subject of large literatures. Seasonality may be modeled using time domain and its presence detected by plotting autocorrelation coefficients versus to finite lags or the untransformed rainfall series. Deterministic seasonality, in the form of model parameters that vary deterministically with season offers no great conceptual problems but many practical ones. Stochastic seasonality is modeled with the form of seasonal unit roots and seasonal differencing of the data removes the unit roots to render stationary data. The autocorrelation function has the appearance of a sinusoidal function with a 12 month period and spectral density function exhibits a discrete spectral component at the Frequency of 1/12 cycle per year.

A deterministic method does not consider the random effect of various input parameters. A stochastic approach, however, employ the concept of probability to a certain extent. Rainfall is periodic and stochastic in nature because it is affected by climatologically parameters i.e. variations of periodic and stochastic climates become periodic and stochastic components of rainfall. As a result, it should be modeled by considering both the stochastic and deterministic parts of the process. Assuming all factors known, rainfall is a function of the stochastic variation of climate (Yevjevich, 1972). Stochastic analysis of rainfall time series should provide a

mathematical model that will account for the deterministic and stochastic parts and will also reflect the variation of the rainfall.

Seasonality is a phenomenon commonly found in environmental, meteorological and other time series. Seasonal series are such that similarities occur at equivalent parts of a cycle. In particular, we say that a series exhibits periodic behavior with period s when similarities in the series occur in s basic time interval. In monthly data series, the basic time interval is one month and the period is $s = 12$ period.

Three types of seasonal time series models are commonly used to model meteorological and hydrological processes which usually have strong seasonality (Hipel and McLeod, 1994): 1) seasonal autoregressive integrated moving average (SARIMA) models; 2) deseasonalized ARMA models; and 3) periodic ARMA models. In another way, annual rainfall data that does not exhibit regular seasonality behavior can usually be modeled using transfer function plus noise and fast furrier transformation methods (Yilma *et al.*, 1994). The stochastic nature of monthly rainfall as a function of time, however, has been frequently modeled with Seasonal Autoregressive Integrated and Moving Average (SARIMA) (Hipel and McLeod, 1994). Usually, 12-lag differencing removes the annual cycle in a monthly time series but may introduce periodicities in the continuous spectral density which are not due to any discrete component of spectral component created by natural cycle (Dulleur and Kavas, 1978).

2.4 Quantitative Forecasting Methods

There are many quantitative methods of model building and forecasting used in various fields. These forecasting methods can be categorized into two groups: causal and time series. The causal method attempts to identify independent variables and their relationship to the variable of interest, the dependent variable. Changes in the independent variables are expected to cause changes in the dependent variable. By finding the proper relationship of the independent to dependent variables, models can be built which will be used to forecast dependent variables, given an input of the independent variables. One of the drawbacks to the causal method is that in some cases it is very difficult, if not possible, to find independent variables that can entirely explain the occurrences of the dependent variable. In addition, even if an accurate model can be

formulated, it is only as good as the ability to predict the future values of the independent variables (Thomopoulos and Nick, 1980).

2.5 Review of Forecasting Methods

The two major forecasting methods commonly in use are univariate and multivariate methods. The univariate method refers to a model based on fitted current and past observation of a given variabke. Different univariate forecasting procedures available include extrapolation of trend curves (long-term forecasting), exponential smoothing, Hot-Winters procedure, and box-Jenkins methods.

Earlier procedures for forecasting a time series from its current and past values were of a somewhat ad hoc nature, though theoretical justifications for their use were generally available. In particular, there exists a group of techniques, classed under the heading "exponential smoothing", characterized by the property that for any one of these techniques a single model was employed for forecasts from any time series. Nevertheless, exponential smoothing method is widely used in business and industry. This method should only be used in its basic form for non-seasonal time series showing no systematic trend or with effects that can be measured and removed to produce a stationary series. Exponential smoothing can readily be generalized to deal with time series containing trend and seasonal variation. The resulting procedure is usually referred to as the Hot-Winters procedure (De Gooije and Hyndman (2006).The remarkably good forecasting performance of smoothing methods has been addressed by several authors. Simple exponential smoothing is optimal for a wide range of data generating processes. Hyndman (2001) revealed that simple exponential smoothing performs better than first order ARIMA models because it is not subject to model selection problems, particularly when data are non-normal. It is common practice to restrict the smoothing parameters to the range 0 to 1. A computer program could be written in such a way that, for any particular series, forecasts by the class of exponential methods could be immediately generated without manual intervention. Such a procedure can be described as fully automatic.

In a series of literature and a subsequent book by Box and Jenkins (1976) outlined in considerable detail a strategy for time series forecasting. In a very important sense their approach represented a radical departure in forecasting methodology. The Box-Jenkins approach is not fully automatic. These authors propose a class of models, and a strategy by which for any given series a particular model is chosen from this class according to the properties of the individual time series under study. Thus the form of the eventual forecast function is dictated, to a large extent, by the data - a principle known as "letting the data speak for it". The user of the Box-Jenkins method is allowed a good deal of freedom of choice and, correspondingly, is required at various stages of the procedure to exercise judgments in the choice of an appropriate model. One might thus expect the Box-Jenkins procedure to be rather more versatile, in terms of its areas of applicability, than many of its competitors. The success of the Box–Jenkins methodology is founded on the fact that the various models can, between them, mimic the behavior of diverse types of series—and do so adequately without usually requiring very many parameters to be estimated in the final choice of the model. The facility afforded by the Box-Jenkins approach for a choice of forecast function appropriate to the particular problem under consideration is, the most distinctive features of the method and its principles virtue (Newbold, 1975).

The univariate Box –Jenkins methods are useful for the analysis of a single time series. In such a case we basically limit our modeling to the information contained in the series' own past. In many cases, however, we may be able to relate the response of one series not only to its own past values, but also to the past and present values of other related (stochastic) time series. In this manner, "transfer function models" can be constructed in such a way that effective merging of the basic concept of general regression model with that of ARIMA model is possible. It is often the case that changes in some series y are anticipated by changes in a related series x, is special class of multivariate methods. Clearly possession of information on some variables is potentially of great value in forecasting. But, in this methodology a great deal of care is needed in the construction of models relating time series, as is illustrated by (Caldwell, 2006). In particular, dangers of "discovering" spurious relationships must be circumvented. A close look at the literature on the subject reveals that there is no unified mathematical model with a universal acceptability (Tsakiris, 1998). But rather, the methods to be used depend on the degree of accuracy, objective of the forecast, and the properties of time series.

2.6 Conceptual Framework for Time Series Modeling

Time series models attempt to forecast the future values by analyzing the past. Time series considers historical data and attempts to derive some process which will explain those occurrences and predict future values. Much statistical theory is concerned with random samples of independent observations. The special feature of time series analysis is the fact that successive observations are usually not independent and that the analysis must take series analysis techniques to identify the patterns which typically exist (Diebold *et.al.* 2006)

There are two broad approaches to time series modeling. The first approach (time domain) represents time series as a function of time and is used to obtain the trend component and, then, to propose a prediction model. The second approach deals with the frequency domain, to determine the periodic components of the series. Inference based on the autocorrelation function, as in Box-Jenkins methodology, is often known as time series modeling in time domain. An analytically equivalent way of viewing the data is to transform the autocorrelation function into the frequency domain, in which the data are analyzed in terms of their cyclical properties. This approach to time series modeling is called spectral analysis, and it provides different insight to the properties of the time series.

Time series plot: When presented with a time series, the first step in the analysis is usually to plot the data and to obtain simple description measures of the main properties of the series via a visual inspection of the time series plot. This may reveal one or more of the following characteristics: seasonality, trends either in the mean level or the variance of the series, long-term cycles, and so on. Time series analysis is mainly concerned with decomposition of the variation of a series into trend, seasonal variation, cyclic and the remaining 'irregular' fluctuations (Chatfield, 1996).

- Trend: A trend is a long-term component that represents a growth or a decline of a time series over an extended period of time.

- Seasonal component: This term of seasonality is used for time series defined at time intervals which are fractions of a year. It is a pattern of change that repeats itself from year to year (example monthly rainfall).

- Cyclical component: Changes in time series sometimes show a wavelike fluctuation around a trend, which shows the possible existence of periodicity with longer time intervals.

- Irregular component: This is a part of a time series represented by residuals, after the above-mentioned components have been removed.

White noise: The fundamental building block of time series models is a white noise series a_t. The symbol a_t represents an anticipating in coming "shock" to the system. The assumption in time series modeling in regard to these a_t sequences is an uncorrelated sequence of random variables with constant variance and means zero. The goal of time series modeling is to capture, with the estimated model, the correlation structure in the series.

Autocorrelation: In a time series, autocorrelation refers to serial dependence measure, that is, the correlation of observations of one variable at one point in time with observations of the same variable at prior time points. This is a convenient point to introduce a concept of fundamental importance in the analysis of time series. It is the object of many forms of time series analysis: To identify the type of dependence which exists, to build a statistical model that emulates the dependence and proceed with forecasting and policy analysis. The array of coefficients ρ_1, ρ_1,... , ρ_k or their sample counterparts $r_1, r_2, \ldots, r_n$ tell us a great deal of the internal structure of the series. And their totality, graphed with k as abscissa and r_k as ordinate is called autocorrelation function plot.

The visual inspection of the autocorrelation function plot provides useful check for time series pattern (Chatfield, 1996).

- If a time series is completely random, then for large n, $r_k = 0$ for every k.

- The stationary time series have only short-term correlations.

- If successive values of a time series tend to alternate, the autocorrelation function would also tend to alternate.
- If a time series has a trend, the values of r_k would not decrease to zero, except for very large values of k.
- If a time series is characterized by seasonal fluctuations, then the autocorrelation function would also shows oscillations along the same period.

Stationarity: A time series is said to be stationary if there is no systematic change in mean (no trend), if there is no systematic change in variance and if periodic variations have been removed. If the statistical property of one section is much like those of any other section of the data then we call the series is stationary (Pankratiz, 1983). Most of the probability theory of time series are concerned with stationary time series, and for this reason time series analysis often requires one to turn a non-stationary series into a stationary one so as to use this theory (Brockwell and Davis, 1996).

There are two main types of transformations to make the series stationary. Firstly, the data can be transformed through differencing if the stochastic process has an unstable mean. This type of transformation is used for the purpose of removing the polynomial trend that is exhibited by the data. The logarithmic and square root transformations are special cases of the class of transformations called the Box-Cox transformation which can be used to induce stationarity. These transformations are used if the series being examined has a non-constant mean and variance (Box *et al.*, 1985). In time series analysis, the most commonly used transformations are variance–stabilizing transformations and differencing. Since differencing may create some negative values, variance–stabilizing transformations should be applied before taking differencing (Cromwell *et al.*, 1994)

Differencing: This method is an integral part of the procedure advocated by Box and Jenkins (1976) and is a data pre-processing step which attempts to de-trend data to control autocorrelation and to achieve stationary by subtracting each datum in a series from its predecessor. Differencing is a special type of filtering, which is particularly useful for removing a trend and seasonal fluctuation (Chatfield, 1996). A differenced stationary series is said to be

integrated and is denoted as I (d) where'd' is the order of integration. The order of integration is the number of unit roots contained in the series, or the number of differencing operations it takes to make the series stationary about the mean (Wei, 1990).

Parsimony: In Box-Jenkins time series modeling the term parsimony refers to a model with least number of parameters. This is not only because simple models are easy to fit and explain, but also because parsimonious models help avoid the problem of parameter redundancy (Pankratiz, 1983).The correlation matrix for estimated parameters provides a means of recognizing the existence of parameter redundancy. Although the estimates of the parameters of a Box-Jenkins model will always have some correlation, very high correlation ($|R| > 0.8$ or 0.9 between the estimates suggest parameter redundancy (Kirkpatrick and Gaynor, 1994).

CHAPTER THREE

DATA AND METHODOLOGY

3.1 Data

The National Meteorological Agency (NMA) is the responsible organization for the collection and issuing of meteorological data. The monthly rainfall series from the period January 1975 – December 2009 used in this study was collected by the agency at Mekele station of Tigray region.

3.2 Methodology

The methodology used for time series analysis in this research is the univariate Box and Jenkins method. This method applies Autoregressive and Moving average to formulate a model used in forecasting future rainfall series. This research exclusively deals with modeling time series forecasting model, in particular, the Seasonal Autoregressive Integrated Moving Average (SARIMA).

The Box and Jenkins methodology is a powerful approach to the solution of many forecasting problems Montgomery and Johnson (1976) and it can provide extremely accurate forecasts of time series and offers a formal structured approach to model building and analysis. There are many quantitative methods of model building and forecasting used in climatology and metrological studies today. With the development of the statistical software packages and its availability, these techniques have become easier, faster and more accurate to use. In this study, we employ SAS and SPSS software packages for the statistical data analysis.

3.2.1 The Box-Jenkins Models

In general, the Box-Jenkins models provide a common framework for time series forecasting. It emphasizes the importance of identifying an appropriate model in an interactive approach. In addition, the framework can cope with non-stationary series by the use of differencing. Box-Jenkins' method derives forecasts of a time series solely on the basis of the historical behavior of the series itself. It is a univariate method which means only one variable is to forecast. The ideas are based on statistical concepts and principles that are able to model a wide spectrum of time series behavior. The basic Box-Jenkins' models can be represented by a linear combination of past data and random variables. The sequence of random variables is called a white noise process. These random variables are assumed to be uncorrelated and normal with mean zero and constant variance.

The first step in developing a Box-Jenkins model is to determine if the time series is stationary, non-random and if there is any significant seasonality that needs to be modeled. Stationarity can be assessed from graphical and statistical tests. Specifically, non-stationarity is often indicated by an autocorrelation plot with very slow decay.

In order to understand the modeling procedure it is useful to briefly introduce the three basic models in non-seasonal and seasonal versions.

3.2.1.1 Non Seasonal Stochastic Models

3.2.1.1.1 Autoregressive (AR) models

Autoregressive models are the most popular time series models, as they can be fully estimated and tested within the framework of least-square regression. An autoregressive model with p AR parameters can be written as:

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \ldots + \phi_p x_{t-p} + a_t, \qquad (1)$$

where x_t is the time series, $\phi_1, \phi_2, \ldots, \phi_p$ are the parameters of the AR, a_t is usually specified as white noise. The subscripts on the ϕ's are the orders of the AR parameters. The highest order p is referred to as the order of the model. Hence, it is called an autoregressive model of order p and is usually abbreviated AR *(p)* (Box et al., 1985). It is convenient to write the model in lag operators,

$$x_t = \phi_1 B x_{t-1} + \phi_2 B\, x_{t-2} + \ldots + \phi_p B\, x_{t-p} + a_t,$$

$$(1 - \phi_1 B x_{t-1} - \phi_2 B\, x_{t-2} - \ldots - \phi_p B\,)x_t = a_t,$$

$$\phi(B)\, x_t = a_t. \qquad (2)$$

The values of ϕ which make the process stationary are such that the roots of $\phi(B) = 0$ should lie outside the unit circle in the complex plane where B is the backward shift operator such that $Bx_t = x_{t-1}$ (Vandaele, 1983). In other words, if all zeros of $\phi(B)$ are larger than one in absolute value, there is a stationary process x_t, which satisfies the autoregressive equation and can be represented as

$$x_t = \sum_{j=0}^{\infty} \psi_j a_{t-j}, \ \psi_j \text{ is the } j^{\text{th}} \textit{ wheigh of } a_t \textit{ at lag } j. \qquad (3)$$

The coefficients ψ_j converge to zero, such that $\sum_{j=0}^{\infty} |\psi_j| < \infty$. If some roots are exactly one in modulus, no stationary solution exists. A typical case is the random walk, defined as

$$x_t = x_{t-1} + a_t$$

$$(1\text{-}B)x_t = a_t. \qquad (4)$$

In this kind of situation, there is no stationary process that satisfies this equation. For the simple AR (1) model, it is to recover them, as $x_t = \phi x_{t-1} + a_t$ is immediately transformed in to

$$x_t = \sum_{j=0}^{\infty} \phi^j a_{t-j}. \qquad (5)$$

The AR (1) model admits a stationary solution –is 'stable' –whenever$|\phi| < 1$. For higher–order models, stability conditions on coefficients become increasingly complex (Diebold *et.al*, 2006).

Stationary autoregressive processes have an autocorrelation function (ACF) $\rho_j = \text{corr.}\,(x_t,\ x_{t-j})$ that converges to zero as $j \to \infty$ at a geometric rate. A plot of the ACF of a stationary AR (p) model would then show a mixture of damping sine and cosine patterns and exponential decays depending on the nature of its characteristic roots. Another characteristics

feature of AR (p) models is that the partial autocorrelation function defined as PACF (j) = corr. $(x_t, x_{t-j} | x_{t-1}, x_{t-2}, \ldots, x_{t-j+1})$ becomes exactly zero for values larger than p (Tsay, 2005).

3.2.1.1.2 Moving Average (MA) Models

A series x_t is said to follow a moving average process of order q, or simply MA (q) process if

$$x_t = a_t + \theta_1 a_{t-1} + \theta_2 a_{t-2} + \ldots + \theta_q a_{t-q}, \qquad (6)$$

where a_t is the white noise, $\theta_1, \theta_2, \ldots, \theta_q$ are the MA parameters. MA (p) models immediately define stationary, every MA process of finite order is stationary. In order to preserve a unique representation, usually the requirement is imposed that all zeros of $\theta(B) = 1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q$ are equal greater than in absolute value. In that case, the MA model corresponds to the wold's representation of the stationary process. Wold's theorem tells that every stationary process can be decomposed in to a deterministic part and purely stochastic part, with the stochastic part being an MA process. The errors of the wold's representation are defined as prediction errors, if x_t is predicted linearly from its past (Diebold *et.al*, 2006).

If all roots of $\theta(B) = 0$ lie outside the unit circle, the MA process has an autoregressive representation of generally infinite order $\sum_{j=0}^{\infty} \psi_j x_{t-j} = a_t$ with $\sum_{j=0}^{\infty} |\psi_j| < \infty$. MA process with an infinite order autoregressive representation are said to be invertible. A property required on occasion in the analysis of such time series is that of invertability.

A characteristics feature of MA (q) is that their ACF, ρ_j becomes statistically insignificant after $j=q$. The property of the ACF should be reflected in the correlogram, which should 'cut off' after q and the PACF converges to zero geometrically.

3.2.1.1.3 Autoregressive–Moving average (ARMA)

In most cases, it is best to develop a mixed autoregressive moving average model when building a stochastic model to represent a stationary time series. The order of an ARMA model is expressed in terms of both p and q. The model parameters relate to what happens in period t to both the past values and the random errors that occurred in past time periods. The general ARMA model can be written as follow:

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \cdots + \phi_p x_{t-p} + a_t + \theta_1 a_{t-1} + \ldots + \theta_q a_{t-q} \quad (7)$$

This expression can be simplified by a backward shift operator B and we obtain

$$\phi(B)\, x_t = \theta(B)\, a_t. \quad (8)$$

An ARMA model is stable–i.e., it has a stationary 'solution'–if all zeros of $\phi(B)= 0$ is larger than one in absolute value. The representation is unique if all zeros of $\phi(B)= 0$ lie outside the unit circle and $\phi(B)$ and $\theta(B)$ not have common zeros. The stable ARMA models always have an infinite order MA representation. If all zeros of $\phi(B)$ is larger than one in absolute value, it has an infinite order AR representation. In practice, ARMA models often permit to represent an observed time series with a lesser number of parameters (more parsimonious) than AR or MA models. To have ARMA (p, q) model, both ACF and PACF should show decaying to zero. The autocorrelogram of an ARMA (p, q) process is determined at greater lags by the AR (p) part of the process as the effect of the MA part dies out. Thus eventually the ACF consists of mixed damped exponentials and sine terms. Similarly, the partial autocorrelogram of an ARMA (p, q) process is determined at greater lags by the MA (q) part of the process (Caldwell, 2006). Thus eventually the partial autocorrelation function will also consist of a mixture of damped exponentials and sine waves.

3.2.1.1.4 Autoregressive Integrated Moving Averages (ARIMA) Models

This section introduces the basic theory of Autoregressive Integrated Moving Average (ARIMA). Generally, most of the time series are non-stationary and Box-Jenkins' method recommends the user to remove any non-stationary sources of variation, and then fit a stationary model to the time series data. In practice, we can achieve stationarity by applying regular differences to the original time series. If differencing a series d times makes it into a stationary ARMA(p, q) and the series is said to be an autoregressive integrated moving average process, denoted by ARIMA(p, d, q) and may be written as

$$\phi(B)(1 - B)^d x_t = \theta(B)a_t, \qquad (9)$$

where $\phi(B)$ is a polynomial of order p, $\theta(B)$ of order q and ϕ and θ obey the relevant stationarity and invertibility conditions, respectively. In this expression the left-hand side has a unit root in the operator $\phi(B)\ (1 - B)^d$. Testing for stationarity is the same as looking for, and not finding, unit roots in this representation of the series. If the presence of a unit root is not obvious it may become obvious from an examination of the sample autocorrelogram and indeed this tool was used for many years to indicate their presence (Hoff, 1983). In recent years Dickey-Fuller tests have been designed to test for a unit root in these circumstances (see page 28 for more details of this test).

ARIMA models are written in the same way as the basic models, except that the differenced (stationary) series w_t is substituted for the original series x_t. In order to express ARIMA models, we have to understand the use of difference operator. For example, the first difference of a series can be expressed as:

$$w_t = x_t - x_{t-1} \qquad (10)$$

The use of a symbol ∇ is used to simplify the expression in equation (10). The first differences of the series x_t could then be written as:

$$w_t = \nabla x_t \qquad (11)$$

If we take the second consecutive difference of the original series the expression would be defined as:

$$w_t = \nabla^2 x_t = \nabla(x_t - x_{t-1}) = \nabla(\nabla)x_t \quad (12)$$

In general, the d^{th} consecutive differencing would be expressed as $\nabla^d x_t$(Vandaele, 1983).

The general form of ARIMA model can written as

$$w_t = \phi_1 w_{t-1} + \phi_2 w_{t-2} + \ldots + \phi_p w_{t-p} + a_t + \theta_1 a_{t-1} + \theta_2 a_{t-2} + \ldots + \theta_q a_{t-q} \quad (13)$$

In ARIMA models, the term integrated, which a synonym is for summed, is used because the differencing process can be reversed to obtain the original time series values by summing the successive values of the differenced series (Hoff, 1983).

3.2.1.1.5 Interpretation of Stationarity and Invertibility of AR and MA Models

A stochastic process is strictly stationary if its probability distribution is invariant to shifts in time, and is weak stationary when the first two moments of the probability distribution of the process are invariant to time shift. In practice, weak stationarity, or simply stationarity, is widely used that represents a critical assumption in the analysis time series data (Diebold *et al.*, 2006). Its importance lies in the fact that the conditions of constant mean, variance and covariance are essential in accurately estimating the parameters and models that describe the data. Given that in most situations only one observation is available at a given time, stationarity ensures that all parts of the series are like the other parts, which allows us to estimate the needed parameters. Therefore, the mean, the variance and the covariance of the series are not functions of time and depend rather on the lag between the observations.

Invertibility condition also the other critical assumption in time series modeling. In this concept, effects of past values on the current value die down the further in to the past.

The ARMA (p, q) model for the x_t series may be defined by the relationship,

$$\left[\sum_{j=0}^{p} \phi_j B^j x_t\right] = \left[\sum_{j=0}^{q} \theta_j B^j a_t\right] \text{ or}$$

$$\phi(B)\, x_t = \theta(B) a_t\,, \qquad (14)$$

where a_t is a white noise, ϕ_j and θ_j are the AR and M A parameters, respectively, B is the back ward shift operator, and p and q are the orders of the polynomials $\phi(B)$ and $\theta(B)$, respectively. Taking the expectation on both sides and remembering that the white noise series a_t has mean zero, it follows that $E(x_t) = 0$. The variance of the ARMA (p, q) model can be represented in terms of its spectrum (Dulleur and Kavas,1978).

$$Var(x_t) = \sigma_\alpha^2 \int_{\frac{-1}{2}\delta}^{\frac{1}{2}\delta} \frac{\left|\sum_{\alpha=0}^{p} \theta_\alpha e^{-i2\Pi\omega\delta_\alpha}\right|^2}{\left|\sum_{\beta=0}^{q} \phi_\beta e^{-i2\Pi\omega\delta_\alpha}\right|^2} d_\omega\,, \qquad (15)$$

Where δ is the sampling time interval, ω the frequency and σ_α^2 the variance of the random shocks series a_t. The polynomial in the numerator converges when θ_α, $\alpha = 0, 1, 2, \ldots p$ are finite. The denominator can be expressed as the product of factors $\left(1 - c_j e^{-i2\Pi\omega\delta}\right)^{m_j}$, where m_j is the multiplicity of the j[th] factor. The variance given by (15) will thus converge when $|c_j| < 1$. If the roots of the denominator polynomial are of the form $1/c_j$, j=1, 2..., p they should be lie outside the unit circle. Therefore, stationarity imposes a restriction on the autoregressive coefficient only, namely, that the roots of $\phi(B) = 0$ be outside the unit circle.

We consider the ARMA (1, 0) model $(1 - c_j B)\, x_t = a_t$. Then, for

$$x_t = c_j x_{t-1} + a_t$$

$$x_{t+1} = c_j^2 x_{t-1} + c_j a_t + a_{t+1}$$

$$x_{t+2} = c_j^3 x_{t-1} + c_j^2 a_{t+1} + c_j a_{t+1} + a_{t+2}, \text{ etc.} \qquad (16)$$

When $|c_j| > 1$, the effect of the past on the present value of the time series increases as the series moves in to the future. When $|c_j| = 1$ the effect of the past on the present value stays the same no matter how far in the future the series have moved.

Although there were no restriction on moving average operator $\theta(B)$ of the ARMA(p, q) model under the stationarity conditions , the invertibility of the $\theta(B)$ is required to assure realizability of time series values in stochastic modeling.

The polynomial $\theta(B)$ may be written as the product of factors $(1 - c_j B)^{m_j}$, where m_j is the multiplicity of the jth root in $\theta(B)$ .The series is said to be invertible if When $|c_j| < 1$, j=1, 2,..., q. If we consider the case when $|c_l| > 1$ then $\theta(B) = (1 - c_l B)\theta_1(B)$ and the ARMA (p,q) model, $\theta(B)x_t = (1 - c_l B)\theta_1(B)$ can be defined as

$$\frac{\theta_1(B)}{\phi(B)} a_t = \frac{1}{1-c_l B} x_t = -\frac{1}{c_l} x_{t+1} - \frac{1}{c_l^2} x_{t+2} - \frac{1}{c_l^3} x_{t+3} - \cdots \qquad (17)$$

Thus for a root of $\theta(B)$ inside the unit circle the future values are used to generate the present values of x_t.Therefore, to have a realizability of the series generation all roots of $\theta(B) = 0$ must be outside the unit circle.

3.2.1.2 Pure Seasonal models

In addition to trend, many time series may contain seasonal periodic component which repeats at every s regular seasonal intervals. The seasonal pattern may additionally frequently display constant change over time. Just as regular differencing applied to the overall trending series, seasonal differencing (SD) is applied to seasonal non stationary as well. Seasonal differencing on series x_t can be defined as: $w_t = (1 - B^S)\, x_t = x_t - x_{t-S}$, where B is the backshift operator and s is the seasonal interval of the series. And as autoregressive and moving average tools are available with the overall series, so too, available for seasonal phenomena using seasonal autoregressive parameters (SAR) and seasonal moving average parameters (SMA). The need for seasonal autoregressive (SAR) and seasonal moving average (SMA) parameters is established by

examining the autocorrelation and partial autocorrelation patterns of a stationary series at lags that are multiples of the number of periods per season. These parameters are required if the values at lags s, $2s$, etc. are nonzero and display patterns associated with the theoretical patterns for such models. Seasonal differencing is indicated if the autocorrelations at the seasonal lags do not decrease rapidly.

In the seasonal box and Jenkins methodology, as in the non seasonal models, requires the similar implementation of the iterative procedure to determine the best model. However, because there is seasonal component present, as well as trend and error, the steps are slightly more involved, the models more sophisticated and the alternatives more numerous. In practice, mixed seasonal models are seldom used (Bower man and O'Connell, 1987), yet the general form of the box-Jenkins model can be used effectively to write the equation for any specific non seasonal and seasonal stationary model.

3.2.2 Handling Outliers and Missing Values

For anomalous data due to various causes, the incorporation of outlier detection and adjustment procedure can ultimately produce more appropriate models, better parameter estimates, and more accurate forecasts.

Moreover, similar to other statistical analyses a problem frequently encountered in data collection is missing observations in a data series. Missing data must also be addressed in the time series context. Special consideration in handling missing data in a time series application is that the missing data cannot simply be omitted from the data series. Missing observations must be replaced by appropriately estimated values so that the alignment of data between time periods will not be offset inappropriately. In order to replace those observations, there are several options available in the literature. Kirkpatrick and Gaynor (1994) discussed that missing data in a time series may be estimated using one of the following methods. Firstly, replace with the mean of the series .This mean can be calculated over the entire range of the sample. Secondly, replace with the naïve forecast. Naive model is the simplest form of a univariate forecast model. It uses the current time value for the next time, that is $\hat{x}_{t+1} = x_t$. Thirdly, replace with a simple trend forecast. This is accomplished by estimating the regression equation of the form $x_t = \alpha + \beta t$, where t is the time for the periods prior to the missing value. Then use the equation to fit the time

periods missing. Finally, replace with an average of the last two or more known observations that bound the missing observation.

3.2.3 Tests for Assumptions in Time Series Analysis

The first step in developing a Box-Jenkins model is to determine if the series is stationary and if there is any chance of randomness. The underlying methodology assumes that the time series is stationary and serially correlated. Thus, before modeling process, it is important to check whether the data under study meets these assumptions.

3.2.3.1 Tests for Stationarity

Stationarity is the first fundamental statistical property tested in time series analysis. If non stationarity is present in a given time series, it is possible to transform the series to a stationary series. Because of most time series data are non-stationary, transformation is needed in stochastically modeling. In this sense, the most common transformation is differencing, that is, subtracting a past value of a variable from its current value (Greene, 2000). But, it is necessary to detect whether non-stationarity is present in a series before differencing. For this reason, there are alternative approaches as graphical method, nonparametric tests and unit root test.

3.2.3.1.1 Visual Tests for Stationarity

The Time Order Plot: The first and simplest type of test one can apply to check for stationarity is to actually plot the time series and look for evidence of trend in mean, variance, and seasonality. If any such patterns are present then these are signs of non-stationarity and different mechanisms exist to turn the series into a stationary one. If the result is a series that does not resemble a roughly flat, horizontal line then a trend in mean is present in the initial series. Statistically, the presence of a trend in the mean can further be detected by applying a seasonal Kendal test to the time series Wei (1990).

The Correlogram Test: One way to characterize a series with respect to its dependence over time is to plot its sample ACF and PACF. Both functions are used in Box-Jenkins modelling as correlograms to reveal important information regarding the order of the autoregressive and moving average factors present in the generating process of the given time series as well as to assess stationarity. Enders (2004) expresses that inspection of ACF serves as a rough indicator of whether non-stationarity is present in a series. If the sample ACF decays very slowly, it indicates that differencing is needed. This inspection of ACF implies that the sample is non stationary.

While the stationarity tests described in the above sections make use of subjective visual inspection of data plots and correlograms, a more recent series of tests were developed to help with determining stationarity. These tests are unit root tests and stationarity tests formulated for the most part on formal statistical tests. And the difference between them lies in the stringency of the assumptions they use as well as in the form of the null and alternative hypotheses they adopt. The standard Dickey-Fuller test (DF) is based on independently distributed errors and has as the null hypothesis the unit root. On the other hand, the Phillips-Perron test is nonparametric and allows for some heterogeneity and serial correlation in the innovations (Enders, 2004). There exist many other unit root and stationarity tests as well as generalizations and combinations of the ones mentioned above. However, in this study, Augmented Dickey Fuller (ADF), the extension of the standard DF test is to be used.

3.2.3.1.2 Unit Root Tests

For a univariate time series, the Unit Root test is frequently employed for testing stationarity. The test first poses the *null hypothesis* that the given time series has a unit root, which means that the time series is non-stationary, and tests if the null hypothesis is to be statistically accepted or rejected in favor of the *alternative hypothesis* that the given time series is stationary. Let us assume that the relationship x_t (the value at time t) and x_{t-1} is given by:

$$x_t = \phi x_{t-1} + a_t, \qquad (18)$$

where a_t is a white noise process and ϕ is a parameter coefficient to be estimated. This model is a first order autoregressive process. The time series x_t converges, as $t \rightarrow \infty$, to a stationary time series if $|\phi| < 1$. If $|\phi| = 1$ or $\phi > 1$, the series x_t is not stationary and the variance of x_t is time dependent (Diebold *et al*, 2006). In other words, the series has a unit root. The Unit Root test based on standard DF subsequently tests the following one-sided hypothesis:

H_0: $\phi = 1$ (has a unit root) versus

H1: $\phi < 1$ (has root outside the unit circle)

The name, unit root, comes from the fact that the coefficient of x_{t-1} is unity in this context. Greene (2000) stated that a non stationary time series could be converted to a stationary time series by taking first or higher order difference. If x_{t-1} is subtracted from the right and left sides of the above equation, resulting equation becomes:

$$\nabla x_t = (\phi - 1) x_{t-1} + a_t \qquad (19)$$

This equation is expressed as a first order difference equation. If ϕ is taken one in the equation, the effect of unit root can be removed from the actual series that has non-stationarity via a first differencing. Note that the autoregressive model in Equation (18) is a simple form. The most common unit root test is the one that employs an autoregressive model containing more than one autoregressive terms (or the Augmented Dickey-Fuller test).

The Augmented Dickey-Fuller (ADF) Test: The stationarity test utilized in this study is the Augmented Dickey-Fuller (ADF) technique which is a general auto-regression model formulated in the following regression equation with trend and drift (Enders, 2004):

$$\nabla x_t = \mu + \beta t + \phi x_{t-1} + \gamma_1 \nabla x_{t-1} + \ldots + \gamma_p \nabla x_{t-p} + a_t, \qquad (20)$$

Where ∇x_t is the first differenced value of x_t, a_t is the error term

∇x_{t-j} is the j^{th} lagged of the first differenced of values of x_t

$\mu, \beta, \phi, \gamma_1, \gamma_2, \ldots, \gamma_p$ are parameters to be estimated .

The Dickey-Fuller-test now estimates $\phi^* = \phi - 1$ *by* $\widehat{\phi^*}$, obtained from an ordinary regression and checks for $\phi^* = 0$ by computing the test statistic:

$$DF_\tau = \frac{\phi^*}{S.e(\phi^*)}$$

where n is the number of observations on which the regression is based. The test statistic follows the so called Dickey-Fuller distribution which cannot be explicitly given but has to be obtained by Monte-Carlo and bootstrap methods (Diebold *et al.,* 2006).

The model hypothesis of interest is:

H_0: $\phi^* = 0$ - the data needs to be differenced to make it stationary, versus the alternative hypothesis

H_A: $\phi^* < 0$ - the data is stationary and doesn't need to be differenced

If the null hypothesis of unit root (non stationary variable) is rejected indicating the variable is stationary and integrated of degree zero. On the hand if the series found to non stationary, a transformation of the variable by differencing is needed until we achieve stationarity (Falk, 2006). In time series models, a linear stochastic process has a unit root if 1 is a root of the process's characteristic equation. And the process will be non-stationary. If the other roots of the characteristic equation lie inside the unit circle, then the first difference of the process will be stationary. The augmented Dickey–Fuller (ADF) statistic, used in the test, is a negative number. The more negative it is, the stronger the rejections of the hypothesis that there is a unit root at some level of confidence (Elliott *et.al.,* 1996). Note that data without a clearly recognizable trend, the term $\mu + \beta t$ in Eq. (20) can be omitted. In order to test for second-order differencing, the test procedure can be repeated for ∇x_t instead of x_t, with $\nabla^2 x_t$ replacing ∇x (Hamilton, 1994).

3.2.3.1.3 Variance Comparisons Test

The behavior of variance associated with different orders of differencing can provide a useful means of deciding the appropriate order of differencing to achieve stationary (Hamilton, 1994). A time series that is non-stationary in mean can be made stationary by the first differencing. But, if the series is also not stationary in the rate of change of the mean (i.e. slope), stationarity can be achieved by taking the second difference, or the first difference of the first difference. It should, however, be borne in mind that each successive differencing will decrease the variance of the series, but at some point, higher-order differencing will have an opposite effect. When variance increases, it means that the series has been over-differenced (Vandaele, 1983).

3.4.2 Tests for Randomness

3.4.2.1 Visual Inspection

A time series, in which the observations fluctuate around a constant mean, have a constant variance and are statistically independent, is a random time series (Chatfield, 1996). In other words, the time series plot does not exhibit any pattern:

- The observations do not trend upwards or downwards,
- The observations do not tend to be larger in some periods than in other periods.
- The variance does not increase or decrease over time,

We can examine whether a time series is random or not by the following procedures:

- Visually, whether the time series plot shows any trend or not.
- Visually by looking at the sample autocorrelation function of the time series.
- Statistically, testing whether the observed series could have been generated by a random stochastic process.

3.4.2.2 Bartlett's Band Test

This test based on individual comparison of the sample autocorrelation and partial autocorrelations to their approximate standard error, $\pm\frac{2}{\sqrt{n}}$. This test was developed by Bartlett (1946), who showed that if a series is generated by a white noise process (cited in Hamilton (1994)). The estimators are approximately normally distributed random variables with mean zero and variance *1/n*, where *n* is the total number of observations. To make a 95% confidence test of the null hypothesis of no autocorrelation or partial autocorrelation lie out sided the band at lag *k*, we simply need to compare the value of the sample coefficient with the critical values $\pm\frac{2}{\sqrt{n}}$. If the value falls outside the bands, the null hypothesis is rejected at the 95% level. Thus, the Bartlett bands can be used as a benchmark for assessing departures from randomness, by providing bounds for the ACF and PACF taken one at a time.

3.4.2.3 Ljung-Box Q Statistic

The Ljung-Box Q or Q statistic can be employed to check independence instead of visual inspection of the sample autocorrelations. The Bartlett band tests the null hypothesis that an individual autocorrelation coefficient is equal to zero. However, if a stochastic process is white noise, then *all* the autocorrelation coefficients up to a specified lags will be zero. Thus, to test for the overall and simultaneous absence of autocorrelation at multiple lags, we need a test of the joint null hypothesis. A test of this hypothesis can be done for serial dependence by choosing a level of significance and then comparing the value of calculated Q with χ^2- table of critical value. The Q-test statistics has an asymptotic x_j^2 distribution (Brockwell and Davis 1996). If the calculated Q value is smaller than the χ^2-table critical value, there is no serial dependence on the basis of available data. The Q statistic is constructed for the first *J* autocorrelations and calculated by using:

$$Q = n\,(n+2)\sum_{j=1}^{J}\frac{r_j^2}{n-j} \qquad (22)$$

where n is the number of available observations, r_j is the j^{th} sample autocorrelation for j=1,2,3,...,J.

3.5 SARIMA Model

In representing x_1, x_2, x_3, ... , x_{t-1}, x_t, $x_{t+1,...,}x_n$ for monthly rainfall time series observed at equal time intervals (month); a general SARIMA $(p, d, q) \times (P, D, Q)_S$ model for x_t is written as:

$$\phi_p(B)\,\boldsymbol{\Phi}_P(B^S)\,(1-B)^d(1-B^S)^D\,x_t^{(\lambda)}) = \mu + \theta_q(B)\boldsymbol{\Theta}_Q(B^S)\,a_t \qquad (23)$$

or

$$\phi_p(B)_P(B^S)\,w_t = \mu + \theta_q(B)_Q(B^S)\,a_t \qquad (24)$$

where $x_t^{(\lambda)}$ refers to some appropriate transformations of x_t such as: a Box-Cox, log, and square root transformations;

t discrete time;

S seasonality interval, equal to 12 for monthly rainfall data;

B backward shift operator defined by $Bx_t^{(\lambda)} = x_{t-1}^{(\lambda)}$ and $B^S x_t^{(\lambda)} = x_{t-S}^{(\lambda)}$;

μ mean level of the process, usually taken as the average of the series w_t (if $D + d > 0$ often $\mu \simeq 0$);

a_t normally independent distributed white noise residual with mean 0 and variance σ_a^2;

$\phi_P(B) = 1 - \phi_1 B - \phi_2 B^2 - \phi_3 B^3 - \ldots - \phi_P B^P$ non seasonal autoregressive (AR) operator or polynomial of order p such that the roots of the characteristic equation $(B) = 0$ lie outside the unit circle for non seasonal stationarity;

$(1-B)^d = \nabla^d$ non seasonal differencing operator of order d to produce non seasonal stationarity of the d^{th} differences ,usually d=0,1,or 2 ;

$\Phi_P(B^S) = 1 - \boldsymbol{\Phi}_1 B^S - \boldsymbol{\Phi}_2 B^{2S} - \boldsymbol{\Phi}_3 B^{3S}, \ldots, - \boldsymbol{\Phi}_P B^{PS}$ seasonal AR operator of order p such that the roots of $\boldsymbol{\Phi}(B^S) = 0$ lie outside the unit circle for seasonal stationarity and the $\boldsymbol{\Phi}_{is}$, i =1,2,3,...,p are the seasonal AR parameters;

$(1-B^S)^D = \nabla_S^D$ seasonal differencing operator of order D to produce seasonal stationarity of the D^{th} differenced data;

$w_t = \nabla^d \nabla_S^D x_t^{(\lambda)}$ stationary series formed by differencing $x_t^{(\lambda)}$ series ($n'=n - d - sD$ is the number of terms in the w_t series);

$\theta_q(B)=1 + \theta_1 B + \theta_2 B^2 + \theta_3 B^3 +...., + \theta_q B^q$ non seasonal moving average MA operator or polynomial of order q such that the roots of θ (B)=lie outside the unit circle for invertibility and θ_i, i=1,2,3,...q are the non seasonal MA parameters;

$\Theta_Q(B^S)=1+\Theta_1 B^S+\Theta_2 B^{2S}+ ...., + \Theta_Q B^{QS}$ seasonal MA operator of order Q such that the roots of $\Theta(B^S)=0$ lie outside the unit circle for invertibility and Θ_i, i=1,2,3,..,Q are the seasonal MA parameters (Hipel *et. al.*, 1977) .

3.5.1 Seasonal Model Development

We often need to analyze time series in which a recurrent pattern having period s occurs. The pattern may change with time. In our case, monthly rain fall series exhibits (as expected), a marked yearly seasonal pattern with s=12.The general model (25) is entirely adequate to accommodate for such seasonal behavior. However, with such behavior there are known special features of the situation which can be exploited to give the model a parsimonious form (Box.*et.al*, 1985). In monthly rainfall series suppose that a value x_t *i*s observed in the month of January. Due to the nature of yearly seasonality, we expect that readings in month January of previous years will show similarities with this January. Consequently, we can link the January values for different years by a model of the form

$$\Phi_P (B^{12}) \nabla_{12}^D x_t = \Theta_Q (B^{12}) \alpha_t \qquad (25)$$

and explains the dependence of the observation x_t at a particular month to the observation taking the same month during previous years. In Eq. (25) the symbol ∇_{12}^D means $(1-B^{12})^D$, the symbol $\Phi_P (B^{12})$ is the seasonal autoregressive operator of order P, and $\Theta_Q (B^{12})$ is the seasonal moving average operator of order Q. There would be one such model for each month.

Although the observation of the monthly rainfall for March, for example, is related to previous March rainfall, it is also related to other monthly rainfalls during the same year. To take care of this serial dependence in the α_t series, Box and Jenkins (1976) introduced the model

$$\phi_p(B)\nabla^d \alpha_t = \theta_q(B)a_t \qquad (26)$$

where $\nabla^d = (1-B)^d$ and a_t is a white noise. Assuming that the parameters Φ and Θ obtained for each month are approximately the same, by combining (25) and (26) arrive at the general multiplicative model

$$\phi_p(B)\,\Phi_P(B^{12})\nabla^d\nabla_{12}^D x_t = \theta_q(B)\Theta_Q(B^{12})\,a_t \qquad (27)$$

This multiplicative model assumes that in addition to the year to year serial dependence of the multiplicative ARIMA model, there is a correlation structure with in the months of the same year (Dulleur and Kavas, 1978). When s=12, described in (Box *et.al.*,1985) the model embodies parameters which describe month-to- month variation and parameters which describe year - to - year variation.

The optimal forecast of future values of a time series are determined by the nature of the stochastic models which describes that series. The main effort in statistical analysis directed to forecasting must be in obtaining a suitable stochastic model for the series. An important principle in the choice of such models is that they should, whilst adequately representing the data, contain as few parameters as possible. In practice such mode development must be done iteratively using process of identification, estimation and diagnostic checks.

3.5.1.1 Model Identification

The purpose of the identification stage is to determine the differencing and transformation required achieving stationarity, and also the order of both the seasonal and the non seasonal AR and MA operators for the w_t series. The Univariate Box-Jenkins model requires stationary series. The stationary series have the following characteristics: (1) their mean, variance, and autocorrelation coefficients are essentially constant through time; (2) the autocorrelations of the series in ACF (autocorrelation-function) converge rapidly toward zero. However, in practice most series are not stationary, requiring a transformation to be converted into stationary series (Box and Jenkins, 1976). Having determined the correct order of differencing required rendering the series stationary, the next step is to find an appropriate model to the stationary series. There are a number of identification methods proposed in the literature. The autocorrelations function (ACF) and the partial autocorrelation functions (PACF) are the two most useful tools in any attempt at time series model identification (Granger and Newbold, 1986). Examination of the autocorrelation function of some suitable difference of the form $w_t=\nabla^d \nabla_s^D x_t$ of the series will usually indicate which model to fit in the first instance. This model can then be fitted, checked and modified if the fit is not adequate .This type of analysis has been also used with success to model a variety of seasonal series (Box *et.al*, 1985).

A plot of original data portrays an overview of how the time series is generally behaving. However, the sample Autocorrelation Function and Partial Autocorrelation Function transform the given information into a format whereby it is possible to detect the order of Auto Regressive, Moving Average, Seasonal Auto Regressive and Seasonal Moving Average terms required in the tentative model to be identified.

*Autocorrelation Function (*ACF*)*: The sample ACF r_k measures the amount of linear dependence between observations in a time series that are separated by a lag k. To use the ACF in model identification, estimate r_k using Eq. (29) (see next page) and then plot r_k series against lag k up to a maximum lag of about five times the seasonality interval and this should be less than to one fourth of the series under study (Hipel *et al.*, 1977). Relying on the plotted autocorrelation function we examine the presence of non stationarity in the x_t series. In case

where the series seem to be non-stationary non-seasonal and seasonal differencing should be applied and check the theoretical pattern of the corresponding ACF if the times of differencing produces optimum series stationarity.

To identify the number of non seasonal and seasonal autoregressive, and non seasonal and seasonal moving average parameters, we examine the ACF based on the theoretical pattern for the identified parameters. When the process is a pure SARIMA *(0, d, q)* × *(0, D, Q)* $_s$ model, r_k truncates and is not significantly different from zero after lag $q+sQ$. If r_k spikes out at lags that are multiples of s, this implies the presence of a seasonal autoregressive component. The failure of the autocorrelation function to truncate at other lags may imply that a non seasonal autoregressive term is required.

The autocorrelation of order k is simply the correlation between w_t and w_{t-k}, i.e.

$$\rho_k = \frac{E\{(w_t-\mu)(w_{t-k}-\mu)\}}{E\{(w_t-\mu)^2\}} \qquad (28)$$

In practice, one never knows the true autocorrelations and partial autocorrelations and at the identification stage, one has to rely on the sample autocorrelation and partial autocorrelation functions imitating the behavior of the corresponding parent quantities. In doing so, ρ_k can be estimated by

$$r_k = \frac{1/n \sum_{t=k+1}^{n}(w_t-\bar{w})(w_{t-k}-\bar{w})}{1/n \sum_{t=1}^{n}(w_t-\bar{w})^2} \quad , \qquad (29)$$

where $\bar{w}$ is the sample mean of the w_t.

Partial Autocorrelation Function (PACF): The partial autocorrelation function, denoted by PACF, is similar to the ACF and can be described as the correlation between x_t and x_{t-s} (observations of the time series recorded at two moments in time s time units apart) after controlling for the common linear effects of the intermediate lags. Following the general rules that may be helpful for interpreting the PACF of the stationary series, partial autocorrelation

function can also be used for determining the possible order of seasonal autoregressive, non-seasonal autoregressive, moving average and seasonal moving average that should be in corporate in the model. When the process is a pure SARIMA $(p, d, 0) \times (P, D, 0)$ model, r_{kk} cuts off and is not significantly different from zero after lag $p+SP$. If r_{kk} damps out at lags that are multiples of s, this suggests the incorporation of a seasonal moving average component in to the model. The failure of the partial autocorrelation function to truncate at other lags may imply that a non seasonal MA term is required (Hipel *et al.*, 1977).

To obtain an estimate r_{kk} for partial autocorrelations (ρ_{kk}) at lag k, we can employ successesive autoregressive estimation procedure. The first step is to model the w_t series by finite autoregressive models of order K given by

$$w_t = \rho_0 + \sum_{k=1}^{K} \rho_{kk} w_{t-k} \qquad (30)$$

where ρ_{kk} is the k^{th} autoregressive coefficients, w_t represents for the stationary monthly rainfall series and k=1,2,..,K.

Estimate of these coefficients by ordinary least squares or maximum likelihood estimation method gives the k^{th} sample partial autocorrelation (Hipel *et al.*, 1977).

3.5.1.2 Model Parameters Estimation

Model parameter estimation is usually carried out on a computer using a nonlinear least-squares approach by minimizing the sum of squared errors. The method of least-squares criterion is suggested by Box and Jenkins(1976) in which least squares refers to the parameter estimates associated with the smallest sum of squared residuals (SSE). The estimation-stage results will be used to check: (i) Parameter estimates, (ii) the appropriateness of coefficient estimates which includes the statistical significance of estimated coefficient and standard error, correlation matrix and coefficient near redundancy, and (iii) closeness of fit-root-mean squared error, AIC and r^2 = 1-SSE/SST refers to the coefficient of determination.

Maximum likelihood estimation for the model parameters, Brockwell and Davis (1996) suggest that the approximate maximum likelihood estimates for the SARIMA model parameters is obtained by employing the unconditional sum of squares method. When using this technique the unconditional sum of squares function is minimized to get least squares parameter estimates.

In Box and Jenkins modeling, the residual a_t are assumed to be independent, have constant variance, and usually normally distributed. The independence assumption is the most important of all and its violation can cause drastic consequences. However, if the constant variance and normality assumption are not true, they are often reasonable well satisfied when the observations x_t are transformed by logarithmic transformation. On the other hand, the normality assumption of the residuals is usually not critical for obtaining good parameter estimates as long as the a_t are independent and have a positive finite variance. We can obtain reasonable estimates called Gaussian estimates of the parameter can be obtained (Hipel *et al.*, 1977).

In maximum likelihood methods the likelihood function is maximized in order to obtain the parameter estimates. The likelihood of a set of data is the probability of obtaining that particular set of data, given its distribution. The philosophy behind maximum likelihood estimates is to find a set o f parameters which maximize the likelihood of observing the data to which the model is being fitted. This maximum likelihood estimates have desirable property that they give unbiased minimum variance estimates asymptotically. In time series analysis, the errors are generally assumed to follow a normal distribution. The likelihood function is created using all the observation available up to a particular time. After this a linear optimization algorithm is used to maximize the likelihood function with respect to the parameter space (Shumway and Stoffer, 2000).

Based on the treatment of initial values, estimation methods are termed as conditional or unconditional. If the initial errors are fixed as zero and initial observations are also regarded as true value, then the estimation is sad to be conditional maximum likelihood. If no assumption is made about the initial observations then estimation is unconditional or exact maximum likelihood. In general, unconditional estimates have lower error variance than the corresponding conditional estimates. However, the downside with unconditional estimation is that it is

computationally more demanding because initial values are not fixed and need to be determined along with the parameters.

3.5.1.3 Model Diagnostic Checks

At this stage the estimated model is tested to determine if it is statistically adequate. If the model proves to be inadequate then the identification stage is revisited to tentatively select one or more other models. In addition, diagnostic checking provides clues about how an inadequate model might be reformulated. A number of diagnostic checks on the adequacy of representation of the estimated model can be applied.

One class of diagnostic checks is devised to test model adequately by overfitting. This test is perhaps the simplest approach to fit a model which is rather more general (in the sense of containing more parameters) than that which has been identified. An examination of the statistical significance of the estimated of the additional parameters will indicate whether or not they should be included in the model. The need to use as few model parameter as possible (i.e., the model should be parsimonious) so formulation that model passes all the diagnostic checks .The Akaike Information Criterion(AIC) is a mathematical form of the parsimonious criterion of model building. When there are several competing models to choose from, select the model that gives the minimum of the AIC defined by

$$\text{AIC}=-2\ln(\text{maximum likelihood})+2m \qquad (31)$$

where m is the number of seasonal and non-seasonal autoregressive and moving average parameters to estimate.

Most diagnostic tests deal with the residual assumptions in order to determine whether the a_t are independent, have a constant variance, and are normally distributed. Residual estimates are needed for the tests used in checking the three aforementioned residual assumptions.

I. Tests for white noise: The most important test of the statistical adequacy of an estimated model also involves the assumption that the residuals are independent, or are not autocorrelated. The basic analytical tools to test this assumption are the residual ACF and residual PACF. The residual ACF for a properly built model will ideally have autocorrelation coefficients that are all near zero (Hipel *et al.*, 1977). If some of the residual autocorrelation coefficients are significantly different from zero; this may an indication that the present model is inadequate. The residual autocorrelation function to examine are the residual autocorrelation coefficients at the first few lags for non-seasonal model and the residual autocorrelation coefficients at the first couple of lags and also at lags that are multiples of s for seasonal model. If the present model is insufficient, a proper model can be selected either by changing the model Box and Jenkins (1976) or by repeating the identification, estimation and diagnostic checks stages of model construction.

Another method that can be used to evaluate the adequacy of a model is the plot of the errors over time. If a visual inspection of the errors reveal that they are randomly distributed over time, then we have a good model. The basic analytical tool test over this assumption is the residual ACF. The residual ACF for a properly built model will ideally have autocorrelation coefficients that are all near zero. Statistically, a t-test indicates if coefficients are significantly different from zero. In practice, if the absolute value of a residual ACF t-value is less than approximately 1.25 at lags 1, 2 and 3, and less than about 1.6 at large lags, it can be concluded that the random shocks at that lag are independent (Pankratz ,1983).

A chi-squared test can also be used to check for the assumption that the true errors a_t are white noise. The sample residual autocorrelation or $r_k(\hat{a}_t)$ series can indicate any departure from a typical white noise behavior in the residual and may suggest an alternative model specification. The overall test for white noise due to Ljung-Box (known as Q-Statistic) and provided by comparing

$$Q = n\,(n+2)\frac{\sum r_k^2(\hat{a}_t)}{n-k} \qquad (32)$$

with tabulated values of χ^2 for $K-m$ degrees of freedom at α significance level; where m is number of parameters, n number of estimated residuals resulted from the model and k =1, 2,... K. This can be further confirmed through plotting the p-values against different lags graphically to notice the probability strength of residuals being white noise.

II. Test for normality of the residuals: Many standard tests are available to check whether residuals are normally distributed. For instance, employing the normal curve plot of the residuals can help to examine whether the data under considerations show significant skewness. If significant skewness could be noticed, it reveals that the residuals series do not possess the normal distribution characteristics.

3.5.1.4 Evaluating Forecasting Accuracy of the Model

To assess the out-of-sample forecasting accuracy of the model, it is advisable to retain some values at the end of the sample period (usually known as validation period) and are not used to estimate the model (Farnum and Stanton, 1989).

These can be used to measure the accuracy of a forecasting model depending on how close the forecasting values ($\hat{x}_t$) are to the actual value(x_t). In practice, we define the difference between the actual and the forecast value values as the forecast error.

$$\hat{a}_t = (x_t) - (\hat{x}_t) \qquad (33)$$

If the model is performing well in forecasting the actual data, the forecast error will be relatively small. Because of the mathematical entity in the definition of forecast error in Eq. (33), we can evaluate a model's accuracy by looking at various quantitative methods. Most frequently: Mean Square Error (MSE) $=\frac{\sum_{t=1}^{v}\hat{a}_t^2}{v}$; Mean Absolute Error (MAE) $=\frac{\sum_{t=1}^{v}|\hat{a}_t|}{v}$;

Mean Absolute Percentage Error$=\frac{\sum_{t=1}^{v}\frac{|\hat{a}_t|}{x_t}}{v}$;

Root Mean Squared Error (RMSE) $= \sqrt{\frac{\sum_{t=1}^{v} \hat{a}_t^2}{v}}$; where v is the number of retained values ; and Thiele's U Statistic are used. Thiele's U statistic calculates the ratio of the RMSE of the chosen model to the RMSE of the 'naive' model. Thus, a value of one for the Thiele's statistic indicates that, on average, the RMSE of the chosen model is the same as the 'naive' model. A Thiele's statistic in excess of one would lead to reconsider the model as the simple 'naive' model performs better on average. A Thiele's statistic less than one does not lead to automatic acceptance of the model, but does indicate that, on average, it performs better than the 'naive' model (Kirkpatrick and Gaynor, 1994).

By Plotting the forecasting values ($\hat{x}_t$) and actual values x_t on the same graph is another way to check the closeness of the actual and forecasted values. The closer together the two plots, the better the model in forecasting.

CHAPTER FOUR

DATA ANALYSIS AND FORECASTING

The main purpose of this data analysis and forecasting section is to develop a forecasting Seasonal Autoregressive Integrated Moving Average (SARIMA) model using the historical monthly rainfall data recorded at Mekele station by the National Meteorological Agency of Ethiopia (NMAE).

This chapter comprises four main parts: The descriptive analysis, forecasting model building, making forecasts and discussions of results. The descriptive mode of analysis shall focus on summarizing the rainfall data series in order to obtain some clues about the statistical characteristics of the data. This can be performed based on some computed descriptive statistics, plotted statistical graphs and tabulations that are commonly used in time series analysis. The SARIMA forecasting model building which is the main area of interest shall be the second part of the statistical data analysis .Within the subsequent main subsections for this part of forecasting model development: i) Tests of stationarity, ii) tests of non randomness for the stationary series, iii) tentative model identification, estimation and diagnostic checks analysis shall be performed. In the forecasting section, point and interval forecast values of monthly rainfall will be generated for the periods from January, 2010-September, 2011 by using the model developed. Finally, results and discussion will be presented at the end of the chapter.

The statistical software package used for most of the analysis is SAS 9.2. The descriptive analysis, however, used the SPSS 16.0 for windows applications.

4.1 Summary of Descriptive Statistics

The aim of this section focuses on summarizing the historical monthly rainfall series using some commonly used statistical characteristics. For this purpose, we used thirty-five years of monthly rainfall data recorded from January, 1975 to December, 2009 by National Meteorological Agency in Tigray region at Mekele station.

With this data, however, we come across some months of different years where their rainfall values were not recorded for various reasons. Missing values are a common problem faced in the analysis of time series data. Particularly, these missing are embedded in the time series rather than occurring at the beginning or the end of the series. Unfortunately, in our data most of the missing values were not in either of the two extremes of the data. As a result, before data analysis and modeling, estimating these values was important. In this study, we estimated these missing values by taking average of two adjacent observations depending on the seasonality behavior of the series. Having used the data from (Gebrerufael, 2008) for the year 1991 missing values due to political instability, and then we estimated the values for the year 1990 also by the estimation method stated above.

To summarize the data, we estimated an average of thirty five years for every month to get mean monthly rainfall series. The other statistical characteristics of the historical data also obtained with the help of SPSS as shown in Table 1 below. The monthly rainfall values vary on average from 1.1 mm in December to 246.1 mm in August. The coefficient of variation (CV) varies from 0.3 in the month of August to 2.7 December. There is large variability among the monthly values of rainfall of different years. If we look at the estimated CV for the months of February, November and December, i.e. 2.2, 2.4 and 2.7 and compared with 0.8, 0.4 and 0.3 for June, July and August respectively; variability is maximum during months of dry seasons but minimum during the rainy season. This may indicate that climate instability during dry season is higher than in rainy seasons.

Table 1: Descriptive statistics of the historical data (monthwise)

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Mean:	2.9	7.2	24.3	35.4	39.3	43.5	193.6	246.1	40.7	6.4	6.1	1.1
S.D.:	5.3	15.5	19.6	39.3	35.0	36.6	78.2	79.6	33.9	11.4	14.7	3.0
CV :	1.8	2.2	0.8	1.1	0.9	0.8	0.4	0.3	0.8	1.8	2.4	2.7
Min.:	0.0	0.0	0.0	0.0	0.0	0.0	32.4	78.9	1.3	0.0	0.0	0.0
Max.:	22.0	83.1	79.6	135.0	126.7	151.9	380.3	430.7	130.3	53.1	59.9	15.7

The mean annual rainfall in Tigray region, Mekele station is computed to be 643.8 mm. The minimum annual rainfall 293.2mm was recorded during the year 1984 and the maximum 918.1 mm in the year 1980.

Table 2: Summary statistics

n	Mean	Standard deviation	Minimum	Maximum
420	53.0	84.4	0.0	430.7

Without seasonal consideration, the summary statistics of the rainfall series for 420 months is given in Table 2 above. The minimum rainfall values was the most frequently marked in months of dry season and maximum rainfall value was recorded in August month of the year 1989.

We plotted the monthly rainfall data against time from January, 1975 to September, 2009 as shown in Figure1. Since there is no noticeable long lasting upward or down ward movement over the span of the rainfall data, the indicated plot seems to be trend free. It appears rather, fluctuating horizontally around its mean ($\approx$-0.017). However, as it is expected, there is yearly seasonal variation due to the periodic high peaks and low peaks observed.

4.2 SARIMA Modeling of Monthly Rainfall Series

In this section, we are dealing with modeling monthly rainfall series using Time Domain Univariate Box-Jenkins methodology of time series analysis. In applying this forecasting model building methodology, the stationary and non randomness assumptions should be tested before we start with model building process (i.e. Identification, estimation, diagnostics and forecasting). In addition, other relevant statistical tests of assumptions will be carried out throughout this section until a suitable forecasting model shall be obtained.

4.2.1 Tests for Assumptions

4.2.1.1 Tests for Stationarity

Graphic Inspection: In time series model building, the first step is to plot the data, calculate sample autocorrelation, partial autocorrelation coefficients and drawn them graphically from which we obtain important information about the series stationary. If the time plot of the series shows that the data scattered horizontally around a constant mean, then the series is called stationary at its level (Chatfield, 1996). On the other hand, if the time plot does not seem to be horizontal, then the series is not stationary. The pattern of the time series plot in Fig.1 does not show apparent systematic change about the mean. The periodic peaks in the plot, however, reflect the yearly regular seasonality (with seasonality interval s=12) of the rainfall values. The series is, therefore, seasonal due to a high rainfall values during the rainy months and a relatively lesser peak rainfall time in the other months. This indicates that the rainfall data have seasonal unit root (i.e., seasonally not stationary).

From the autocorrelation function plot in Figure 2, the presence of seasonality behavior and seasonally non stationarity of the rainfall series is clear. Because there is a sinusoidal wave pattern at the multiple of seasonal intervals and declining slowly while non seasonal lags are relatively decaying quite rapidly. However, the time series

models are usually fitted to the decaying part of autocorrelation (Dulluer and Kavas, 1978). It is, thus, necessary to remove the circularly component of the time series corresponding to the sinusoidal periodic component of the autocorrelation function to make series seasonally stationary. Kavas and Dulluer (1975) have investigated three methods of removal of periodicity in the monthly time series: non-seasonal differencing, seasonal differencing and monthly mean subtraction. For our data, we need only seasonal differencing to make the series stationary.

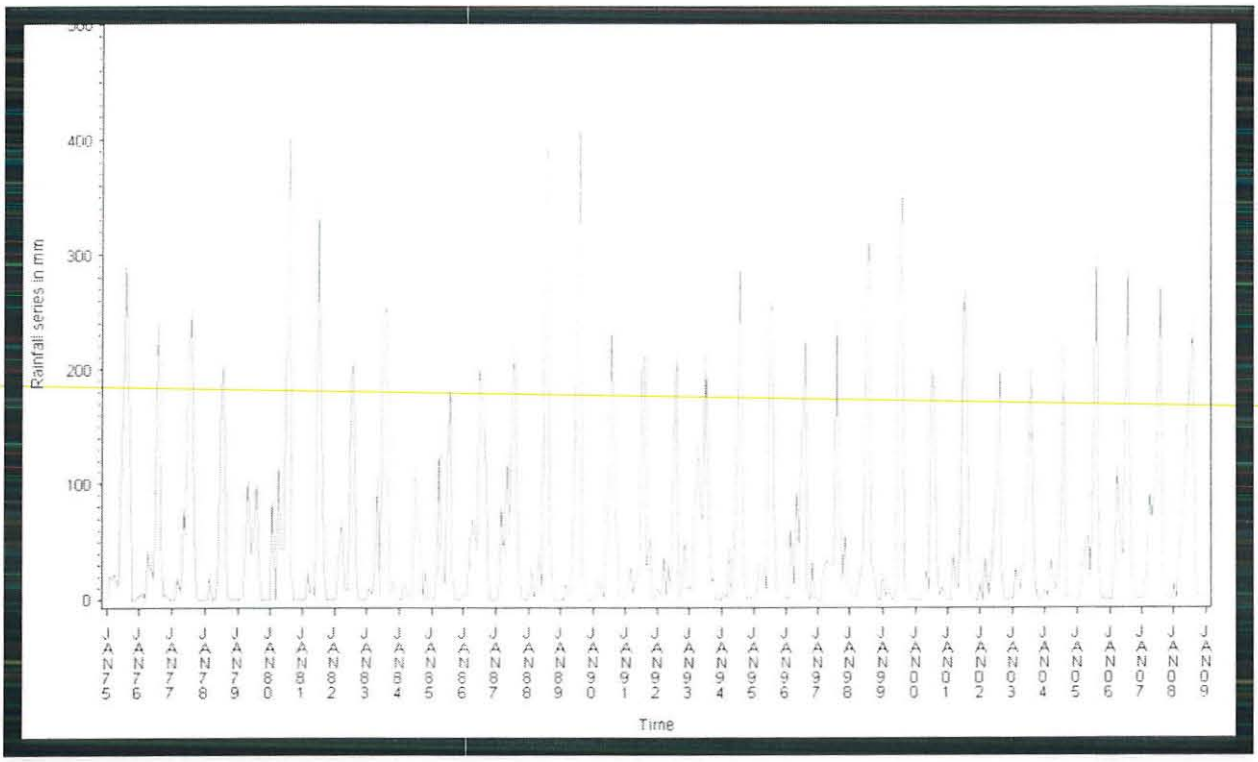


Figure 1: Plot of the original rainfall data

The patterns of monthly rainfall series plot and autocorrelation function suggest the need of seasonal differencing but not–regular differencing.

We have also several ways to ascertain this. In what follows, we perform some tests of series stationarity.

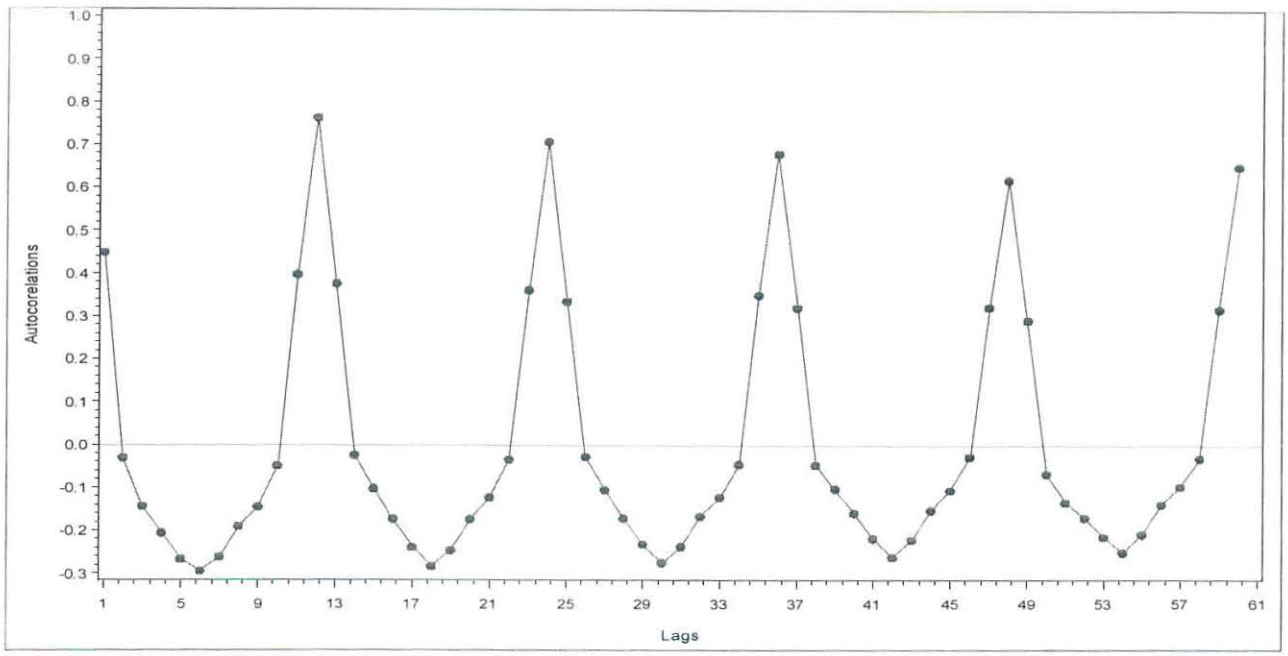


Figure 2: Autocorrelation plot for the untransformed monthly rainfall series.

Augmented Dickey-Fuller Test: A statistical test for stationary is the most widely used Dickey-Fuller test. This test is based on the estimate of the following regression equation with no deterministic trend.

$$\nabla x_t = \phi^* x_{t-1} + \gamma_1 \nabla x_{t-1} + \ldots + \gamma_p \nabla x_{t-p} + a_t \qquad (34)$$

where p is the number of autoregressive terms

To test the hypothesis that the series x_t stationary, we formulate the following hypothesis

H_O: The series is non-stationary versus $(\phi^* = 0)$

H_1: The series is stationary $(\phi^* < 0)$ at α=0.05.

Usually the order of p in the regression equation is set to three. Then if the estimate of ϕ^* is nearly zero in the fitted regression Eq. (34), the original series x_t needs first differencing, and if the estimate of $\phi^* < 0$, then the original series is already stationary (Makridakis et al., 1998). Based on the regression procedure in SAS application, it was found that the estimated value for

ϕ^* to be (ϕ^*=-0.41, $p < 0.001$). At 5% significance level, these figures further confirm that original time series plot without obvious trend in Fig.1 represent series stationary about a constant mean and the autocorrelation function in Fig.2 exhibits non-seasonally decaying quite rapidly. As a result, both tests appear to agree resulting to avoid first non seasonal differencing.

Variance Comparison: The behavior of variance associated with different orders of differencing can provide a useful means of deciding the appropriate order of differencing (Mills, 1999) cited also in (Alamarew Belay and Eshetu Wencheko, 2009). The sample variance decreases until a stationary series has been found. While we increase the differencing order tends to increase in the variance, there will be an indication of overdifferencing. To examine the monthly series via this test associated with the candidates of first differencing and non-differencing for non seasonally and seasonally stationarity, we computed the sample variance for each of x_t, ∇x_t, $\nabla_{12} x_t$, and $\nabla_{12}^2 x_t$ series and obtained the following results:

(i) Var(∇x_t)=8064.1, Var(x_t)=7106.5, Var($\nabla_{12} x_t$)=2745.8, and $\nabla_{12}^2 x_t$=7850 values;

(ii) Var(∇x_t) > Var(x_t);

(iii) Var ($\nabla_{12}^2 x_t$) > Var(x_t) > Var($\nabla_{12} x_t$).

These results suggest that non-seasonal first differencing (∇x_t) has been over-differenced and hence the original series is non-seasonally stationary .The first seasonal differencing would rather be important but not $\nabla_{12}^2 x_t$, because the Var ($\nabla_{12}^2 x_t$) is greater than Var($\nabla_{12} x_t$).
These tests for stationarity all seem to agree and suggest that the first seasonal differencing in the series can achieve stationarity around a constant mean, which is approximately zero and calculated its standard deviation to be 52.4 mm (see Figure 3). Moreover, the ACF and PACF shown in Figure 4 are in support of monthly rainfall series stationarity in both the mean and the variance after having first seasonal difference.

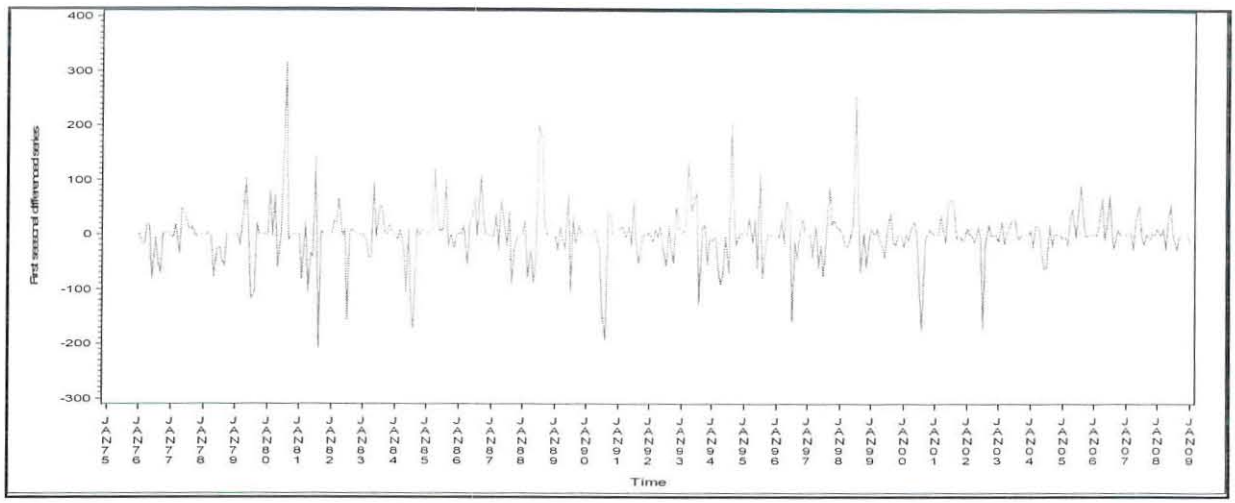
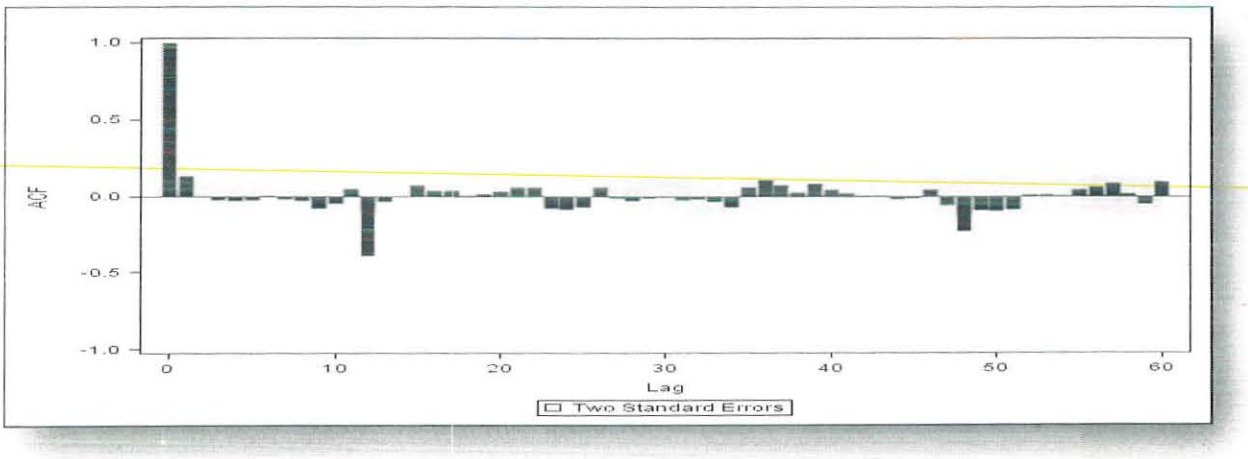
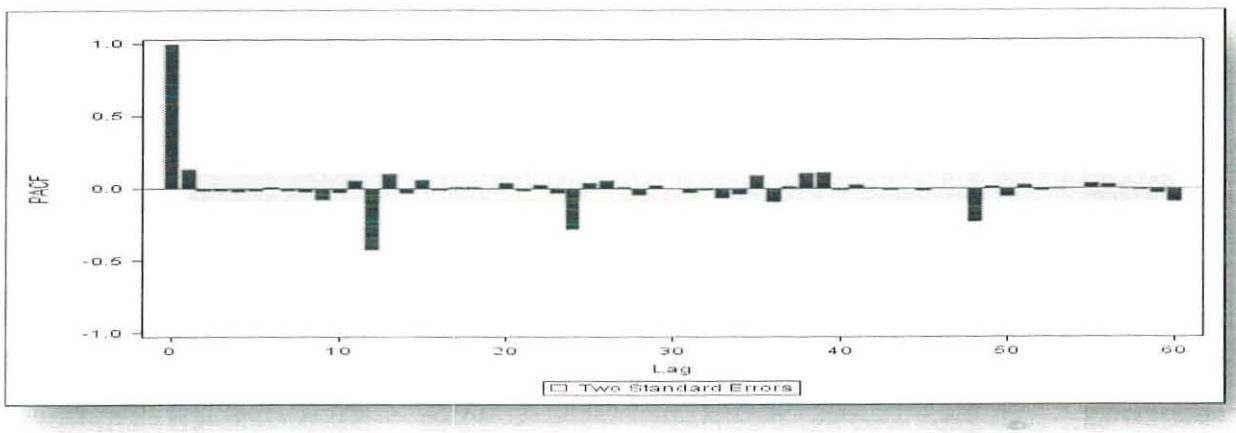


Figure 3: Plot for First seasonal differenced monthly rainfall series



(a)



(b)

Figure 4: **(a):** Autocorrelation Function (ACF) **(b)**: Partial Autocorrelation Function (PACF) for the first seasonal differenced monthly rainfall.

Table 3: Estimated Autocorrelations and Partial Autocorrelations magnitudes for x_t and $\nabla_{12}x_t$ series.

Lags	**Autocorrelations for** x_t and $\nabla_{12}x_t$											
1-12	0.45	-0.03	-0.14	-0.21	-0.27	-0.30	-0.26	-0.19	-0.15	-0.05	0.40	0.76
13-24	0.46	-0.03	-0.10	-0.17	-0.24	-0.28	-0.25	-0.17	-0.12	-0.03	0.36	0 .71
25-36	0.35	-0.03	-0.11	0.17	-0.23	-0.28	-0.24	-0.17	-0.12	-0.05	0.35	0.68
37-48	0.32	-0.05	-0.10	-0.16	-0.22	-0.26	-0.22	-0.15	-0.11	-0.03	0.30	0.62
49-60	0.29	-0.07	-0.13	-0.17	-0.22	-0.25	-0.21	-0.14	-0.10	-0.03	0.32	0.65
1-12	0.14	-0.01	-0.02	-0.03	-0.03	0.01	-0.02	-0.03	-0.09	-0.05	0.00	-0.40
13-24	-0.04	-0.01	0.07	0 .04	0.04	0.01	0.02	0.03	0.06	0.06	-0.09	-0.09
25-36	-0.07	0.06	-0.01	-0.03	-0.02	-0.01	-0.02	-0.02	-0.04	-0.07	0.06	0.10
37-48	0.07	0.03	0 .09	0.05	0.02	0.00	0.01	-0.02	-0.02	0.04	-0.06	-0.23
49-60	-0.09	-0.10	-0.10	0 .01	0.01	0 .01	0.04	0.06	0.09	0.02	-0.05	0.09
	Partial Autocorrelations for x_t and $\nabla_{12}x_t$											
1-12	0.45	-0.29	0.01	-0.18	-0.16	-0.20	-0.18	-0.20	-0.26	-0.22	0.36	0.51
13-24	-0.15	-0.05	0.07	-0.02	0.01	-0.06	-0.01	-0.03	0.04	-0.03	0.05	0.27
25-36	0.12	-0.01	-0.02	-0.02	0.00	-0.03	-0.01	-0.04	-0.01	-0.05	0.04	0.16
37-48	-0.08	-0.08	0.03	0.00	-0.01	-0.02	0.01	-0.02	0.05	0.01	-0.07	0.04
49-60	0.01	-0.12	-0.09	0.02	-0.03	-0.04	-0.02	-0.06	-0.04	-0.04	0.00	0.18
1-12	0.16	-0.02	-0.04	-0.04	-0.02	0.01	-0.02	-0.03	-0.08	-0.03	0.05	-0.41
13-24	0.09	-0.02	0.04	-0.02	0.01	0.01	0.00	0.04	-0.02	0.03	-0.04	-0.21
25-36	0.07	0.07	-0.02	-0.05	0.02	0.00	-0.03	-0.01	-0.07	-0.04	0.07	-0.08
37-48	0.03	0.12	0.10	-0.02	0.03	0.01	0.01	-0.01	-0.02	0.01	-0.01	-0.19
49-60	0.04	-0.01	0.01	-0.02	-0.01	-0.01	0.03	0.02	-0.01	0.01	-0.02	-0.10

4.2.1.2 Tests for Randomness

Testing for randomness is a fundamental aspect of time series analysis once stationarity of the series has been established. The simplest time series is a random model, in which the observations vary around a constant mean, have a constant variance, and are probabilistically independent (Harvey,1993). In other words, a random time series does not have time series pattern, meaning that there is no point in attempting to fit a time series model to such type of data. Therefore, it is important to perform tests of randomness before any attempt to modeling process with the underlining methodology.

We pursue checking through the following tests to investigate the hypothesis that the first-seasonally differenced monthly rainfall series are serially uncorrelated.

Graphic Inspection: The visual inspection of the autocorrelation function plot provides useful in formations to identify the type of time series (Chatfield, 1996). For example, if a time series is completely a random series, then for large n, $r_k \approx 0$ for every k. This can be examined after the array of autocorrelation coefficients r_k, plotted with k as abscissa and r_k as ordinate.

Figure 4(a) exhibits the graph of sample autocorrelations given in Table 3 against different lags from which we can observe visually that the autocorrelations are not all insignificant. This indicates that there is some sort of dependence between values of $\nabla_{12}x_t$ series.

The graphical analysis and interpretations can be further ascertained by the following formal statistical tests of randomness.

Bartlett's Band Test: If a stationary time series is completely random, then for large n, the sample autocorrelations (r_k) and partial autocorrelations (r_{kk}) are approximately zero for all values of k (Chatfield, 1996). Statistically, if the time series is random, 19 out of 20 of the values of r_k and r_{kk} are expected to lie between the approximate Bartlett's bands $\pm 2/\sqrt{n}$.

Based on the assumption that the values in $\nabla_{12}x_t$ series are serially independent, we have s.e(r_k) =(r_{kk})$\approx \pm 2/\sqrt{408} = \pm 0.1$ for all lags k=1,2,3,..K and n=408. Here, the constant k should be specified in advance. As a rule of thumb for determining for this constant is to choose k in the neighborhood of $\sqrt{n}$ (Harvey, 1993) cited also in (Alamarew Belay and Eshetu Wencheko, 2009). Hence, in our case we set k=20 and we refer the first 20 autocorrelation and partial autocorrelation estimated values listed in Table.2 for this comparison. Accordingly, since autocorrelations 1 and 12 and partial autocorrelation values at lags 1, 12 and 13 are significantly lie outside the estimated Bartlett's band ± 0.10 , we reject the assumption that the $\nabla_{12}x_t$ series are serially independent in favor of the alternative that the series is serially correlated. Therefore, based on the magnitudes of individual autocorrelation coefficients test, we can conclude that the first seasonally differenced monthly rainfall series have not shown randomness behavior.

Box-Ljung Test Statistic: This statistic is used for collectively testing the magnitude of the autocorrelation of stationary time series for significance. For this test, we used the sample autocorrelation coefficients of the first seasonally differenced monthly rainfall as well.

The hypothesis to be tested is

Ho: All autocorrelations up to lag J are zero

Versus

H_1: Not all up to lag J are zero at α= 0.05.

The statistic for this test problem is $Q = n\,(n+2)\sum_{j=1}^{J}\frac{r_j^2}{n-j}$ and has a chi-square distribution with J degrees of distribution. Q-Statistic is usually computed for J= 6, 12 24, 36 and 48 by most of the statistical packages. However, J=12 or 24 will prove to be satisfactory (Kirkpatrick and Gaynor, 1994). In this regard, we compute the test statistic above for the first J=12 lag autocorrelation values and n=408 observations. The value of the calculated Q-Statistic is found to be 43.72 and the tabulated value for chi-distribution with 12 degree of freedom at 0.05 significance level is 21.02. The decision to reject H_o is based on whether the value of Q-Statistic > $\chi^2_{0.05,J}$;if that does not hold we do not reject H_o. Since for the data set considered Q-statistic=43.7>

$\chi^2_{0.05}$=21.2, we reject H_o, we conclude that the seasonally first differenced monthly rainfall series are serially correlated.

In summary, the above tests regarding to stationarity and series randomness all appear to agree in that the monthly rainfall data has been achieved stationarity after seasonal first differencing and the data correlated.

4.2.2 Model Identification

Having established that the monthly rainfall data are serially correlated and stationary, the next step in the identification process is to find the initial values for the order of seasonal and non seasonal parameters *p, q, P*, and *Q* that should be incorporated in the SARIMA $(p, 0, q) \times (P, 1, Q)_{12}$ model. A basic model can now be identified with the help of sample autocorrelation and sample partial autocorrelation functions characteristics. To use the sample autocorrelation and sample partial autocorrelations functions for tentative model parameters identification, we consider the ACF and PACF shown in Figure 4. The first step in this direction is identifying the significant autocorrelations and partial autocorrelation from the ACF and PACF plots of the underlying stationary series (Hipel *et al.*, 1977). Hence in our case, for the $(1\text{-}B^{12})x_t$, we find significant correlations at lag k=1, 12 and k=48 at the autocorrelation function plot and k=1, 12, 24 and k=48 partial autocorrelation in Figure 4(a) and 4(b), respectively. Hence, we suggested that a model of the form

$$(1\text{-}\Phi_1 B^{12})(1\text{-}B^{12})x_t = (1\text{-}\theta_1 B)(1\text{-}\Theta_4 B^{48})a_t \qquad (35)$$

to fit in the first instance with one non seasonal moving average, one seasonal moving average and one seasonal autoregressive, denoted by θ_1, Θ_1 and , Φ_4 parameter coefficients to be estimated.

Another alternative model seems to be appropriate tentatively at this stage is made based on the principle that when the process is a pure SARIMA $(p, d, 0) \times (P, D, 0)_{12}$ model, r_{kk} cuts off and

is not significantly different from zero after lag $p + SP$. If r_{kk} damps out at lags that are multiples of s, this suggests the incorporation of a seasonal MA component into the model. The failure of the PACF to truncate at other lags may imply that a non-seasonal MA term is required (Hipel *et al.*, 1977). Accordingly, we specified SARIMA $(1, 0, 0) \times (4, 1, 1)_{12}$ model also tentatively to be checked through the estimation and diagnostic stage in order to take the best model.

4.2.3 Model Estimation and Diagnostics Analysis

Estimating the parameters for Box-Jenkins models is a quite complicated non-linear estimation. Parameter estimates are usually obtained by maximum likelihood method, which is more appropriate for time series (Brockwell and Davis, 1996). Estimates are usually sufficient, efficient, and consistent for Gaussian distribution and are asymptotically normal. Using maximum likelihood estimation method, the parameters of SARIMA $(0, 0, 1) \times (1, 1, 4)_{12}$ and $(1, 0, 0) \times (4, 1, 1)_{12}$ are Estimated and given in Table 4.

Table 4: Parameter estimates for suggested SARIMA models.

(a): $(1-\Phi_1 B^{12})(1-B^{12})x_t = (1-\theta_1 B)(1-\Theta_4 B^{48})a_t$ or $(0, 0, 1)\times (1, 1, 4)_{12}$

(b): $(1-\phi_1 B)(1-\Phi_4 B^{48})(1-B^{12})x_t = (1-\Theta_1 B^{12})a_t$ or $(1, 0, 0)\times(4, 1, 1)_{12}$

(c): $(1-\phi_1 B)(1-\Phi_4 B^{48})(1-B^{12})x_t = (1-\Theta_1 B^{12}-\Theta_2 B^{24})a_t$ or $(1, 0, 0)\times(2, 1, 4)_{12}$

Model	Parameter	Estimate	St.error	t-value	P-value	Fit statistics
	θ_1	-0.15	0.05	-2.98	0.0030	AIC=4302.17
	Θ_4	0.25	0.04	4.15	<0.0001	RMSE.=38.62
(a)	Φ_1	-0.42	0.05	-8.95	<0.0001	r^2 =0.81
						AIC=4291.80
	ϕ_1	0.83	0.07	10.06	<0.0001	RMSE.= 37.46
(b)	Φ_4	0.13	0.05	2.71	0.0070	r^2 =0.81
	Θ_1	-0.29	0.04	-5.74	<0.0001	

(c)	ϕ_1	0.13	0.08	10.04	<0.0001	AIC =4178.73
	Φ_4	0.34	0.05	-1.98	0.0700	RMSE.= 37.58
	θ_1	0.79	0.03	2.70	0.0071	r^2 =0.83
	θ_2	0.07	0.06	6.68	< 0.0001	

Table 5: Correlations of Parameter Estimates for the two models

Model (a): SARIMA (0,0,1) x $(1,1,4)_{12}$				Model b:SARIMA (1,0,0) X $(4,1,1)_{12}$			
Parameter	θ_1	Θ_4	Φ_1	Parameter	ϕ_1	Φ_4	Θ_1
θ_1	1.00	-0.01	-0.03	ϕ_1	1.0	-0.07	0.03
Θ_4		1.00	0.02	Φ_{48}		1.00	0.08
Φ_1			1.00	Θ_{12}			1.00

In the estimation procedure, two types of outliers (5 additive and 1 shift outliers) were detected and adjusted in the fitted models by SAS software.

Diagnostic checking includes overfitting, information criteria and residual analysis. We start with overfitting by including one more seasonal moving average parameter (which measures the error dependency effect at 24 distant or lag 24 and denoted by Θ_2) to the SARIMA model (b) to examine whether this model with more parameters would adequately be fitted to the seasonally first differenced monthly rainfall data. The inclusion of this parameter can be determined by testing its significance and the improvement in the measures of goodness of fit of the model. All substantial parameters in all the models in Table 4 showed statistically significance except the SARIMA model (c) with ϕ_1, Φ_4 Θ_1 and Θ_2 parameter coefficients. One estimated parameter (Θ_2 = 0.06, P-value=0.07 >0.05) which is insignificant because p-value > 0.05, highly supports the probability power to be the estimated parameter statistically approximately zero at α=0.05 level of significance. As a result, including this parameter (Θ_2) to the model (b) that has been identified in the model identification

stage will have no visible contribution in the model (c). Thus, the model will have one parameter less than in (c).

Information criteria which are the relative measures goodness of fit that is grounded in the concept of entropy, when a given model, is used to describe the data under study. In deciding on the number of parameters in the model, we use AIC as computed by Eq. (31). In Table 4, models (a) and (b) which have one parameter less than model (c). In this context, since all these three models also have almost the same RMSE and AIC values, these results are indications to drop the third model (c) which is less parsimonious and we proceed to check the adequacy of the remaining two models using residual analysis. As models may perform reasonably similar, a number of alternative formulations may have to be retained at this stage to be further assessed at the forecasting stage.

Assessment of each of the suggested models will involve a rigorous assessment of diagnostic tests for the competing models. There are a number of diagnostic tools available for ensuring a satisfactory model is arrived at. The residual analysis parts of diagnostic checking are Tests for white nosiness and normality of residuals.

The Autocorrelations Functions (ACF) and Partial Autocorrelations Functions (PACF) of the residuals resulted from the fitted models should not show any pattern (trend or seasonality pattern). Moreover, for a correctly fitted model the residuals correlation coefficients should not lie outside the two standard error at a given significant level. It is clear, as shown in Figure 5 and Figure 6 that there is no pattern in residuals ACF and PACF plot for model (a) and (b) respectively. In addition, no autocorrelation partial or partial autocorrelation coefficient lie outside the two standard errors significantly at 5% level of significance for both fitted models. The scatter plot of residuals from the fitted model should not possess an obvious pattern (trending or seasonality behavior) as well. Because the graphical analysis shows that the residuals in the model appeared to fluctuate randomly around zero with no apparent pattern as it can be noted from Figure 7 that was in support of both models adequacy. In examining the model residual

histogram (normal curve), we can visually inspect that the Figure (5c) do not show any violation of the models' assumption that the residuals are normally distributed with mean zero and constant variance.

Another way to accomplish the residual analysis for white noise residuals test is based on chi-square statistic which is denoted by Q-Statistic. The Q-statistic at each grouped in six lags are computed using Eq. (32) and we obtained the following results given in Table 6a and 6b.

Table 6a: White noise check for residuals Autocorrelations of SARIMA (0, 0, 1) × $(1, 1, 4)_{12}$

To Lag	Q_{cal}	DF	p-value	Autocorrelations					
6	2.51	3	0.4729	0.003	-0.000	-0.061	-0.047	-0.011	-0.001
12	9.79	9	0.3681	-0.016	-0.014	-0.079	-0.019	0.086	0.053
18	14.91	15	0.4580	0.010	0.083	0.062	-0.011	0.029	-0.015
24	20.76	21	0.4737	-0.015	-0.004	-0.027	0.055	-0.041	-0.088
30	34.49	27	0.1523	-0.056	0.145	0.075	-0.004	0.023	-0.031
36	36.52	33	0.3084	0.001	-0.024	-0.018	-0.030	-0.015	-0.050
42	41.32	39	0.3696	-0.024	0.027	0.082	0.032	0.029	-0.026
48	44.19	45	0.5060	-0.004	-0.022	-0.022	0.041	-0.054	-0.024
54	50.40	51	0.4974	-0.104	-0.014	-0.028	0.006	0.030	-0.021
60	58.04	57	0.4367	0.010	0.031	0.029	0.028	-0.109	-0.037

Degree of freedom (DF) associated with $\chi^2_{\alpha,(k-m)}$, for k = 6, 12,... , 60 lags in months ,for m=3 number of parameter in the model and at α=0.05.

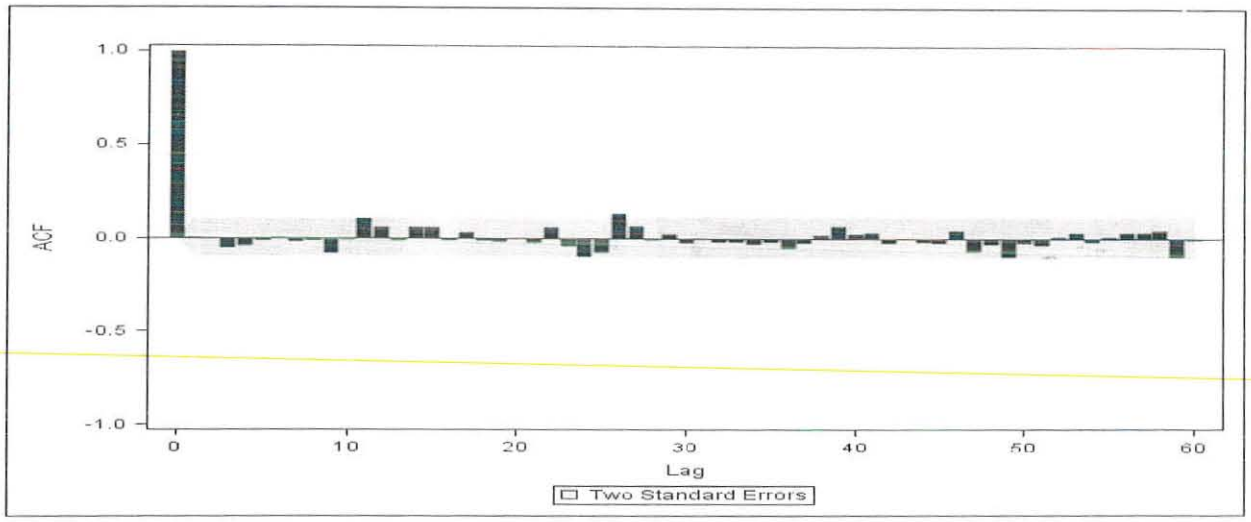
Table 6b: White noise check for Residuals Autocorrelations of SARIMA $(1, 0, 0) \times (4, 1, 1)_{12}$ model.

At lags	Q_{cal}	DF	p-value	Autocorrelations					
6	2.95	3	0.3991	0.008	0.031	-0.061	-0.046	-0.015	-0.003
12	10.10	9	0.3421	-0.020	-0.015	-0.077	-0.018	0.085	0.055
18	15.51	15	0.4154	0.016	0.085	0.064	-0.007	0.030	-0.015
24	20.95	21	0.4622	-0.016	-0.003	-0.029	0.051	-0.043	-0.083
30	34.19	27	0.1605	-0.054	0.142	0.075	0.000	0.025	-0.032
36	36.44	33	0.3115	0.001	-0.027	-0.019	-0.033	-0.016	-0.051
42	41.27	39	0.3715	-0.021	0.026	0.083	0.032	0.031	-0.025
48	44.28	45	0.5023	-0.004	-0.022	-0.024	0.039	-0.057	-0.025
54	50.76	51	0.4832	-0.107	-0.015	-0.031	0.005	0.029	-0.019
60	58.57	57	0.4175	0.012	0.032	0.026	0.027	-0.112	-0.036

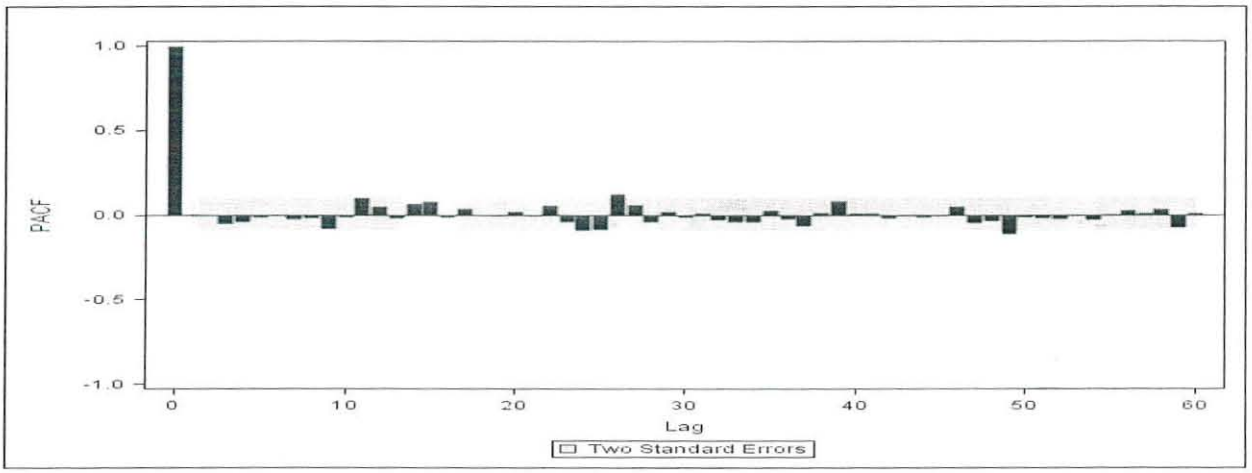
The computed values of Q-statistic are compared to critical values from chi-square distribution with the indicated respective degrees of freedom in Tables 6a and 6b. Note that the P-value is the probability that $\chi^2_{\alpha,(k-m)}$ is higher than our observed test statistic (Q_{cal}). If a model is correctly specified, residuals should be uncorrelated and Q_{cal} should be small (probability value should be large). A significant value indicates that the chosen model does not fit well. The P-value in the area of under the curve of the Chi-square distribution having the degree of freedom (DF) to the right of the Q-statistic demonstrates also the same conclusion as comparing the Q-statistic with $\chi^2_{\alpha(k-m)}$. In this test, the results in Table 6a and 6b show that the P-values associated with Q_{cal} for lags up to k=6, 12... 60 are all greater than 0.05 for the two seasonal models. As shown from plots in Figure 5d and Figure 6c that represent the white noise probability function for the P-values against the various set of lags in agree with that of all probabilities for the residuals of being white noise lie above the α=0.05. Therefore, the set of autocorrelation for residuals are not significant and we cannot reject the hypothesis that the autocorrelations of the residuals are zero.

The Ljung-Box results based on collective test confirmed with what the plots of the Autocorrelation Function (ACF) and Partial Auto Correlation Function (PACF) for

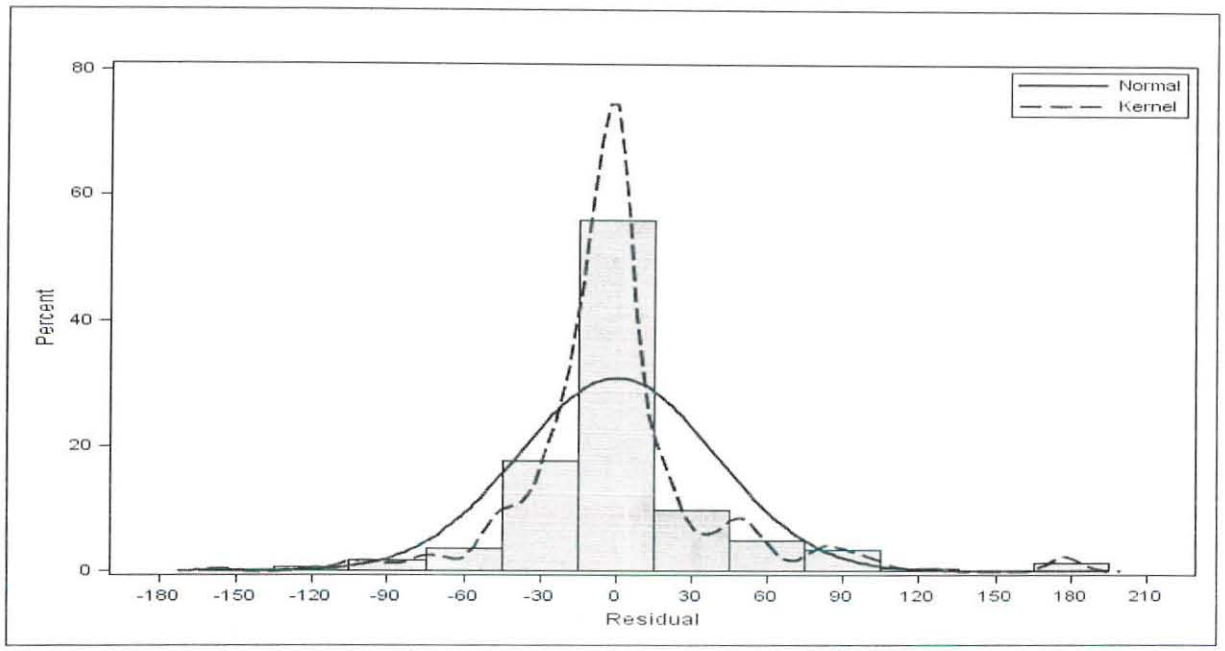
residuals in Figures 5 and 6 reveal that there are no distinctive correlation coefficients that lie outside of the two standard error confidence limits at 0.05 significance level for both models. These results are in agreement with the hypothesis that the residuals resulted from each of the possible models do not show any correlation or pattern (i.e. what is left after fitting each model are just noise) and these are normally distributed, we conclude that the two SARIMA (0, 0, 1) × (1, 1, 4) and (1, 0, 0) × (4, 1, 1) models are found to be adequately fitted to the seasonally first–differenced monthly rainfall series.



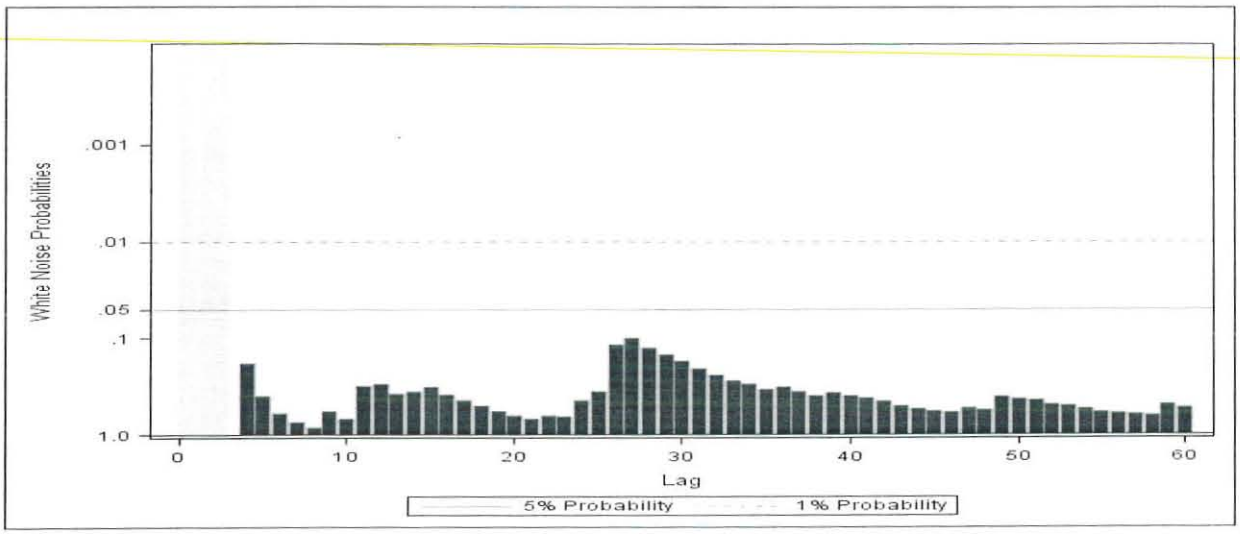
(a)



(b)

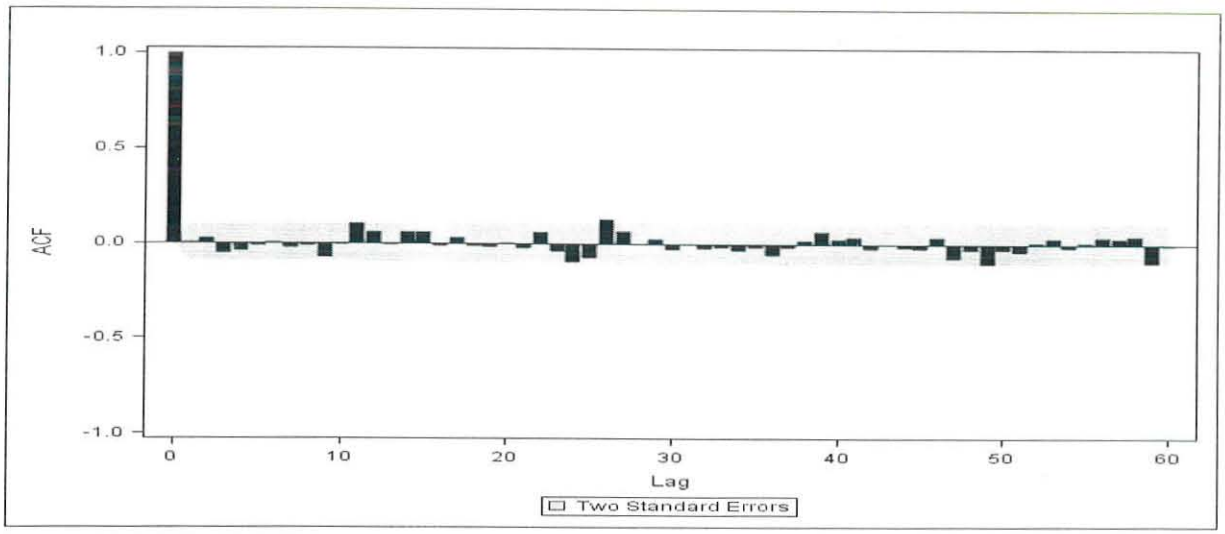


(c)

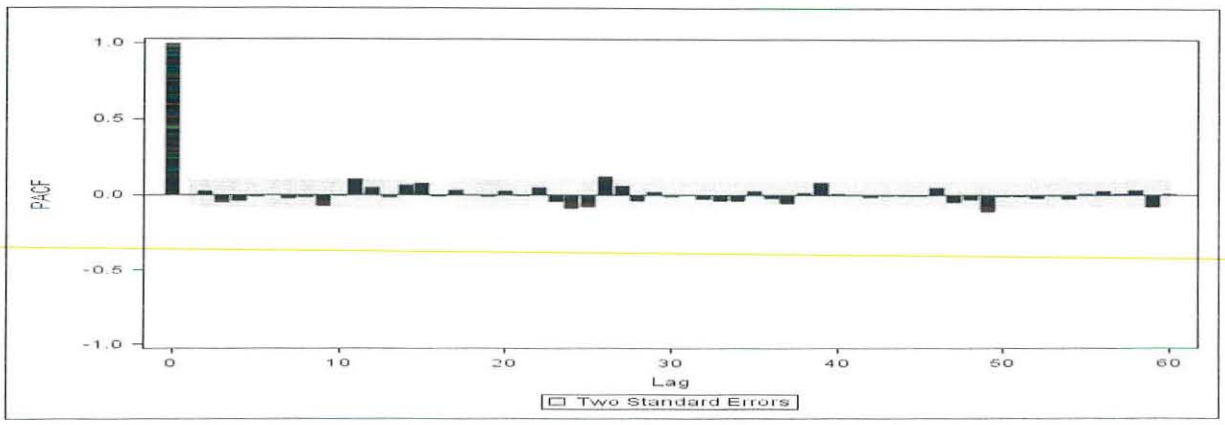


(d)

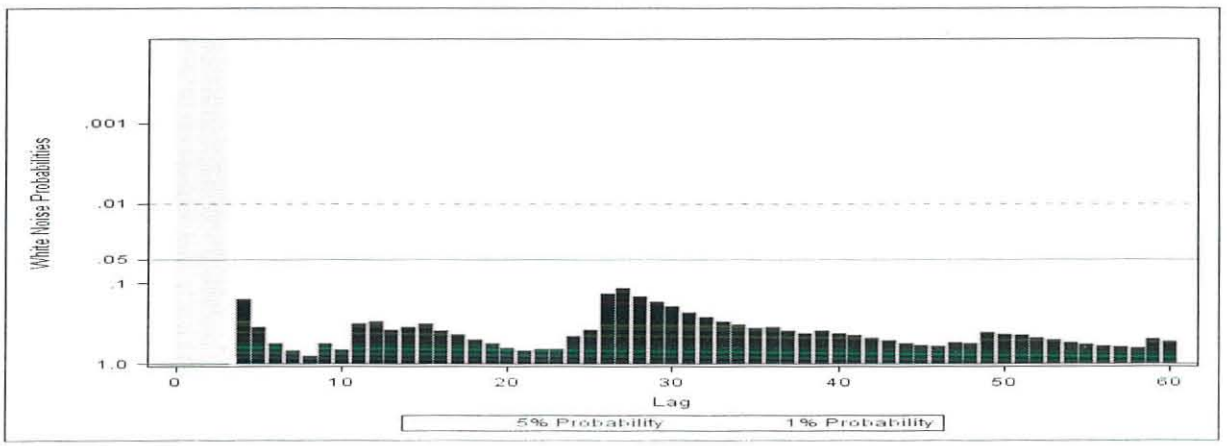
Figure 5: (a): Residual Autocorrelation (RACF) **(b):** Residual Partial Autocorrelation (RPACF) **(c):** Diagnostics for residuals Normality distribution **(d):** White noise test p-values Plot for Residuals resulted from SARIMA $(0, 0, 1) \times (1, 1, 4)_{12}$ model.



(a)



(b)



(c)

Figure 6: (a): Residual Autocorrelation (RACF) **(b):** Residual Partial Autocorrelation (RPACF) **(c):** White noise Test P-values Plot for Residuals resulted from SARIMA $(1, 0, 0) \times (4, 1, 1)_{12}$ model.

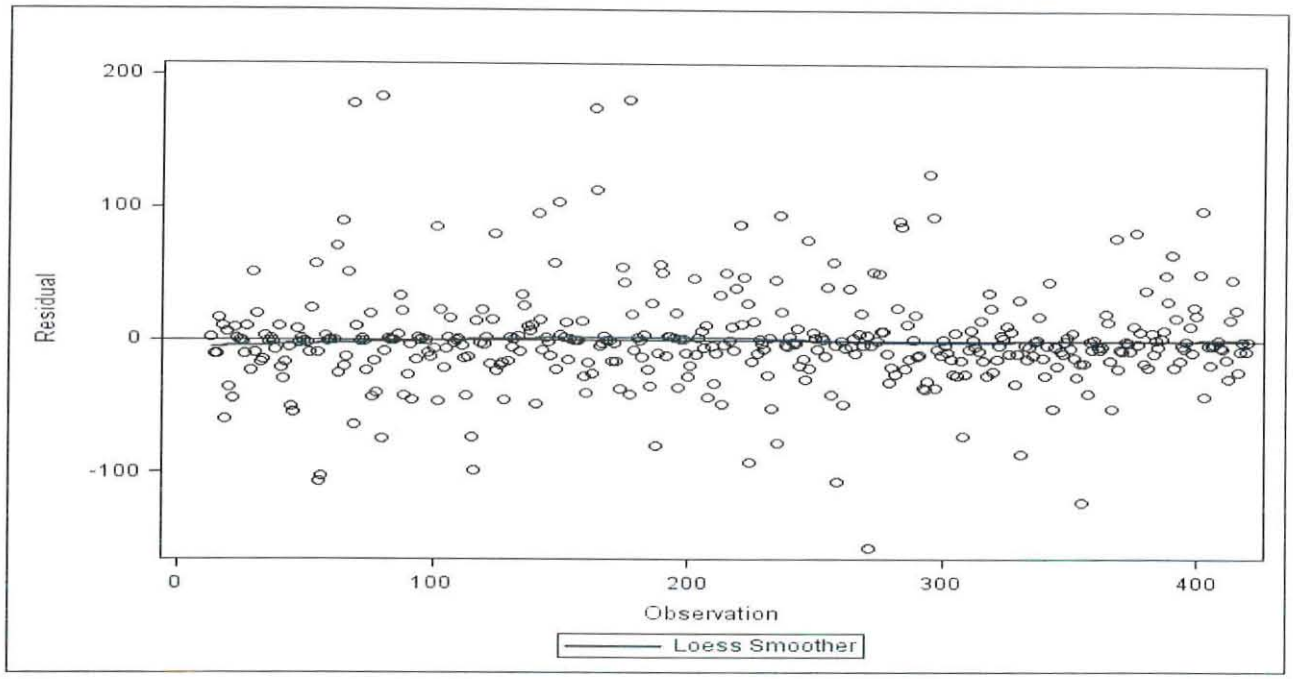


Figure 7: Scatter plot of residuals from the fitted model.

The statistical residual analysis for both models (a) and (b) showed similarities of model adequacy. Thus, the results are supportive of the randomness of residuals in both models at 5% level of significance.

In summary, the diagnostic checks section, retaining the two models up to the forecasting stage; further tests would be necessary to select the better of the two models with relative better forecasting accuracy. Therefore, further tests should be done based on the forecasting reliability of the two competing models that are adequately fitted.

4.3 Forecasting

4.3.1 Forecasting Accuracy Assessment of the models

As we have two models that meet the various diagnostic checks criteria, we proceed to compare their forecasting performance using the various accuracy evaluating measures. These evaluating measures can be carried out through quantitative measures. To calculate the model evaluating

results, we retained the last values from (Sep2004 to Dec 2009) monthly rainfall values which are not used for the model development. Then, we computed the series of forecasting errors using Eq. (33) for these hold-out monthly rainfall series and served as a principal function to obtain the results in Table (7).

Table 7: Results of Accuracy Measures for the two models

SARIMA Models	MAE	MAPE	MSE	RMSE	THIEL'S
$(1, 0, 0) \times (4, 1, 1)_{12}$	23.43	265.35	1754.77	41.89	0.15
$(0, 0, 1) \times (1, 1, 4)_{12}$	17.81	219.23	1493.05	38.64	0.12

To measure the forecasting ability of the two models, we have estimated within-sample and out-of-sample forecasts. If the magnitude of the difference between the forecasted and actual values is low then the model has good forecasting performances. In this case, the seasonal ARIMA $(0, 0, 1) \times (1, 1, 4)_{12}$ model has shown better results as is evident from the Table 7. Thicle's U-Statistic is computed to be 0.15 and 0.12 for SARIMA$(1, 0, 0) \times (4, 1, 1)$ and $(0, 0, 1) \times (1, 1, 4)_{12}$ models, respectively. Both results indicate that the two models are reasonably better than the naïve forecasting model. However, because the Thiele's U-Statistics result= $0.12 < 0.15$ for the seasonal ARIMA $(0, 0, 1) \times (1, 1, 4)$ and seasonal ARIMA $(1, 0, 0) \times (4, 1, 1)_{12}$ model, respectively. This is to suggest that the seasonal ARIMA $(0, 0, 1) \times (1, 1, 4)_{12}$ model performs quite better forecasting accuracy than the second model.

It can be concluded from the findings that the forecasting ability of the SARIMA $(0, 0, 1) \times (1, 1, 4)_{12}$ model is found to be suitable for the future monthly rainfall data forecasting purpose. Forecasting validation using graphical analysis has also assessed to see the closeness of its forecast with the actual data of the hold out monthly rainfall values.

Figure 8 represents the forecasts for the validation period and future forecasts of monthly rainfall series using SARIMA $(0, 0, 1) \times (1, 1, 4)_{12}$ model. Notice that

forecasts in the validation period are reasonably close to the actual series and captured the turning points patterns as well.

The ultimate goal is to model the monthly rainfall series applying the Univariate Box-Jenkins time series methodology and provide forecast for lead times up to about two years. In this regard, using the selected model that passed all the stages of model estimation, diagnostic checks and forecasting criteria's statistically, the next task would be producing the point and interval future forecasts of monthly rainfall series at Mekele in Tigray region.

4.3.2 Forecasting Monthly Rainfall values

Forecasting refers to the process of predicting future values from a known time series. In this section, having built an adequate model that appropriately fitted to the historical monthly rainfall series of Mekele station showing better accuracy. Forecasting was performed using the difference equation approach as follows. The SARIMA $(0, 0, 1) \times (1, 1, 4)_{12}$ model can be written as in Eq. (23) of the form

$$(1-\Phi_1 B^{12})(1-B^{12})\, x_t=(1-\theta_1 B)(1-\Theta_4 B^{48})a_t \qquad (36)$$

This equation can also be multiplied out and rewritten in a form that is used in forecasting as shown in Eq. (37) below.

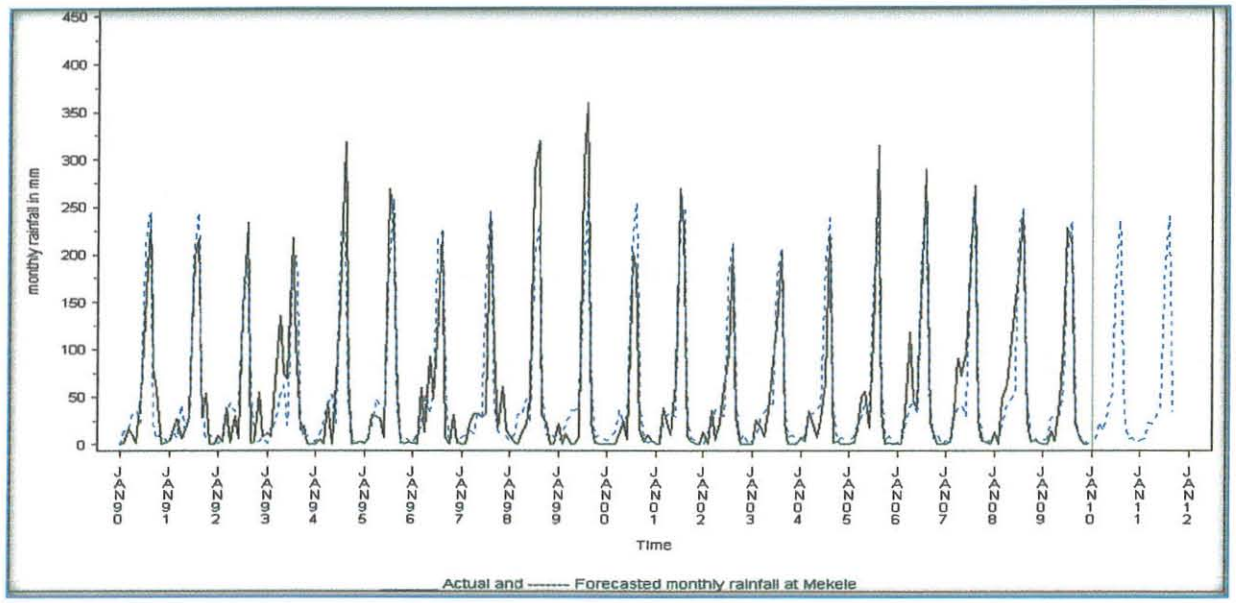
$$x_t = x_{t-12} + \Phi_1(x_{t-12}-x_{t-24}) + a_t - \Theta_4 a_{t-48} - \theta_1 a_{t-1} + \theta_1\Theta_4 a_{t-49} \qquad (37)$$

After substituting the estimated parameter values to Eq. (37) above, we obtain the following difference equation.

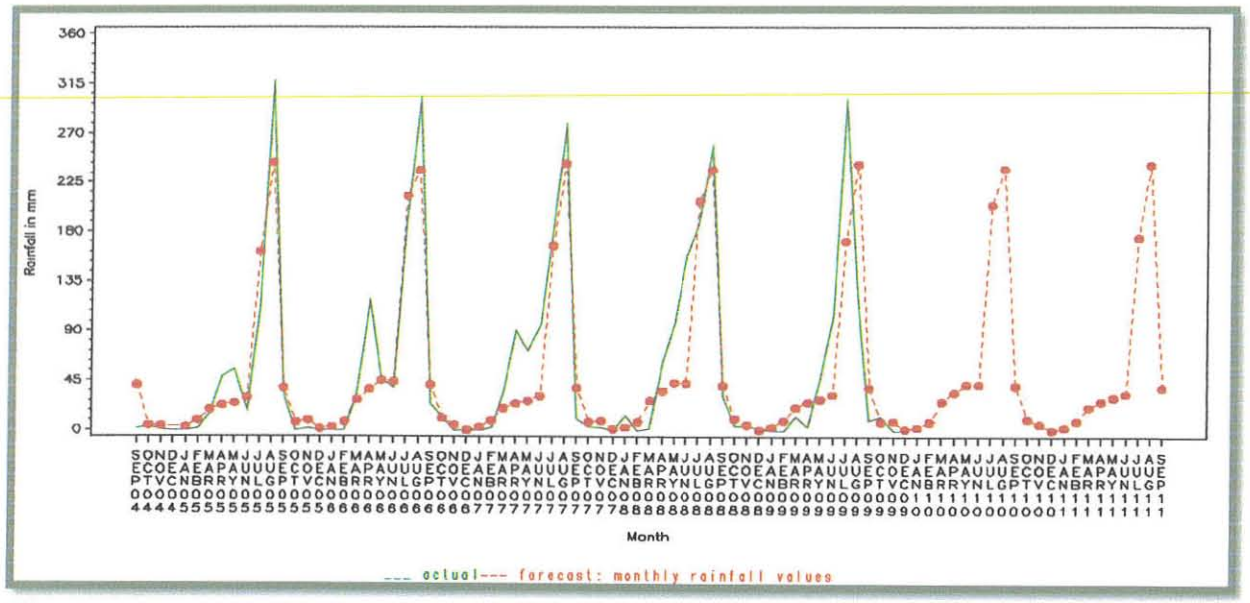
$$x_t = x_{t-12} - 0.42(x_{t-12} - x_{t-24}) + a_t - 0.25a_{t-48} + 0.15a_{t-1} - 0.04a_{t-49} \qquad (38)$$

Table 8: Forecast results of monthly rainfall over the period Jan 2010-Sep2011.

Months	Forecasts	(95%Lower Limit)	(95 % upper Limit)
Jan 2010	1.4	0.0	11.6
Feb 1200	6.2	0.8	23.0
Mar 2010	27.9	3.8	51.9
Apr 2010	36.3	6.4	67.5
May 2010	48.8	21.5	49.7
Jun 2010	49.3	22.4	97.1
Jul 2010	217.3	103.2	254.2
Aug 2010	230.1	109.6	310.4
Sep 2010	35.3	10.2	78.7
Oct 2010	10.1	0.0	19.5
Nov 2010	3.4	0.1	11.0
Dec 2010	0.1	0.3	7.9
Jan 2011	3.9	0.5	17.7
Feb 2011	8.2	3.8	13.7
Mar 2011	18.9	14.9	59.2
Apr 2011	34.6	12.1	37.7
May 2011	31.8	9.8	53.2
Jun 2011	39.1	13.9	68.9
Jul 2011	165.8	117.8	285.0
Aug2011	251.0	143.5	324.9
Sep2011	38.9	9.4	71.0
Mean	**53.5**		
SDV.	**78.4**		



(a)



(b)

Figure 8: (a) Plot of the model estimation, validation and Forecasts **(b):** Plot of the model validation periods (Sep04-Dec2009) and forecasted monthly rainfall series for the periods from (Jan2010-Sep2011).

4.4 Results and Discussions

Descriptive analysis: From the data in Table 1 that represents the 420 monthly rainfalls values, we calculated the minimum, maximum, average and standard deviation to be 0.0, 430.7, 53.0 and 84.4 in mm respectively. In addition, summary statistics of rainfall values were also computed for each month. Descriptive statistics results on the historical month rainfall are presented in Tables 1 and 2.

Results show that there is a considerable variability among the monthly rainfall values and among the monthly values of rainfall of different years. Rainfall values of months of dry seasons showed greater variance relative to months of rainy seasons.

Modeling and Forecasting: Before we start with modeling procedures, we made an attempt to examine what characteristics the original historical monthly rainfall series behaves using graphical analysis. The plot of the original monthly rainfall series against time presented in Figure1 gives some idea about the characteristics of the series. Since the plot of the series does not show any noticeable long lasting upward or downward movement, it appears to be trend-free. The series, however, exhibits seasonality due to the rainfall values in the months of the rainy season (June, July and August) observed high and low in the months of dry seasons (Example: January, February,...and so on).

To examine the presence of seasonal behavior in monthly rainfall series, we plotted a Autocorrelation Function (ACF) of the original series. ACF is a graphical representation of serial correlation coefficients (r_k) as function of lag k in which the values of r_k are plotted against respective value of k. The resulting oscillating shape of the autocorrelation function plot shown in Figure 2 confirms the presence of seasonal component in the monthly rainfall as well. Further, the ACF plot has peaks at lags equal to 12 and its multiples.

The application of the Box-Jenkins methodology in building a SARIMA model requires that the series must be stationary. Therefore, we started the process with testing the series for stationarity using Augmented Dickey-Fuller and variance comparison in addition to the simple inspection of

the data and autocorrelation function plots. The graphical representation of the series in Figure 1 exhibits strong seasonal variation and Autocorrelation Function plot in Figure 2 showed sinusoidal wave with high peaks at 12 and its multiples. The plots thus, suggest that the monthly rainfall series is seasonally non-stationary. We adjusted for seasonality by first seasonally differencing the series (i.e., each observation is replaced by the difference between it and the observation a year before) in the analysis. However, due to the absence of a clear apparent trend regular differencing was not suggestive. In order to confirm these results, we performed statistical tests of stationarity.

The Augmented Dickey-Fuller and Variance Comparison test results are found in support of the graphical analysis that the monthly rainfall series is stationary about its mean and no need of regular differencing. The Autocorrelation Function (ACF) in Figure 4a and Partial Auto-correlation (PACF) in Figure 4b of the first seasonally differenced monthly rainfall series showed the behavior of decaying rapidly for both seasonally and non seasonal correlation coefficients of the series indicating that stationarity is achieved after first seasonal differencing. Figure 3 exhibits seasonal variation free series that fluctuates almost around zero mean and constant variance.

Once the first seasonally differenced monthly rainfall series becomes stationary, next we dealt with testing its non-randomness applying graphical, brattlet's band and Ljung-Box tests. All these tests are found to be in support of the underlying stationary series is non random (i.e. serial dependence is there).

Having the stationary and non-random series in place, it is clear that we can develop a SARIMA model using Box-Jenkins methodology of model identification, estimation and diagnostic checks. In order to Identify the number of AR, SAR, MA and SMA terms in the SARIMA $(p, 0, q) \times (P, 1, Q)$ model, we look for significant sample Autocorrelations and Partial Autocorrelation Function coefficients employing the plots in Figure 4 and Table 2. The autocorrelation coefficients at lag 1, 12 and 48 as indicated in the Autocorrelation Function (ACF) plot in Figure 4a seem to be statistically significant.

For seasonal series model identification of the type $(0, d, q) \times (0, D, Q)_{12}$ using Autocorrelation function, (Hieple *et al.*, 1977) demonstrates that the autocorrelation coefficients at lag k should be

truncated and is not significantly different from zero after lag $q+sQ$. The significance of at other lags multiples of s imply the incorporation of seasonal autoregressive terms. Hence, the ACF plot in Figure 4a suggested the form $(0, 0, 1) \times (0, 1, 4)_{12}$ model to be identified because the autocorrelation coefficients of lags after $q+SQ$ (1+12×4) were insignificant except at lag 12 that imply the seasonal auto regressive term to be included. As a result, the SARIMA $(0, 0, 1) \times (0, 1, 4)_{12}$ model become $(0, 0, 1) \times (1, 1, 4)_{12}$ at the first instance to be fitted as a tentative model. In addition, we identified another tentative SARIMA model of the form $(1, 0, 0) \times (4, 1, 1)_{12}$ using the characteristics of PACF shown in Figure 4b.

Results of the maximum likelihood parameter estimates and correlation matrix of estimates of the models are shown in Table 4 and Table 5, respectively. The P-values associated with each parameter estimates of (a) and (b) models are found to be less than $\alpha=0.05$, indicating that these parameters should be retained for both models. The correlation matrix of the parameter estimates for the respective models has confirmed that there is no parameter redundancy in either of the models. By incorporating a seasonal moving average term at lag 24 to the model $(1, 0, 0) \times (4, 1, 1)_{12}$, we obtained model (c) of the form shown in Table 4 with the principle of overfitting to check if the added parameter has a significant contribution to the model. However, the overfitted parameter estimation resulted in an estimate of the added parameter $\Theta_2 = 0.07$ which is not significant (P-value $=0.07 > \alpha=0.05$) without significant improvement in the measures of goodness of fit. Therefore, results of the estimation stage suggests a further test is important in order to select one of these competitive two models that meet the diagnostic checks via residuals analysis.

For this comparison, we analyzed the various residual analysis tools including tests randomness of residuals and normality tests. Figure 5 and Figure 6 show that there is no significant auto-correlation between residuals at different lags; scatter plot of residuals appeared to fluctuate randomly around zero with no obvious pattern as shown in Figure 7. Figure 5c showed that the residuals are normally distributed as well. The residual randomness test results based on Ljung-Box in Table 6a and Table 6b confirmed the graphical analysis of residual randomness for both models. Since the graphical inspection and Ljung-Box tests are supportive of model adequacy for

both models, and there is no firm evidence to drop one of these models at this stage. Thus, both models can adequately fit to the seasonally first differenced monthly rainfall data.

With the hold-out last monthly rainfall values from Sep2004 to Dec2009 for evaluating forecasting accuracy and model validation, a slight difference in forecasting accuracy between the two fitted models is noticed. The results of forecasting accuracy tests for the two seasonal models in Table 7 showed that the SARIMA $(0, 0, 1) \times (1, 1, 4)_{12}$ model has a better forecasting accuracy in almost of the measures of accuracy indicated. Further validation of the selected model for its forecasting reliability using the validation periods also assessed by graphical examination whether the forecasted values are close to the corresponding actual values, in particular, and if the pattern of observed values captured in general. This graphical analysis of model validation was found to hold reasonably well with this model as indicated in Figure 8b that represents the plot of periods of the model forecasting verification and future forecasted monthly rainfall series.

Since the more parsimonious SARIMA $(0, 0, 1) \times (1, 1, 4)_{12}$ model passed all the identification, estimation, diagnostic checks and forecasting procedures of SARIMA modeling, we provided the future point and interval forecasts of monthly rainfall values in Table 8 after the model reformulated in the form of difference equation in Eq.(38).

CHAPTER FIVE

CONCULUSIONS AND RECOMMENDATIONS

5.1 Conclusions

With an aim to investigate the practical applicability of the Univaraite Box-Jenkins methodology in forecasting model development based on the actual monthly rainfall data at Mekele and its surrounding area, the steps of model identification, estimation, diagnostic checks and forecasting are performed as recommended in vast literatures (example: Box-Jenkins, 1976, Markridkis *et al.*, 1998, and Hipel *et al.*, 1977) in this research paper. Moreover, this research mainly intended to forecast the future values of monthly rainfall with better accuracy so that decision makers gain insight for better preparations. Descriptive analysis, on the other hand, has also been discussed based on the historical data of the underlying variable.

Based on the overall results of the research, therefore, the following conclusions could be drawn:

i) Since SARIMA models are identified according the autocorrelation structure of the appropriate monthly rainfall stationary series, the estimation and diagnostic analysis results revealed that models' are adequately fitted to the historical data. In particular, the residual analysis, which is important for diagnostic checking confirmed that there is no violation of assumptions in relation to model adequacy. Further comparison based on the forecasting accuracy of the models is performed with the hold-out some rainfall values. The point forecast results showed a very closer match with the pattern of the actual data and better forecasting accuracy in validation period.

Therefore, the results of the research indicate that SARIMA model of Box-Jenkins methodology allows capturing more complex description of the seasonality, autocorrelation structure and non-stationariy of the series and appears to be suitable in forecasting the monthly rainfall values based on Mekele station.

ii) The forecasting results reveal that there is no tendency of decreasing or increasing pattern of monthly rainfall over the forecast period from January 2010 to September 2011.

5.2 Recommendations

i) The developed forecasting model is recommended to be used as input in decision–making.

ii) The 95% confidence limits for point forecast have shown high variability in distance among months. Moreover, the point forecasting percentage errors have also varied considerably from month to month. The reason for this may largely be due to the high variability noticed in the descriptive analysis results. However, the limitations in data quality (example, because of the missed values) in the data might be an issue worth paying attention to. There is no doubt that this methodology without any missing value or applying better method of estimating those missing values would improve the efficiency of parameter estimates and forecasting accuracy. In our study, we used the simple method of computing the two adjacent values to estimate the missing values. Nevertheless, applying more advanced methods for estimating missing values may improve the results. Therefore, the filtering technique of estimating missing for Box and Jenkins Methodology suggested by (Mahir and Al-Kahalh, 2008) is recommended for further improvement of our forecasting model.

iii) To bring real and dependable work to the ground consistent and free of errors data is un doubtfully needed. Thus, great attention should be given to maintain quality of meteorological data from human and instrumental errors.

References

Alamerew Belay and Eshetu Wencheko, 2009. Assessment of Local Climate in Addis Ababa. *Journal of Ethiopian Statistical Association*, Vol. **18**, 55-68.

Balek, J., 1977. Hydrology and Water Resources in Tropical Africa; Elsevier, Amsterdam.

Bartlett, M.S., 1946. On the Theoretical Specification of Sampling Properties of Auto correlated Time Series; *Journal of the Royal Statistical Society*, Vol. **27,** 41.

Box, G. E. and Jenkins, G.M., 1976. *Time Series Analysis: Forecasting and Control*; Holden day.

Box, G. E., Jenkins, G.M. and G. Reinsel, 1985. *Time Series Analysis: Forecasting and Control*; Prentice-Hill, Englewood Cliffs.

Bowerman, B. and R.O'connel, 1987. *Time Series and Forecasting*: An applied approach; Duxbury Press, Boston.

Brockwell, P.J. and Davis, R. A., 1996. *Introduction to Time Series and Forecasting*; Springer, New York.

Caldwell, J.G., 2006. The Box-Jenkins Forecasting Technique Posted at Internet websites http://www.foundationwebsite.org.

Chatfield, C., 1996. *The Analysis of Time Series*; Chapman & Hall, London.

Conway, D., 2000. Some Aspects of Climate Variability in the Northeast Ethiopian Highlands Wollo and Tigray. *SINET – Ethiopian Journal of Science,* Vol. **23,** 127-139.

Cromwell, J.B., Labys, W.C. and Terraza, M., 1994. *Univariate Tests for Time Series Models*; A Sage Publications, London.

De Gooije, G. and Hyndman, J.R., 2006. 25 Years time series forecasting; *International Journal of forecasting,* Vol.**22,** 643-647.

Diebold, F.X., Kilian, L., and Nerlove, M., 2006. Time Series Analysis; Working Paper No. 2006-01, University of Maryland.

Dulluer, J. W and Kavas, M.L., 1978. Stochastic Models for Monthly Rainfall Forecasting and Synthetic Generation; *Journal of Applied Meteorology*, Vol.**17,** 1528-1535.

Elliott, G., Rothenberg, T. J. and J.H. Stock, 1996. "Efficient Tests for an Autoregressive Unit Root", *Econometrica*, Vol. **64**, 813–836

Enders, W., 2004. *Applied Econometric Time Series*; John Wiley and Sons, Inc., New York.

Farnum, N. and L., Stanton,1989. *Quantitative Forecasting Methods*; PWS-Kent Publishing Company, Boston.

Falk, M., 2006. *A First Course on Time Series Analysis:* Examples with SAS; University of Wurzburg.

Gebrerufael Hailu, 2008. Groundwater Assessment and Modeling, Aynalem Wellfield, Mekele, Ethiopia; International Institute for Geo-Information Science and Earth Observation Enscheda (Unpublished M.Sc.Thesis), the Netherlands .

Greene, W.H., 2000. *Econometric Analysis*; Prentice-Hall International, Inc., New Jersey,USA.

Granger, C. and P. Newbold, 1986. *Forecasting Economic Time Series*; Academic Press, Sandiago.

Harvey, A., 1993. *Time Series Models*; Harvester Wheatear, London.

Hamilton, J., 1994. Time Series Analysis; Princeton University Press, New Jersey.

Hipel, K., W.McLeod, A.I and Lennox,W.C., 1977. Advances in box and Jenkins modeling; *Water Resources Research*, Vol.**13**, 567-572.

Hipel, K. W. and McLeod, A. I.,1994. Time series modeling of water resources and environmental systems; Elsvier, Amsterdam.

Hoff, C.J., 1983. *A Practical Guide to Box-Jenkins Forecasting*; Lifetime Learning Publications, California.

Hyndman, R. J. 2001. It's Time to Move from What to Why; *International Journal of Forecasting*, Vol.**17,** 571.

Kavas, M.and J.Dulluer, 1975. The Stochastic and Chronological Structure of Rainfall Sequence –Application to India; *Water resources Research Center No*. 57, Perdue University.

Kirkpatrick, R. C. and Gaynor, P E., 1994. *Introduction to Time Series Modeling and in Business and Economics*; McGraw-Hill, Inc., New York.

Lincare, E. and B. Geerts, 1997. Climates and Weather Explained; Rutledge, London.

Makridakis, S., Wheelwright, S.C. and Hyndman, R.J., 1998. *Forecasting methods and Application*; John Wiley & Sons Inc., New York:

Mahir, A. and Al-kahalh, A.M., 2008. Estimation of Missing Data by Using the Filtering Process in a Time Series Modeling; [Stat ME], Malaysia.

Mills, T., 1999. The Economic Modeling of Financial Time series; Cambridge University Press, Cambridge.

Montgomery, D.C. and Johnson, L.A, 1976. *Forecasting and Time Series Analysis*; New York, McGraw – Hill.

Nyssen, J., Vandenreyken, H., Poesen, J. and Moeyersons, J., 2005. Rainfall Erosivity and Variability in the Northern Ethiopian Highlands; *Journal of Hydrology*, Vol.**311,** 172-187.

Newbold, P., 1975. The principles of the Box and Jenkins approach; University of Nottingham, Pergamon Press.

Naill, P.E. and Momani M., 2009. Time Series Analysis Model for Rainfall Data in Jordan: A Case Study for Using Time Series Analysis; *American Journal of Environmental*, Vol. **5**, 599-600, Jeddah, Kingdom of Saudi Arabia.

Osman,M. and P. Sauerborn, 2002. A Preliminary Assessment of Characteristics and Long- term Variability of Rainfall in Ethiopia - Basis for Sustainable Land Use and Resource Management; International agricultural research for development (unpublished).

Pankratiz, A., 1983. *Forecasting with Univariate Box-Jenkins: Concepts and Cases*; John Wiley & Sons, Inc., New York.

Shumway, R.H. and Stoffer, D. S., 2000. *Time Series Analysis and Its Applications*; Springer, New York.

Tamiru, F., 2009. Impact Assessment of global climate change on Some Components of Hydrometeorology in Ethiopia; Kochi University of Technology (unpublished).

Tsakiris, G., 1998. Stochastic Modeling of Rainfall Occurrences in Continuous Time; *Hydrological Science Journal*, Vol. **21**, 112, Athens, Greece.

Thomopoulos, P. and Nick, T., 1980. *Applied Forecasting Methods*; Prentice-Hall, Inc.

Tsay, S. R., 2005. *Analysis of Financial Time Series*; John Wiley & Sons, Inc., Hoboken, New Jersey.

U.S. Library of Congress, 2005. *Ethiopia, Country Studies Handbook*, Retrieved July 17, 2006, from http://countrystudies.us/ethiopia.

Vandaele, W., 1983. *Applied Time Series and Box-Jenkins Models*; Academic Press, New York.

Wei, W.S., 1990. *Time Series Analysis;* Addison-Wesley Publishing Company, Inc., New York, USA.

Wing, H., Gabriel, B., and Ashbindu, S., 2008. Trends and Spatial Distribution of Annual and Seasonal Rainfall in Ethiopia; *International Journal of Climatology*, Published online in Wiley Inter Science www.interscience.wiley.com

Yevjevich,V., 1972. Stochastic Process in Hydrology; Water Resources Publications, Fort Collins.

Yilma Seleshi, W. Delleur, and R. Demarke, 1994. Sunspot Numbers as a Possible Indicator of Annual Rainfall at Addis Ababa, Ethiopia: *International Journal of Climatology*, Vol. **14**, 911-923.

Annex. Monthly Rainfall Data at Mekele Station

Year	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Total
1975	1.4	20.8	18.0	21.5	10.6	100.0	185.1	293.1	120.7	0.0	0.0	0.0	771.2
1976	3.4	6.0	0.0	41.7	28.1	17.5	180.1	243.2	49.6	2.1	3.2	0.0	570.9
1977	0.0	0.0	19.3	5.4	77.0	56.5	204.6	251.2	62.9	1.1	0.0	0.0	678.0
1978	0.0	0.0	22.3	0.0	0.6	32.6	180.8	203.6	5.6	0.0	0.0	0.5	445.8
1979	0.0	0.0	2.5	40.9	103.8	38.8	63.3	98.4	28.5	0.0	0.0	0.0	376.2
1980	0.0	83.1	0.0	112.3	43.5	42.1	206.2	413.2	17.7	0.0	0.0	0.0	918.1
1981	0.0	0.0	24.0	5.6	11.3	0.0	349.0	205.3	22.9	0.0	0.0	0.0	618.1
1982	0.0	24.4	45.2	70.8	9.3	8.0	192.6	207.1	31.2	0.0	0.0	0.0	588.6
1983	0.0	11.0	4.6	27.5	106.3	1.8	244.3	255.1	35.0	1.6	17.4	0.0	704.6
1984	0.0	0.0	14.2	2.2	0.0	9.7	117.6	78.9	44.7	0.0	25.9	0.0	293.2
1985	1.8	1.0	24.7	126.8	37.4	14.6	129.4	180.8	20.6	0.0	0.0	0.0	537.1
1986	3.5	2.7	42.0	68.1	61.2	42.2	199.2	176.0	130.3	32.0	0.0	0.0	757.2
1987	0.0	2.0	79.6	37.2	126.7	56.6	177.2	220.2	36.3	1.9	0.0	0.0	737.7
1988	0.0	29.3	0.0	10.1	37.6	6.7	380.3	394.9	59.0	0.0	0.0	0.0	917.9
1989	0.0	15.4	8.7	10.5	19.3	81.7	273.6	430.7	40.5	16.5	0.0	0.0	870.6
1990	0.0	7.7	4.3	5.2	9.1	40.5	235.8	323.5	34.8	33.9	0.0	0.0	639.1
1991	0.0	0.0	0.0	0.0	0.0	0.0	198.0	216.3	28.2	53.1	0.0	0.0	593.6
1992	8.7	2.1	38.3	1.0	30.7	6.2	140.7	233.1	1.3	2.1	54.4	8.3	526.9
1993	11.7	7.7	63.9	135.0	74.7	69.0	217.2	106.5	15.2	20.0	0.0	0.0	720.9
1994	0.0	5.3	0.4	43.8	0.8	67.6	147.9	317.8	70.1	0.0	1.8	2.0	859.5
1995	0.0	5.9	31.2	29.2	27.1	6.8	268.2	237.7	51.4	3.0	0.0	2.7	663.2
1996	1.4	0.0	59.5	12.5	92.2	47.9	109.2	224.0	7.1	0.0	31.4	1.1	586.3
1997	0.0	0.0	20.4	32.4	32.6	29.8	32.4	243.1	100.5	16.3	59.9	15.7	583.1
1998	10.0	1.2	0.0	10.6	22.0	48.0	289.0	318.8	31.8	22.0	0.0	0.0	753.4
1999	22.0	0.3	10.9	0.0	0.0	7.4	293.6	359.2	22.8	0.9	0.0	0.0	717.1
2000	0.0	0.0	0.0	10.4	24.6	5.4	201.4	182.0	15.8	2.2	10.3	3.5	455.4
2001	0.0	0.0	38.1	18.7	8.7	65.5	267.9	226.3	9.2	2.9	0.0	0.0	635.3
2002	12.9	0.0	35.5	4.2	23.0	60.8	95.5	208.6	28.0	0.0	0.0	0.3	443.6
2003	0.0	25.9	18.2	8.4	35.2	87.5	125.6	201.8	23.4	0.7	0.0	0.1	526.8
2004	7.4	3.7	35.2	20.5	7.1	25.4	64.3	221.1	1.4	3.1	0.8	0.2	390.2
2005	0.0	1.4	15.6	48.9	55.3	18.2	110.5	314.0	34.3	0.0	1.3	0.0	599.5
2006	0.0	0.0	31.3	117.6	46.3	38.1	187.1	298.9	23.6	12.0	0.0	0.7	755.6
2007	0.1	2.3	34.5	90.1	71.6	95.0	184.6	271.2	25.8	3.7	2.5	0.0	781.4
2008	13.2	0.0	47.0	62.6	97.0	151.0	182.0	243.5	28.0	2.4	4.5	1.6	832.8
2009	0.2	0.0	13.5	3.1	46.2	104.0	296.7	226.8	78.6	17.5	0.0	0.0	783.5

Source: National meteorological Agency, Addis Ababa, Ethiopia.

Declaration

I, the undersigned, declare that the thesis is my original work, has not been presented for degree in any university and all source materials used for the thesis have been duly acknowledged.

Name: Amaha Gebretsadikan Gebreegziabher

Signature .....

Date: 04/11/2010

Statistics Department, Science Faculty, Addis Ababa University.

The matter embodied in this thesis work has not been submitted earlier for award of any degree or diploma to the best of my knowledge and belief. This thesis is an authentic work carried out by Mr. Amaha Gebretsadikan under my guidance and it has been submitted for examination with my approval as a university advisor.

Name: Mahendra Kumar Sharma

Signature: M.K. Sharma

Date: 04/11/10