Impact of Atmospheric Tides on Climate Model

Francis Gachari

A thesis submitted in partial fulfillment for the degree of Doctor of Philosophy in Physics in the Jomo Kenyatta University of Agriculture and Technology.

2014

DECLARATION

This thesis is my original work and has not been presented for a degree in any other university.

Signature:…………………......………… Date: ………………......……………..

Francis Gachari

JKUAT, Kenya.

This thesis has been submitted with our approval as the University Supervisors

Signature:………………......………… Date: ………………......……………...

Prof. David M. Mulati

JKUAT, Kenya.

Signature:………………......………… Date: ………………......……………...

Dr. J.N. Mutuku

JKUAT, Kenya.

DEDICATION

To my wife Lucy, daughter Carol and son Eric.

ACKNOWLEDGEMENTS

I would like to single out the contribution by my Supervisors Prof. D. M. Mulati and Dr. J. N. Mutuku. I have learned a lot by working closely with them and it is through their guidance that this work has remained on course and has eventually reached a conclusion in good time. They remained patient with me especially at the early stage of this work when the progress was slower than expected and have been very resourceful when difficulties arose along the way and provided solutions as to what resource materials were needed.

I thank the following for providing data: the Kenya Meteorological Department (Nairobi, Kenya), the National Oceanic and Atmospheric Administration and the National Aeronautics and Space Administration (NASA), the Climate Research Unit of the University of East Anglia (UK), and the Solar Influences Data Analysis Center of the Royal Observatory of Belgium. I also thank the following for providing software: R Foundation for Statistical Computing at the National Centre for Atmospheric Research and Unidata for the Integrated Data Viewer at the Program Center, University Corporation for Atmospheric Research (UCAR).

TABLE OF CONTENT

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

LIST OF ABBREVIATIONS

AMO - Atlantic Multidecadal Oscillation

ASALs - Arid and Semi-arid Lands

CM12.3 - Climate Model 12.3. JKUAT, Nairobi.

CRU - Climate Research Unit .University of East Anglia, UK.

ECHAM - ECMWF Atmospheric Model.

ECMWF - European Center for Medium Weather Forecasting.

EDM - Exponential Distribution Model.

ENSO - El-Nino Southern Oscillation

GCM - General Climate Model.

GDP - Gross Domestic Product.

GEE - Generalized Estimating Equation.

GFDL - The Geophysical Fluid Dynamics Laboratory at NOAA. USA.

GISS - Goddard Institute for Space Studies, NASA, USA.

GLM - Generalized Linear Model.

GPCP - Global Precipitation Climatology Project (GPCP) World Data Center for Meteorology, Asheville.

ICTP - International Centre for Theoretical Physics, Abdus Salam

IPCC - Intergovermental Panel on Climate Change.

IRI - International Research Institute,

ITCZ - Inter-tropical Convergence Zone.

JJA - June July August

JKUAT - Jomo Kenyatta University of Agriculture and Technology.

KenMet - Kenya Meteorological Department, University of Nairobi

LOD - Length of Day.

MAM - March April May.

MAX-DOAS -Multi-AXis Differential Optical Absorption Spectrometer.

MI - Meteorological Institute, Goddard Space Flight Center, NASA.

MJOs - Madden–Julian oscillations.

MLE - Maximum likelihood Estimates.

Model SMS12.12 - Model Sunspots, Mld and Sdec of December 2012.

Model 12.3 - Model of March 2012.

Model CM13.1 - Model Climate Model of January 2013.

MP - Model Prediction

MP ECHAM - Max Planck ECMWF Atmospheric Model, Germany.

MPI - Max-Planck Institute of meteorology, Germany.

MRI - Meteorological Research Institute, Japan.

MRI-RCM - MRI regional climate model (RCM), Japan.

NASA - National Aeronautics and Space Administration, U.S.A.

NCAR - National Center for Atmospheric Research, Boulder Colorado, USA.

NCAR NCSV - National Center for Atmospheric Research NCSV.

NCC - National Climate Centre.

NOAA - National Oceanic and Atmospheric Administration.

OND - October November December.

PCMDI - Program for Climate Model Diagnosis and Inter-comparison.

PWC - Physics of Weather and Climate

QBO - Quasi-Biennial Oscillation.

QIC - Quasi-likelihood under the Independence model information Criterion.

R^2 - Marginal R-squared.

RCM - Regional Climate Model

SCIAMACHY - SCanning Imaging Absorption spectroMeter for Atmospheric CHartographY

SOI - Southern Oscillation Index

SRES - Special Report on Emission Scenarios.

TRMM - Tropical Rain Measuring Mission (NASA).

UKMO HADCM3 - UK Met Office Hadley Climate Model 3

UNDP - United Nations Development Programme.

ABSTRACT

The main factor that determines the weather and climate on the surface of the earth is the time variation of the position of the overhead sun. This single factor determines the time of the day or night, variation of earth’s surface temperature, prevailing wind direction and therefore precipitation, weather and climate. The locus of the overhead sun as described by the solar-declination from a reference point on the earth surface can be accurately calculated astronomically at all times. This makes it possible to predict most weather parameters, using weather and climate models. We have in this study used a second important factor to account for the natural climate variability as the time variation of the overhead moon as described in a similar manner by the lunar declination.

This study demonstrates that the presence of enhanced atmospheric tides resulting from lunar-solar geometry is a key factor when used to predict the temporal distribution of rainfall amounts. Solar and lunar declination values obtained from ephemeris available from National Aeronautics Space Agency (NASA) have been used to compute the relative magnitude and duration of the tidal effect in the atmosphere for the period 1959 to 2005 over Nairobi. The impact of the tidal effect has been assessed by statistical modeling of Kenya rainfall against the conventional climate variability indices such as the Southern Oscillation Index (SOI) and Quasi-Biennial Oscillation (QBO) as well as modeling against parameters derived from the tidal effect. We have found that while conventional variability indices provide a method to explain past variability, their values are unknown for the purpose of projection into the future. We have therefore in this study used statistical modeling technique to obtain future rainfall amounts with covariates and factors derived from the lunar-solar geometry. The main advantage of lunar-solar parameters is that their values can be calculated accurately at all times and have therefore been used to carry out a projection of monthly rainfall amounts in Kenya for the period 1901 to 2020. The statistical model reveals an increase in frequency and intensity of severe hydrology events for the period 2018 to 2020.

CHAPTER ONE

1.0 INTRODUCTION

Three important factors have combined to bring about a large increase in interest in atmospheric science. First, man's increasing industrial activity has brought into focus the problems of pollution and the possibility of artificial modification of the environment. Secondly, concern about world food resources in the face of rapidly increasing population has made us much more aware of the critical effect of fluctuations in climate, particularly in parts of the developing world. Thirdly, the use of computers has made available much more powerful tools for atmospheric research.

One major factor which determines the climate on the surface of the earth is the time variation of position of the overhead sun. This single factor determines the time of day and night, variation of surface temperature, prevailing wind direction, precipitation, weather and climate at any position of the earth surface. Luckily, the locus of this position can always be determined astronomically with accuracy at all times and makes it possible to predict most weather parameters.

When the expected weather and climatic condition is not forthcoming, one is forced to look outside this factor for the cause. An unexpected drought condition, as the one of years 1984 and 2004 or a spike in rainfall such as the El-Niño rainfall of 1997, are examples. In this country droughts have been more devastating than floods and therefore more emphasis has been placed on the predictability of droughts. Natural disasters disrupt people’s lives through displacements, destruction of livelihoods and property, deaths and injuries. Consequently they take back years of development thus posing a major challenge to the achievement of the goals that target alleviation of extreme poverty. In this section we begin by discussing the drought situation, first in the Sahel region and then in Kenya. The variation of climate in the Sahel has is seen to influence the drought situation in this country. A drought condition is determined by the proportion of the country occupied by the Sahel type of climate during the year. [Cook, 2011].

1.1 THE SAHEL DROUGHT

The Sahel drought is a series of historic droughts, beginning in at least the 17th century affecting the Sahel region, as in Figure 1.1, is a climate zone sandwiched between the African savanna grasslands to the south and the Sahara desert to the north across West and Central Africa.

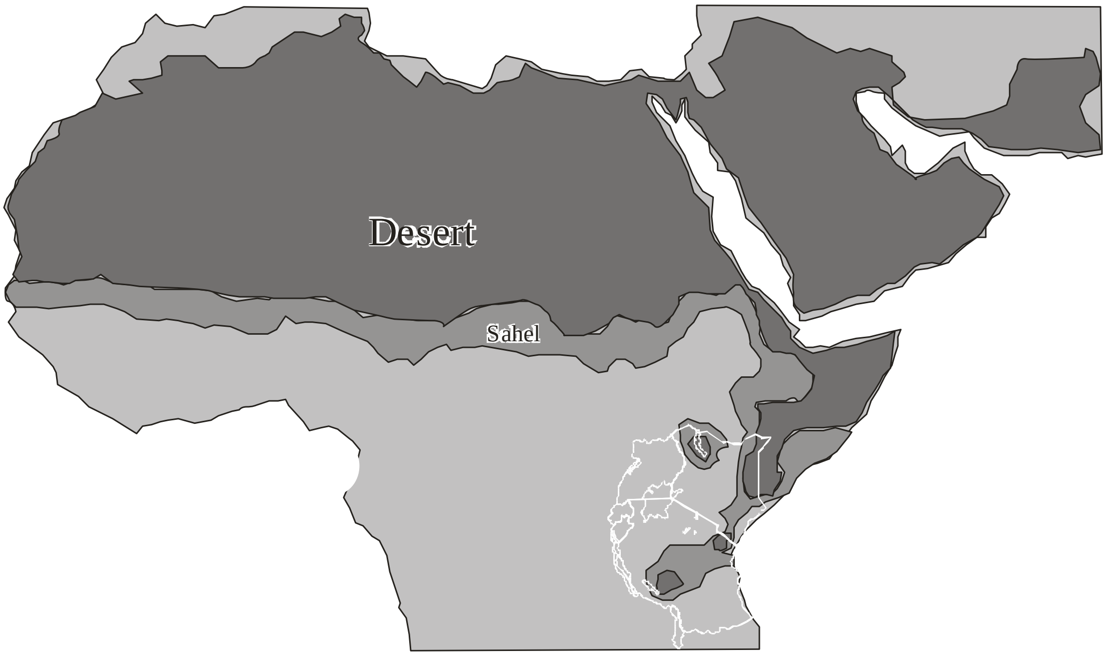
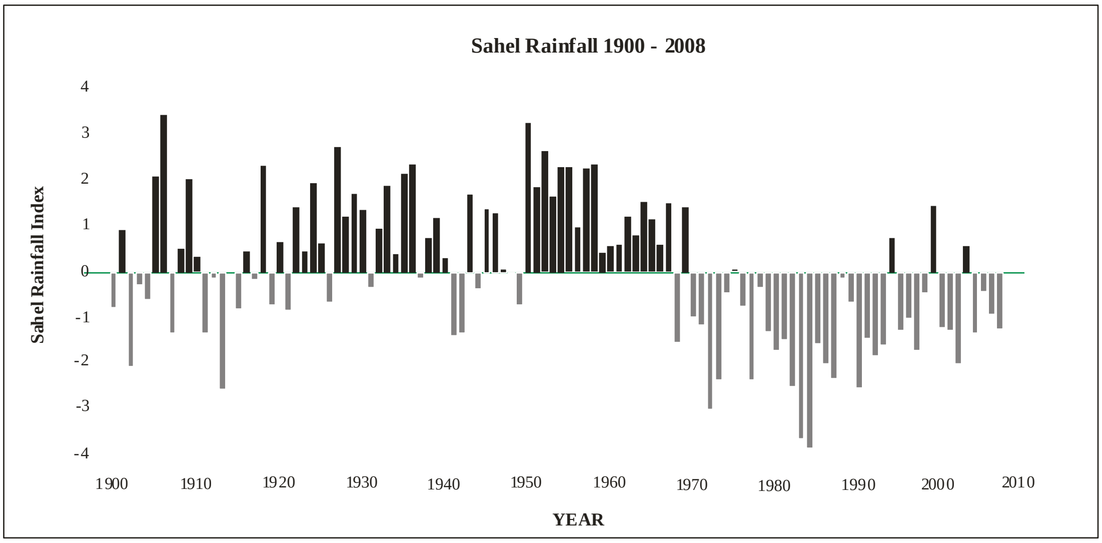


Figure 1.1
Africa: Climate Regions showing the Desert and the Sahel. [Cook, 2011]

While the frequency of drought in the region is thought to have increased from the end of the 1980s, three long droughts have had dramatic environmental and societal effects upon the Sahel nations. Famine followed severe droughts in the 1910s, the 1940s, and the 1960s, 1970s and 1980s, although a partial recovery occurred from 1975-80. While at least one particularly severe drought has been confirmed each century since the 17th century, the frequency and severity of recent Sahelian droughts stands out. In figure 1.2, severe conditions are taken to occur when the rainfall index is above 1.3 or below -1.3. [Wikipedia, March 2012].



<u>Figure 1.2</u>
Sahel rainfall variation in the twentieth century [wikipedia, Feb 2014].

Famine and dislocation on a massive scale from 1968 to 1974 and again in the early and mid 1980s—was blamed on two spikes in the severity of the late 1960-1980s drought period [Batterbury, 2001]. From the late 1960s to early 1980s famine killed at least 100,000 people, left 750,000 dependent on food aid, and affected most of the Sahel's 50 million people [UNEP, 2002]

The economies, agriculture, livestock and human populations of much of Mauritania, Mali, Chad, Niger and Burkina Faso were severely impacted. As disruptive as the droughts of the late 20th century were, evidence of past droughts recorded in Ghanaian lake sediments suggest that multi-decadal mega-droughts were common in

West Africa over the past 3,000 years and that several droughts lasted far longer and were far more severe, [Shanahan et al, 2009]

Because the Sahel's rainfall is heavily concentrated in a very short period of the year, the region has been prone to dislocation when droughts have occurred ever since agriculture developed around 5,000 years ago. The Sahel is marked by rainfalls of less than 100 mm a year, all of which occurs in a season which can run from several weeks to two months.

Despite this vulnerability, the history of drought and famine in the Sahel do not perfectly correlate. While modern scientific climate and rainfall studies have been able to identify trends and even specific periods of drought in the region, oral and written records over the last millennium do not record famine in all places at all times of drought. One 1997 study, in attempting to map long scale rainfall records to historical accounts of famine in Northern Nigeria, concluded that "the most disruptive historical famines occurred when the cumulative deficit of rainfall fell below 1.3 times the standard deviation of long-term mean annual rainfall for a particular place [Aondover and Woo, 1997]. Towards an Interpretation of Historical Droughts in Northern Nigeria. Climatic Change, no 37, 1997. pp.601-613 . The 1982-84 period, for instance, was particularly destructive to the pastoral Fula people of Senegal, Mali and Niger, and the Tuareg of northern Mali and Niger. The populations had not only suffered in the 1968-74 period, but the inability of many to rebuild herds destroyed a decade earlier, along with factors as various as the shift of political power to settled populations with independence in the 1960s, Senegalese-

Mauritanian border relations, and Niger's dependence upon falling world uranium prices coinciding in a destructive famine.

The first rain gauges in the Sahel date from 1898 and they reveal that a major drought, accompanied by large-scale famine, in the 1910s, followed by wet conditions during the 1920s and 1930s reaching a peak with the very wet year of 1936. The 1940s saw several minor droughts especially, in 1949 but the 1950s were consistently wet and expansion of agriculture to feed growing populations characterized this decade and many have thought it contributed to the severity of the subsequent Sahel droughts

Based on Senegal river cycles, precipitation cycles of various El Sahel stations which are related to Solar(89–120 years) Wolf-Gleissberg cycles, and on relations to Nile floods and Equatorial lake levels, Yousef and Ghilly anticipated that there is a considerable probability that drought will occur in El Sahel Zone in 2005±4 years. This forecast was correct as drought occurred in El Niger in 2005 and again in 2010. [Yousef and Ghilly, 2000].

In the early 2000s (decade), after the phenomenon of global dimming was discovered, it was speculatively suggested that the drought was likely caused by air pollution generated in Eurasia and North America. The pollution changed the properties of clouds over the Atlantic Ocean, disturbing the monsoons and shifting the tropical rains southwards.

In 2005, a series of climate modeling studies performed at NOAA/Geophysical Fluid Dynamics Laboratory indicated that the late 20th century Sahel drought was likely a climatic response to changing sea surface temperature patterns, and that it could be viewed as a combination of natural variability superimposed upon an anthropogenically forced regional drying trend.[Held, 2005]. Using GFDL CM2.X, these climate model simulations indicated that the general late 20th century Sahel drying trend was attributable to human-induced factors; largely due to an increase in greenhouse gases and partly due to an increase in atmospheric aerosols. In IPCC future scenario A2 (CO_2 value of ≈860 ppm) Sahel rainfall could be reduced by up to 25% by year 2100, according to climate models.

A 2006 study by NOAA scientists Zhang and Thomas suggests that the Atlantic Multidecadal Oscillation plays a leading role. An AMO warm phase strengthens the summer rainfall over Sahel, while a cold phase reduces it [Zhang and Thomas, 2006]. The AMO entered a warm phase in 1995 and, assuming a 70-year cycle (following peaks in 1880 and 1950), will peak around 2020.

So, what caused the Sahel drought and what has led to its recovery?

Analysis of the model results shows that in the 1980s, subsidence over the Sahel due to Indian ocean warming suppressed convection and the outflow blocked the flow of moisture from the Atlantic Ocean into the continent. The relatively warm Atlantic Ocean also contributed to the Sahel drought by competing for the moisture. But the Indian Ocean continues to warm so that the 1990s recovery was due to an increase of the scale of the Indian Ocean warming moved the subsidence to the Tropical Atlantic and led to the recovery, [Cook, 2011]. Thus the Sahel droughts were a result of natural variability. In this study we demonstrate that the Kenya droughts can be predicted by means of a climate model such as the one developed in this study.

1.2 KENYAN DROUGHTS AND FLOODS

The cyclic droughts and floods in Kenya have constantly eroded the recovery capacity of communities especially in the Sahel and Savannah type of climate. These areas are more commonly referred to as the Arid and Semi-arid Lands (ASALs). These natural disasters affect their economic development year in year out and require more vigorous attention and planning to mitigate the effects as they have impacted greatly on the country's fight against poverty and efforts to reduce the number of people living below the poverty line. The economic cost of the impact of floods, droughts and landslides in the past have been estimated in millions of shillings, [Paul, 2004].

Kenya's landscape is grouped into geographical zones including; the Savannah Lands covering most of the arid and semi- arid areas, the Coastal Margin, the Rift Valley, the Highlands and the Lake Victoria Basin [UNDP, 2004]. Kenyan population was reported as 38.6 million in 2009, compared to 28.7 million in 1999, 21.4 million in 1989 and 15.3 million 1979, an increase by a factor of 2.5 over 30 years, or an average growth of more than 3% per year [Wikipedia, 2012]. The population growth rate has been reported as somewhat reduced during the 2000s and was estimated at 2.7% in 2010, resulting in an estimate of a total population 41 million in 2011. The population is predominantly rural and relies on agricultural or other related activities for daily income although only 17% of the country's territory is arable.

The next 15 or 20 years are likely to see a rapid reduction in the rate of growth of Kenya’s population. Having been close to 4% per annum in the 1970’s (when it was widely claimed to be the highest in the world), by the year 2010 it was less than 2% and possibly under 1% if fertility fell as rapidly as envisaged in the “fast fertility decline” projections**.** [AFRICA ENVIRONMENT OUTLOOK, 2014].

Agriculture supports up to 75% of the Kenyan population including those who reside and work in urban centres, accounts for approximately one third of the Gross Domestic Product (GDP), employs more than two thirds of the labour force and about 70% of the export earnings [Kenyaweb, 2003]. It generates almost all the country’s food requirements and provides a significant proportion of raw materials for the agro-based industries. Overall, the smallholder sub-sector contributes about 75% of the country’s total value of agricultural output, 55% of the marketed agricultural output and just over 85% of total employment within agricultural sector. For this reason, it has a major role in the economy and consequently on the design of poverty eradication programmes.

Declining economic growth in general, coupled with a high population growth have lowered living standards and left sizeable numbers of the population poor and vulnerable to both natural and man-made disasters. The country’s geographical set up has also contributed much to regular if not permanent hazards in some areas. When these disasters interact with vulnerable communities they cause suffering of varying magnitudes. This has affected the economic development effectively lowering the human development of these areas.

Generally, natural hazards include drought, floods earthquakes, volcanic eruptions, landslides cyclones, storms etc. These occur all over the world and are, on their own, not harmful. However when these natural hazards interact with people, they are likely to cause damage of varying magnitude resulting in a disaster. Disasters thus occur when the natural hazards interact with vulnerable people, property, and livelihoods causing varying damage depending on the level of vulnerability of the individual, group, property or livelihoods.

The impact of such natural hazards is compounded by poverty and lack of adequate resources to develop the affected areas rendering the populations becoming more vulnerable. There is need to take up a proactive strategy in the management of natural disasters in Kenya, which would improve the coping capacity of communities, lessen the impact and therefore improve the lives of Kenyans in the areas prone to harsh weather conditions. A clear perspective on future rainfall situation inevitably necessary to set the pace for development programmes aimed at mitigating the impact of natural hazards. We should all be committed towards improving the lives of communities in Kenya and our hope is that this study will provide a useful tool towards this goal.

Drought, the most prevalent natural hazard in Kenya affects mainly Eastern, North Eastern, parts of Rift Valley and coast Provinces. Floods seasonally affect various parts of the country especially along the flood plains in the Lake Victoria basin and in Tana river while landslides are experienced during the long rains season running from March to May especially in Murang’a county and areas surrounding the Mount Kenya region. See Tables 1.1 and 1.2.

Table 1.1

Kenya Hazard areas

Drought prone provinces	Eastern, North Eastern, coast, parts of Rift Valley
Flood prone areas	Budalangi, Nyando, Rachuonyo, Tanariver
Landslide prone zones	Muranga county, parts of Kiambu, Thika, Maragua, Nyeri, Kirinyaga, Nyandarua and areas around mount Kenya region

The country covers a total area of 582,644 square kilometers of which less than 3% of the total is forest. 75% of Kenya’s population earns its living from agriculture which in turn depends on rainfall. Due to the vast areas prone to drought, Kenya’s vulnerability to food insecurity is highest among the pastoralists and small-scale agriculturalists in ASALs of the country. Extreme weather and climate events influence the entire economy, which depends mostly on agricultural products like cash crops, food crops and animals. Arid and semi arid lands carry 30 % of the country’s total human population yet they are characterized by uncertainty of rainfall, high evapo-transpiration rates, low organic matter levels and poor infrastructure.

Kenya has in the past recorded deficits of food due to drought resulting from a shortfall in rainfall in 1928, 1933-34, 1937, 1939, 1942-44, 1947, 1951, 1952-55, 1957-58, 1984-85, and 1999-2000. The 1983-84 drought and the 1999-2000 ones are recorded as the most severe resulting in loss of human life and livestock, heavy government expenditure to facilitate response and general high economic losses of unprecedented levels. After the El Nino induced rains of 1997 and 1998 Kenya experienced prolonged drought in many areas leading to famine and starvation [UNDP, 2004].

There are two rainy seasons in Kenya, the long rains in March to May and the short ones in October to December. The extreme climate and weather conditions are associated with anomalies in the general circulations of the seasonal northward and southward movement of the Inter-tropical Convergence Zone (ITCZ).

The ASAL areas in Kenya are categorized as follows. 11 regions are classified as arid, 19 as semi arid and 6 as those with high annual rainfall but with "pockets" of arid and semi-arid conditions. This gives a total of 36 areas. The various droughts that occurred in Kenya since 1883 and their characteristics are shown in Table 1.2 [Gathara, 1995], [Republic of Kenya, 2004].

<u>Table 1.2</u>

Chronology of drought and Floods incidences in Kenya.

Droughts		
Year	Region	Characteristics
1883-84	Coast	Worst famine in 30 years
1889-1890	Coast	One year of drought and famine
1894-1895	Coast	Information not available
1896-1900	Countrywide	Failure of three consecutive rainy seasons, human deaths
1907-1911	Lake Victoria, Machakos, Kitui and Coastal, Eastern and coastal provinces	Minor food shortages
1913-1919	Coastal areas	Impacts exacerbated by warfare
1921	Rift valley Central and Coast	A record dry year at the coast Local food shortages, crop and livestock losses (50% in Baringo).
1925	Northern Rift Valley and central provinces	Heavy loss of livestock, Lorian swamp dried up; deaths occurred
1938-1939	Countrywide	Food shortages, about 200 deaths
1942-1944	Central and Coast Provinces Eastern, central, Coast	Very severe drought in Coast Province

1947-1950	Nyanza, western and rift valley provinces	Mombasa reported driest, water shortages in Nairobi
1952-1955	Eastern, south/north rift Valley	Droughts followed by floods, cattle mortality at about 70-80 % in Maasai land.
1960-1961	Widespread	Rains of about 50% long-term mean, Nairobi hit by water shortage. Wildlife deaths in Nairobi national park.
1972	Most of Kenya	Human and livestock deaths in the northern counties Maasai cattle losses of about80%
1973-1974	Eastern Central, northern provinces	Crop production paralyzed. 16,000 people affected.
1974-1976	Central, Eastern, Western, coast	Famine in eastern province Water shortages, migration of people and livestock
1977	Widespread	20,000 people affected.
1980	Eastern province	Large food deficits. 40,000 people affected.
1981	Countrywide	Severe food shortages in Eastern province, less in central province
1983/84	Central, Rift Valley, Eastern and North Eastern	Moderately Severe in Eastern Province, Relief food imported. 200,000 people affected.
1987	Eastern and Central	4.7 million people dependent on relief power and water rationing
1991-94	Arid and semi-Arid Areas of NE, Rift Valley, Eastern and Central, Coast Provinces	1.5 Million people affected
1995/96	Widespread	1.41 Million people affected
1999-2000	Countrywide except west and coastal belt.	4.4 Million people affected
2004	Widespread	2-3 Million people affected

Floods

2002	Meru Central, Muranga, Nandi	2,000 people affected
2002	Nyanza, Busia, Tana river basin	150,000 people affected
1997/1998	Widespread (El Nino Floods)	1.5 million people affected
1985	Nyanza and Western	10,000 people affected
1982	Nyanza	4,000

The impact of disasters can either push more people below the poverty line or impoverish further the existing poor people due to injuries, displacements, destruction of property and livelihoods among other effects. Most communities in the Kenyan arid and semiarid lands depend on pastoralism and agriculture for survival. These economic activities in turn depend on rainfall for water and pasture.

In Kenya the economic parameters that affect the severity of drought making the communities more susceptible to drought and famine are rise in food prices, fall in animals prices, depletion of food reserves without replacement, deterioration of health due to lack of food and clean water among other issues.

Poor infrastructure including impassable roads, poor telecommunication lines and inaccessibility of some regions hampers the transportation of food to these regions either for commercial purposes or relief aid. Poor communication also hampers action in terms of response to distress calls, poor publicity and inability to air the plight of the people.

Once the effects of the drought begin to be felt the health of animals begins to deteriorate due to inadequate pasture and water. The animals also experience Tsetse flies infestation and foot and mouth disease, which are common in drought conditions. This requires use of veterinary medicines, which are expensive and sometimes not accessible to the pastoralists.

1.3 THE DROUGHT AND THE SOLAR ECLIPSE

During the great Sahel drought period of the 80’s (1981-1986) only a single eclipse was observable in Kenya over a period of 20 years (1981-2000), the one that occurred on December, 4 1983. The path of the eclipse coincided with the areas much ravaged by the drought, the Northern Kenya and the Turkana corridor as in Figure 1.3.

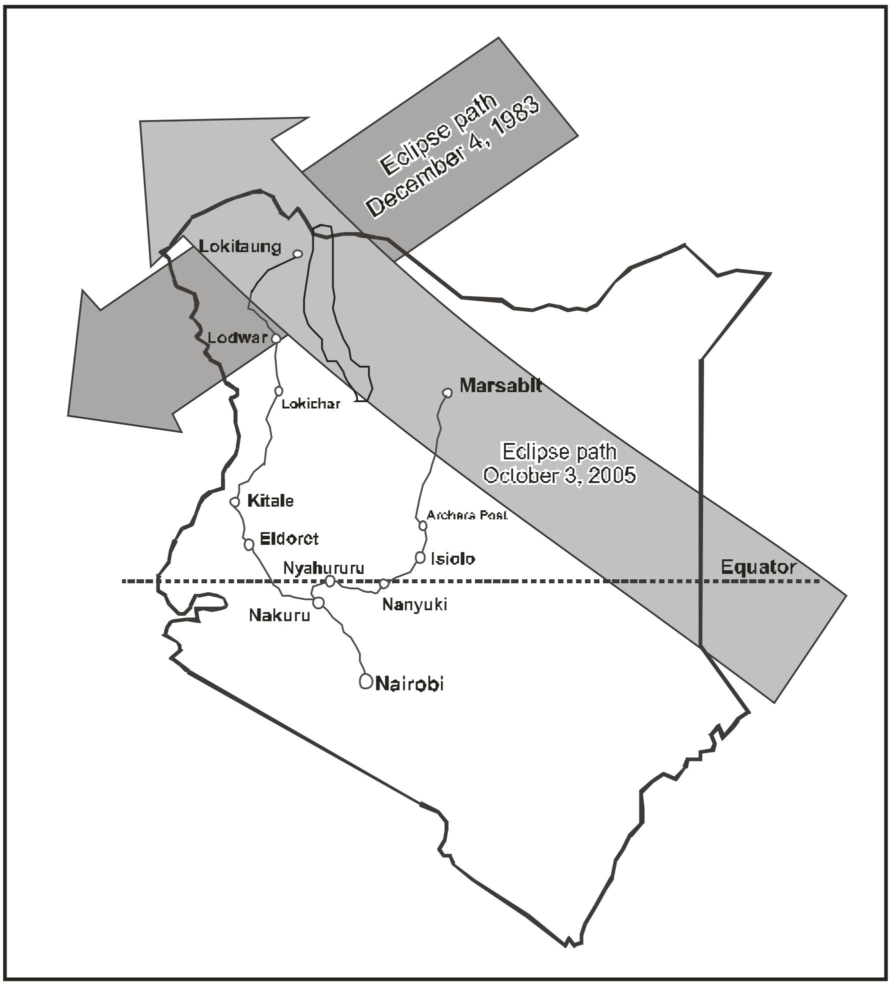


Figure 1.3
Eclipse paths during the solar eclipse on December 4, 1983 and October 3, 2005

Scientists documenting the eclipse took a flight from Nairobi to Kitale and travelled by road on December 4, 1983 through Lodwar, Lokichar to the lakeside town of lokitaung to arrive before 12.31 pm the time for the greatest eclipse. We were unable to travel to Lokitaung so we set up a site at Nyahururu and made observations from there although our site was outside the path a shown in Figure 3.9 of Chapter 3.

It was a partial eclipse but because of the rarity of the event local and foreign scientists flocked the area to make real time observations. No similar opportunity was expected in this country until October 3, 2005. Not only did they observe the eclipse but also the devastating effect of greatest Sahelian drought of the twentieth century. But alas! In 2005 scientist flew to Nanyuki and travelled by road through Isiolo, Archers Post to arrive in Marsabit before 10.30 am., in the time for greatest eclipse. On the way again they found the area still ravanged by another drought; the 2002-2004 droughts similar to the one that had happened twenty years before. Both the towns of Lokitaung and Marsabit lie within the region affected by the Sahelian drought. This observation prompted the question as to whether the drought and eclipse episodes were mere coincidences. The Sahel region did not fully recover from the drought state until 1990. Thereafter it underwent droughts in 1998 and 2002-2004. A study carried out by Cook, [Cook, 2011] shows that the Sahel region as a whole seems to have recovered from the drought after 1990. Even then, another less severe drought has devastated this country in 2011. The predictability of severe hydrology events therefore remains a challenge in this country. That is why it is the subject of this study.

In 2005 we began a study on measurement of atmospheric water content over Nairobi by means of both a ground based MAX-DOAS Spectrometer and SCIAMACHY a satellite based spectrometer. Due to the prevailing drought situation, we found it necessary to take the opportunity and address the drought issue. We responded by investigating whether a relationship did exist between atmospheric water vapour content of the atmosphere, the drought and the occurrence of an eclipse. It was felt that perhaps an eclipse condition led to occurrence of a drought by influencing the amount of water vapour in the atmosphere. However, by the end of the study in 2009, we found that while atmospheric water vapour content oscillates between 5g/cm^2 and 3g/cm^2 depending on the season, no relationship between water vapour content and the occurrence of the eclipse was identified.. Even then, the drought was suspected to have resulted from enhanced atmospheric tides resulting from the superposition of solar and lunar atmospheric tides as happens during a solar eclipse. We also found no relationship between the occurrence of a solar eclipse and the rainfall distribution, but it was pointed out that the appearance of atmospheric tides during a solar eclipse could indeed affect the expected rainfall amount, [Gachari, 2008]

1.4 THE CLIMATE MODEL

The task to establish a relationship between atmospheric tides and rainfall variability began in this study in 2009. First steps involved obtaining periods when gravity atmospheric tides occur during the cause of the year. This was accomplished by obtaining time variation of the solar and lunar declination angles relative to Nairobi. Like ocean tides, gravity atmospheric tides are dependent on the solar-lunar geometry. However, unlike ocean tides, air is more elastic and therefore surface pressure changes due to a solar atmospheric tide are less than 0.3 mmHg. In this study we have defined two atmospheric tidal states: the ordinary atmospheric tidal state, *atide* and the enhanced atmospheric tide state, *etide.* The enhanced tidal state coincides with the occurrence of the solar eclipse. The location of the tidal state is determined by the angular amplitude of the lunar declination. The lunar angular velocity as observed from the earth's surface is smaller when the moon orbits the earth within the 23.5 degrees limits. The relative maximum lunar angular declinations at various months are defined in this study and allocated numerical values of the factor referred to the maximum lunar declination factor, *mld* with values ranging from -28.5 to 28.5 degrees. We have used a fitting procedure to reproduce the rainfall variability pattern for the period 1901 to 2000. Using this pattern we have projected rainfall for the period 1901 to 2020. This has been achieved by obtaining a statistical rainfall model using rain-gauge data taken both at Dagorreti and Jomo Kenyatta Airport meteorological stations since 1959 together with factors and covariates derived from the solar and lunar geometry. The task was completed by March 2012 and the model named the CM12.3 Model output has since undergone evaluation and found to be consistent with observations. The fact that a model based purely on astronomical variables describing the spatial-temporal variation of atmospheric tide can be used to reproduce the rainfall variability is evidence that atmospheric tides play a key role in determining rainfall distribution on the earth's surface in general. The model may be extended to work out rainfall distribution in multiple sites while the current model results may be used to estimate when future floods and droughts are expected. One important observation from the model results is that heavy floods are expected in 2013 and 2016 while prolonged droughts will be back between 2019 and 2020.

1.5 PROBLEM STATEMENT

The worst of the droughts in Kenya's history occurred in 2004 and the country was still recovering from its devastating effects when this study commenced. To-date, no one knows when the next drought(s) will strike. It was widely thought that the drought had its origin in the solar eclipse that preceded it. Furthermore, another such drought had also occurred in 1984 after another much publicized eclipse in December 1983. Figure 1.40 shows rainfall amounts measured at Kenya Meteorological Department in Nairobi from 1959 to 2004 [KenMet]

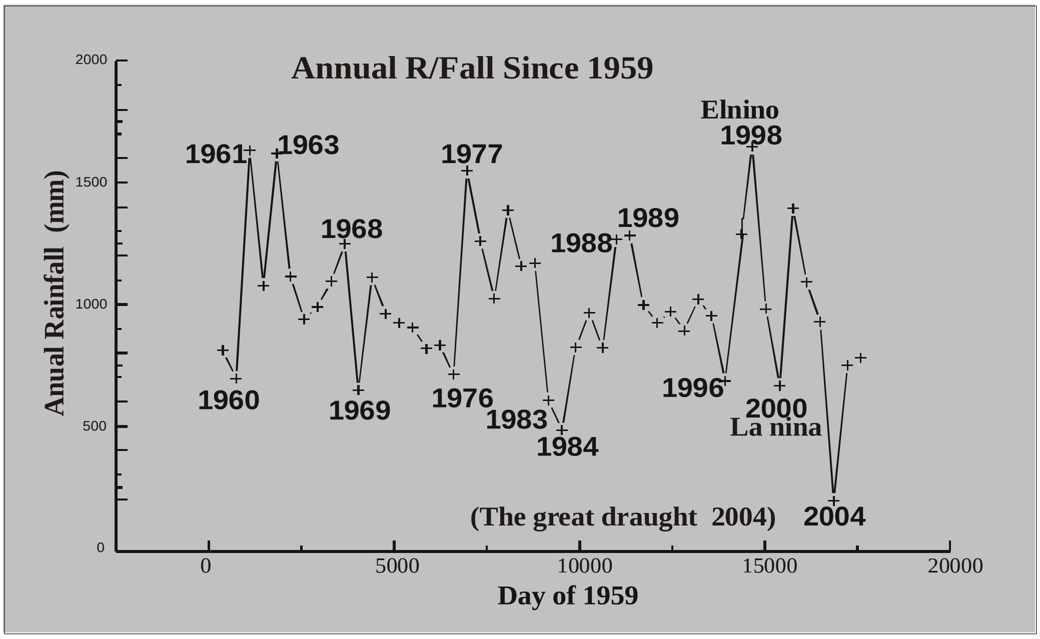


Figure 1.4
Annual rainfall since 1959. [KenMet]

The reduction in solar energy reaching the atmosphere during a solar eclipse was suspected to have lowered atmospheric temperature thereby causing a corresponding reduction in atmospheric water vapour content thereby leading to a drought situation. But, when rainfall amounts measured on each day together with the days on which a solar eclipse occurred were considered, it was observed that while the occurrence of a solar eclipse is a periodic event, a corresponding periodicity in the occurrence of a drought situation could not be identified. It was then observed that the gravitational attraction between the moon and the atmosphere could have affected the expected precipitation. Since then, it has been established from available literature that the presence of atmospheric tides does affect the atmospheric parameters. It is the purpose of this study to design an appropriate rainfall model based on astronomical factors but also to use the model in estimating future rainfall amounts.

1.6 Significance and Justification

The backbone of Kenya's economy is Agriculture and the equatorial location of the country together with her position on the African continent land mass gives the country a s*avannah* type of climate otherwise referred to as the Tropical Wet and Dry according to Köppen classification of world climates [Tarbuck and Lutgens, 1997]. The predictability of climate and weather is therefore critical for the success of rain-fed agriculture.

Natural science is based on the assumption that the natural world behaves in a consistent and predictable manner. It is also suspected that the spikes visible in the rainfall distribution were as a result of enhanced solar-lunar atmospheric tides which affect the usual flow of the prevailing air masses and have prolonged affect. The tides affect the direction and manner of the prevailing winds and therefore precipitation. They can therefore cause an un-expected shift in short-term weather conditions . It also explains why some solar/lunar tides may not affect precipitation – their locality may not be relevant to local weather. The dates and paths where the solar/lunar atmospheric tides are expected can accurately be calculated and documented and used in construction of a rainfall model.

1.7 Hypothesis

The droughts of 1984 and 2004 were as a result of enhanced atmospheric tides resulting from the prevailing Solar-Lunar geometries.

1.8 STUDY OBJECTIVES

The main objective of this study was to design a model which can then be used to estimate rainfall amounts in this country for the next decade. This study therefore aims at achieving the following objectives:

i. analyze eclipse, and precipitation data to establish their relevance to the hypothesis stated above.

ii. design a rainfall model based on astronomical and meteorological parameters.

iii. use the model to determine future rainfall trends.

iv. use model results to make appropriate recommendations for planning and disaster preparedness in this country.

CHAPTER 2

2.0 LITERATURE REVIEW

In this chapter we describe the factors of climate variability in Eastern Africa and then discuss how some of these factors have been used to design a statistical rainfall model based on fitting a generalized linear model of the Tweedie family.

2.1 SOLAR DECLINATION

Solar declination is the Sun-Target-Observer angle: the vertex angle at target center formed by a vector to the apparent center of the Sun and a vector intersecting the observer. This measurable angle is within 20 arc-seconds (0.006 deg) of the reduced phase at observer's location. The difference is due to down-leg stellar aberration affecting measured target position but not apparent solar illumination direction. The phase angles were obtained from the ephemerides [Solar Ephemeries, 2011].

The season corresponds to the solar declination periods. As can be seen in Figure 2.1, solar declination is a periodically stable cycle.

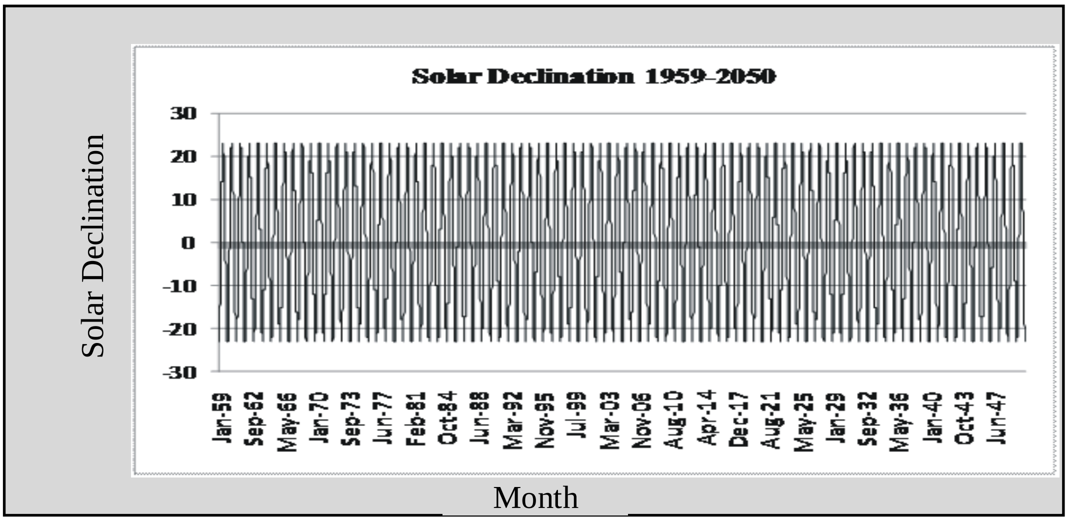


Figure 2.1
Solar Declination 1959 to 2050 temporal variation

Oscillations of the position of the sun relative to Nairobi are uniform on the inter-annual scale.

2.1.1 The Eastern African Monsoon and the Solar Declination

The basic principle behind the monsoon is similar to that of the sea breeze: differential heating over land and sea. During the summer monsoon (Jun –Aug), air over the continents warms and ascends, and moist, colder air flows in from the ocean bringing heavy rain. The latent heat release and continuous solar insolation stabilize the circulation which can continue for months. The winter monsoon (Dec-Feb) has the same origin but opposite direction as the tropical sea is warmer in winter and brings draught rather than rain [Wikipedia, 2012]. Monsoon patterns extend over East Africa, Arabia, India, and the Arabian Sea as seen in Figure 2.2 below.

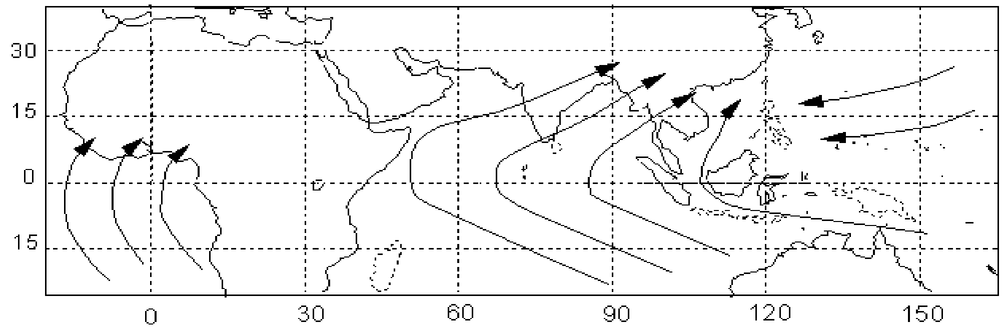


Figure 2.2 (a)
Surface winds during Jun-Aug [Richter A., 2004]

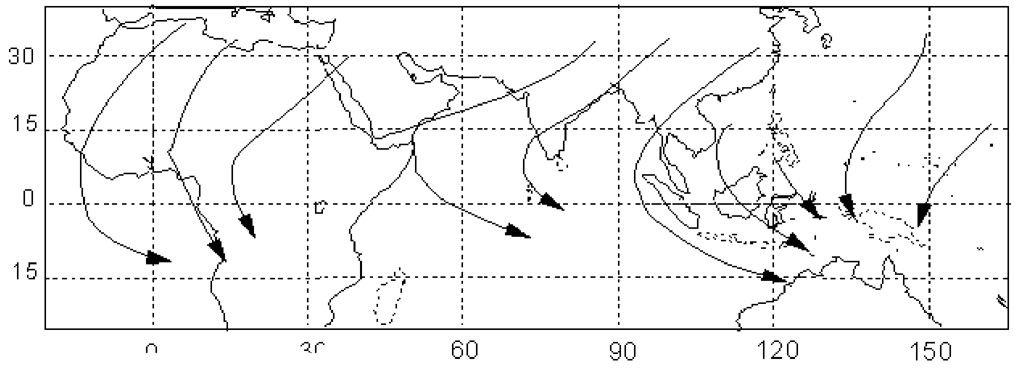


Figure 2.2 (b)
Surface winds during Dec-Feb [Richter A., 2004]

Similar but weaker patterns can also be found in equatorial America. Summer monsoon over India is a result of the low pressure zone over the Asian highlands that move North with the sun, leading to monsoon patterns appearing first in Sri Lanka end of May and moving to the Himalayas by July. They lead to the highest rainfall values observed anywhere on earth. The monsoon is therefore dependent on the season which is in turn determined by the solar declination. Solar declination is a major factor in the rainfall model developed in this study.

2.1.2 General Circulation. The Hadley Cell

The Hadley cell is a circulation pattern that dominates the tropical atmosphere, with rising motion near the equator, pole-ward flow 10-15 kilometers above the surface, descending motion in the subtropics, and equator-ward flow near the surface. This circulation is intimately related to the trade winds, tropical rain-belts, subtropical deserts and the jet streams as seen Figure 2.3 below.

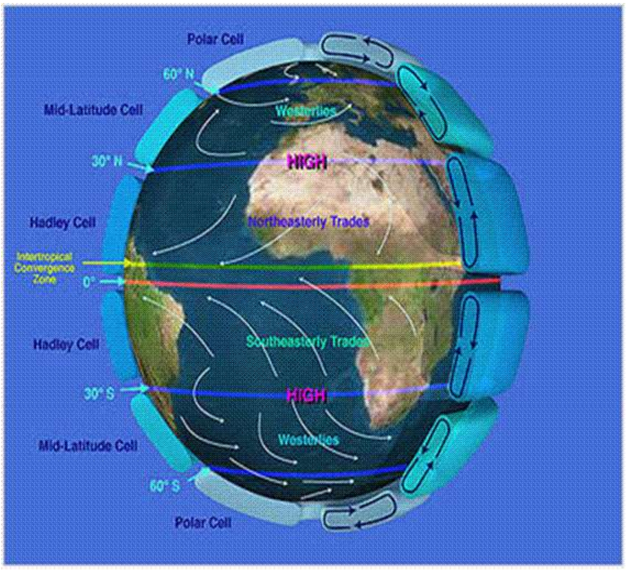


Figure 2.3
Hadley cells in the atmosphere [Wikipedia, October 2011].

Seasonal convective rainfall in the tropics is associated with the ITCZ. The Hadley cells also explain the location of the tropical high pressure zones at 30 degrees north and south of equator as well the zonal and meridianal winds direction. The presence of an overhead moon affects the shape and the height of the Hadley cells as well as the location of the ITCZ. This affects the local precipitation. In this study the incidence where a location is influenced by the presence of the lunar-solar gravitational tide is referred to as the *atide* state or enhanced atmospheric tide state *etide*. An upwards displaced atmosphere due to atmospheric tides may translate the high pressure zone equator-wards leading to lower Sahel rainfall amounts.

2.1.3 Solar declination and precipitation

Wet season occurs when solar declination is increasing with time either -2 to 20 March-April-May (MAM) or decreasing with time -10 to -23 then increasing -23 to -20.October-November-December (OND) as seen in Figure 2.4 below.

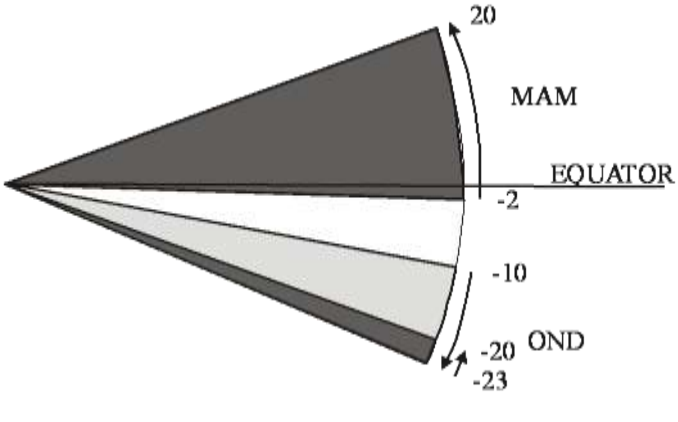


Figure 2.4
Solar angle Restriction of Nairobi Rainfall.
The larger Sector for MAM accounts for longer rains season and the proximity of the sector to the Equator accounts for the larger amounts in MAM.

The sun is at the equator on March 22 and September 22. In both cases the solar declination is zero. Although the solar declination determines the two seasons, other factor come into play to vary the amount of rainfall in the season. Months fail to deliver the expected rainfall creating inter-annual rainfall variability evident in Figure 2.5 below.

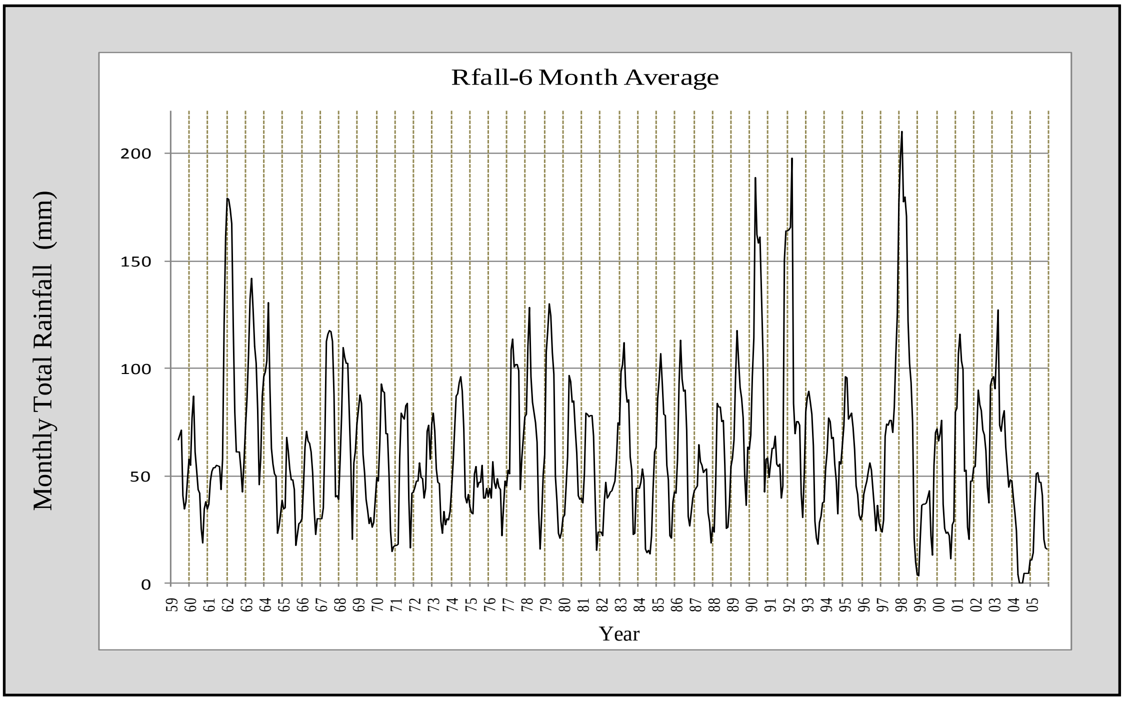


Figure 2.5
Six month average clearly showing inter-annual variation of Nairobi Rainfall.

2.1.4 Lunar Declination and Standstills.

On the inter-annual scale the latitudinal extremities of lunar position oscillates with a wavelength of about 19 years. Standstill periods (ldec>28.5) are therefore separated by a period of about 19 years as shown in Figure 2.6. It can also be seen that episodes of the same lunar declination amplitudes are separated by the same period.

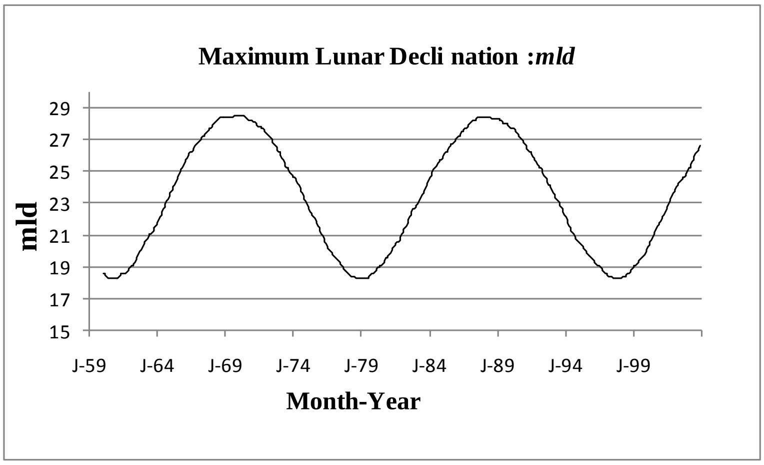


Figure 2.6
Lunar Declination temporal variation

At a lunar standstill, which takes place every 18.6 years, the range of the declination of the Moon reaches a maximum. As a result, at high latitudes, the Moon appears to move in just two weeks from high in the sky to low on the horizon. The Moon changes in declination, but it does so in only a month, instead of a year for the Sun. So it might go from a declination of +28.5° to −28.5° in just two weeks, returning to +28.5° two weeks later. Thus, in just one month the moon can move from being high in the sky, to low on the horizon, and back again. as seen Figure 2.7 below.

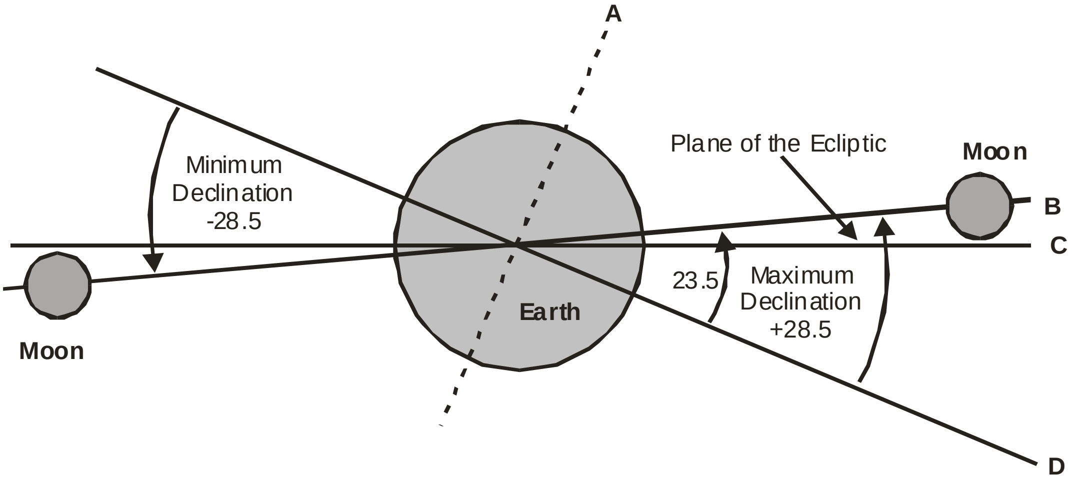


Figure 2.7
A Major Standstill

This is because the plane of the Moon's orbit around the Earth is inclined by about 5° to the plane of the Earth's orbit around the Sun, and the direction of this inclination gradually changes over an 18.6-year cycle, alternately working "with" and "against" the 23.5° tilt of the Earth's axis. As a consequence, the maximum declination of the Moon varies from (23.5° − 5°) = 18.5° to (23.5° + 5°) = 28.5°. The effect of this is that at one particular time (the minor lunar standstill), the Moon will change its declination during the month from +18.5° to −18.5°, which is a total movement of 37°. This is not a particularly big change, and may not be very noticeable in the sky. However, 9.3 years later, during the major lunar standstill, the moon will change its declination during the month from +28.5° to −28.5°, which is a total movement of 57°, and which is enough to take its zenith from high in the sky to low on the horizon in just two weeks (half an orbit).

Strictly speaking, the lunar standstill is an instant in time: it does not persist over the two weeks that the Moon takes to move from its maximum (positive) declination to minimum (negative) declination, and it most likely will not exactly coincide with either extreme. However, because the 18.6-year cycle of standstills is so much longer than the Moon's orbital period, the change in the declination range over periods as short as half an orbit is very small.

During the standstill periods, the moons angular velocity increases and with it the attendant increase in the associated zonal wind velocities. This affects local atmospheric conditions. For the purpose of this study, the standstill state is the period when the lunar declination reaches beyond the 23.5 degrees.

Another consequence of the lunar motion is the earth-moon distance. Moon's orbit is elliptical and so the moon is closest to the earth when it is also collinear with the sun. This geometry generates strong atmospheric tides; the enhanced tide, *etide* factor.

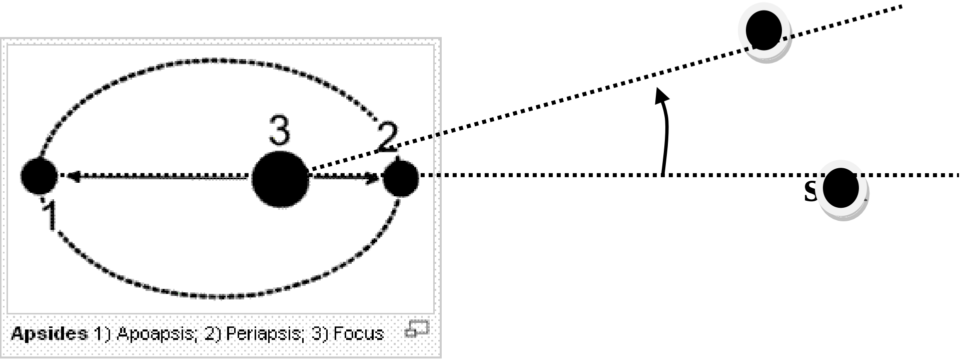


Figure 2.8
Lunar distance variation; 1) Apogee, 2) Perigee, 3) Focus (Earth).

Figure 2.8 shows how solar declination may be a measure of the magnitude of the atmospheric tide, having a maximum value when both values of solar and lunar angles equals zero.

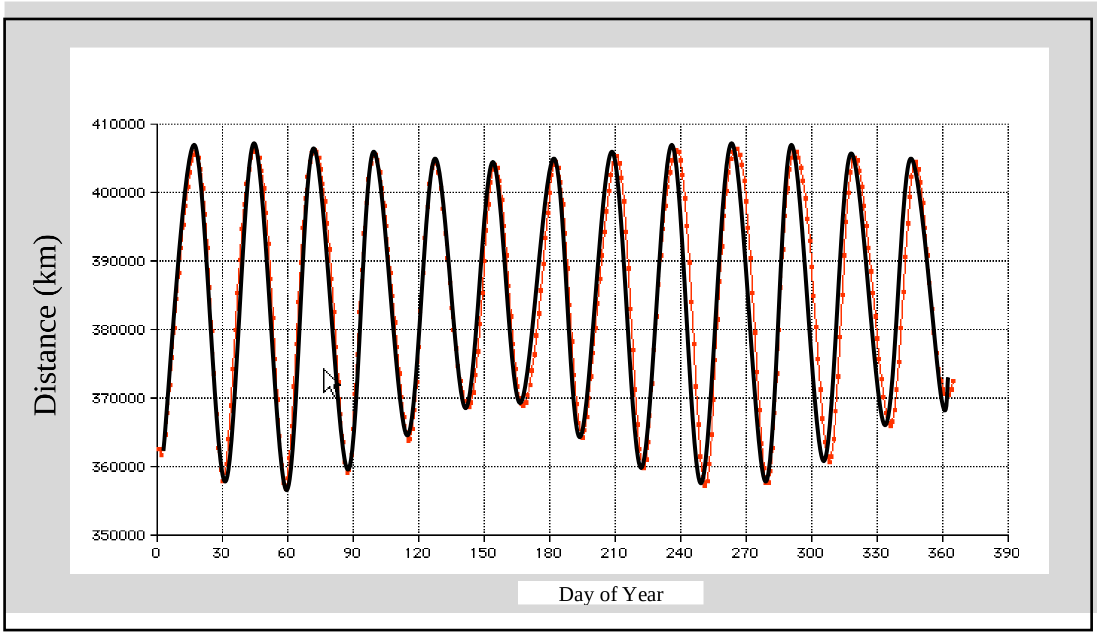


Figure 2.9
Lunar distance variation in Kilometers for 1926

The distance between the moon and the Earth varies from around 356,400km to 406,700km at the extreme perigees (closest) and apogees (farthest) as seen Figure 2.9. Enhanced atmospheric gravity tides are generated due to this linear geometry of the three bodies- earth, moon and the sun.

Enhanced Atmospheric tides, as happens with ocean tides, occur at the sublunar point and at its antinodal point at New Moon and Full Moon. Taking Earth's radius as 6371 km, Moon's mass as 7.349 x 10^{22} kg, mean distance Earth-Moon radius as 384401000 m, Sun's mass M to be 1.989 x 10^{30} kg and the mean distance Sun-Earth as 1.496 x 10^{11} m, the axial tidal acceleration a by the Moon, aMoon and that by the sun, aSun respectively are approximately:

$$\text{aMoon} \approx 1.1 \text{ x } 10^{-6} \text{ m/s}^2$$

$$\text{aSun} \approx 0.50 \text{ x } 10^{-6} \text{ m/s}^2$$

The tidal forces of the Moon are reinforced by the Sun at New Moon and Full Moon so that

$$a \approx \text{aMoon} + \text{aSun} \approx (1.1 + 0.50) \text{ x } 10^{-6} \text{ m/s}^2 = 1.6 \text{ x } 10^{-6} \text{ m/s}^2$$

When the Moon is at first quarter or third quarter (Sun and Moon separated by 90° when viewed from the Earth) the solar tidal force partially cancels the Moon's:

$$a \approx \text{aMoon} - \text{aSun} \approx (1.1 - 0.50) \text{ x } 10^{-6} \text{ m/s}^2 = 0.60 \text{ x } 10^{-6} \text{ m/s}^2$$

Using the ecliptical geocentric longitudes of the Moon and the Sun (neglecting their declinations), the geocentric distances of the bodies the following quantities were calculated: the geocentric and topocentric distance of the Moon from the Earth (kilometers), the apparent angular size (arc minutes), the illuminated fraction of the Moon's disc (per cent), and the Moon's phase have been computed.

Moon phases are denoted as

0.00 = new Moon

0.25 = first quarter

0.50 = full Moon

0.75 = last quarter

1.00 = new Moon

Distances/altitude in kilometers is taking into account the horizontal parallax of the Moon are; Perigee: 356,375, and Apogee: 406,720.

The change of distance may be up to about 6,300 km per day and the mean perigee distance of 284 anomalistic months is 362562.4 km [NASA, 2009].

2.2 EL-NINO SOUTHERN OSCILLATION (ENSO)

In normal years, the Walker circulation is characterized by a low pressure system over the Western Pacific and high pressure over the Eastern Pacific. This leads to, easterly trade winds and upwelling of cold and nutrient rich waters off the coast of Peru. These trade winds “pile up” warm water in the Western Pacific (30 cm) resulting in strong convective activity, storm and precipitation over Indonesian region [Wendell, 2008].

During El Niño conditions, the pressure difference reduces and inverts, and trade winds weaken or invert therefore warm water from the Western Pacific flows back to the East within 2 months (Kelvin wave) creating an upwelling off Peru is interrupted and sea surface temperature (SST) increases. Convective activity moves with the warm water, leading to heavy rain fall at the West coast of SA and draught in the Western Pacific as seen Figure 2.10 below.

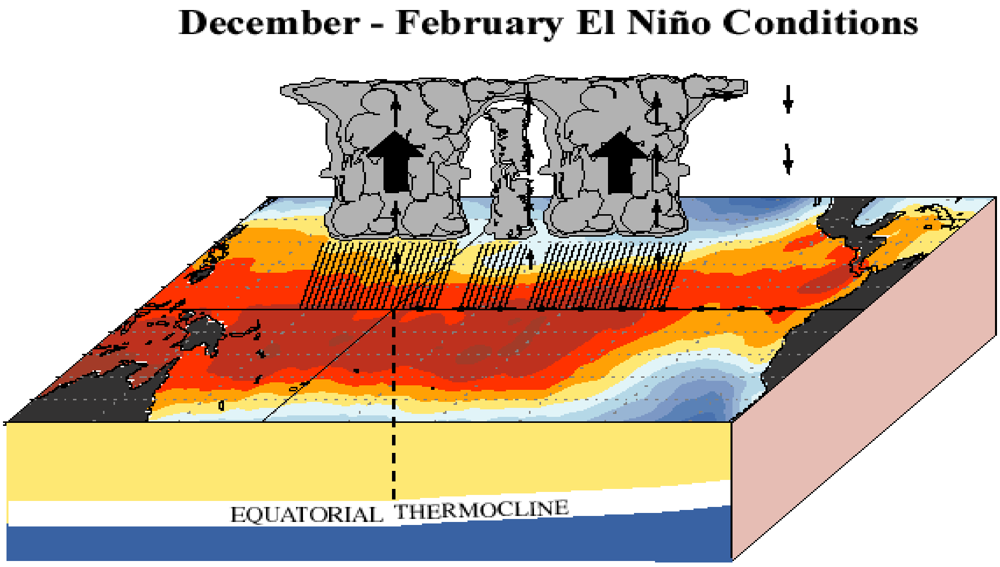


Figure 2.10
Model view of El-Niño Phenomenon [Wikipedia 2009]

The changes during an El Niño event have many effects on the ocean atmosphere system such as changing flow directions, increased storm frequencies in some regions and reduced land falling hurricanes in other regions. Many areas are subject to unusual draughts (Central America, Philippines, Indonesia, Africa and Australia) which lead to large scale fires which are difficult to extinguish because of the lack of rain. At the same time, other regions, Southern America, Southern Europe) experience increased flooding frequencies.

2.2.1Southern Oscillation Index (SOI)

Southern Oscillation Index is the difference in sea-level pressure (slp) between Darwin (Australia) and Tahiti's air pressure, multiplied by a factor of 10 [Troup, 1965]. Records of the monthly average SOI have been collected since January 1879, with missing values being computed by interpolation. Relationships between the SOI and rainfall have been extensively explored and numerous authors have shown its relationship rainfall in Eastern Africa

Despite the depth of research in this relationship, the SOI does not provide a strong predictor of precipitation occurrence [Hyndman 1999]. Furthermore it is proposed that the SOI values prior to 1935 should be used with caution, as there are questions regarding the consistency and quality of the Tahiti pressure values prior to this year. However, as SOI is used as a predictor of rainfall in current meteorological practices, it is considered as a covariate in this study. Its use, though, is approached with caution.

2.2.2 Troup's SOI Calculation

SOI Calculation formula was given by Troup as follows [Troup 1965]:

$$Troup's \quad SOI = \frac{PA\ (Tahiti - PA\ (Darwin)}{Std.Dev.Diff}\ x\ 10 \tag{2,1}$$

where: PA is the Pressure Anomaly equal to monthly mean minus long-term mean (1887-1989 base period). Std.Dev.Diff. is the Standard Deviation of the Difference (1887-1989 base period)

A SOI value of -10 means the SOI is 1 standard deviation on the negative side of the long-term mean for that month. Monthly SOI from 1876 onwards is derived from normalized Tahiti minus Darwin mean sea level pressure (mslp)

2.2.3 NCC SOI Calculation

The National Climate Centre (NCC) has a revised SOI calculation although still based on the Troup formula [Troup, 1965]. However, the base period for calculating the NCC SOI is 1933-1992..

2.2.4 SOI Phases

Research into the SOI has also found that an index which classifies seasons into 5 phases depending on the value and rate of change in the SOI would be useful when modeling rainfall [Stone and Auliciems, 1992] used a principal components analysis and cluster analysis to group all sequential two-month pairs of the SOI into five groups called the SOI phases. The SOI phases are recorded monthly, indicating which phase each month appears to be in. Generally, the use of SOI phases to calculate future seasonal rainfall probabilities gives a more accurate result than using SOI averages. The five phases can be stated generally in the following terms [Dunn and Lennox ,2006].

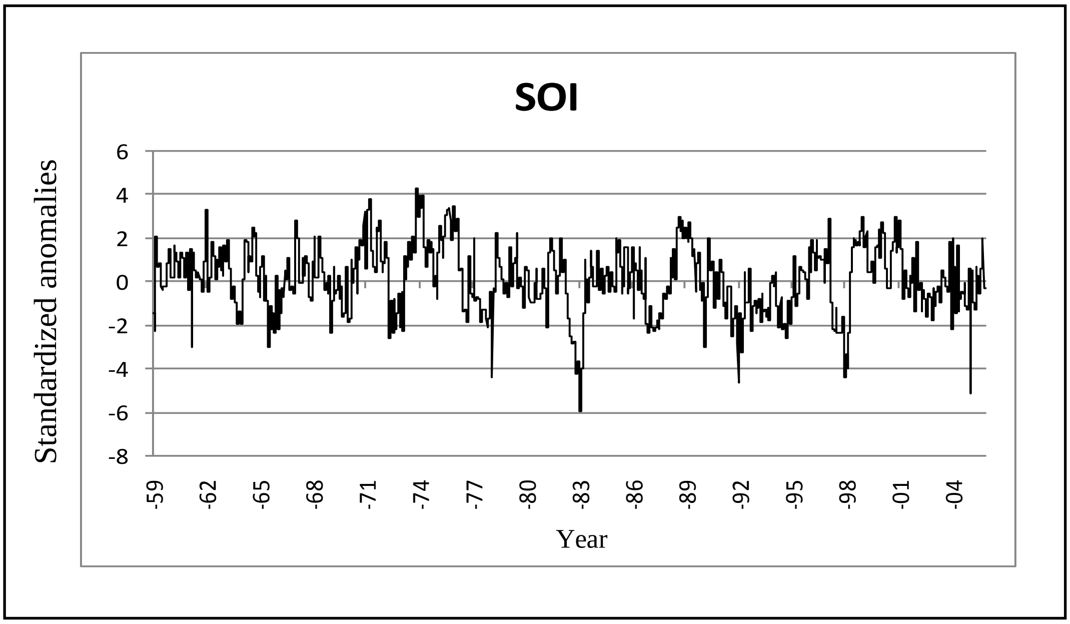


Figure 2.11
SOI index 1959-2005

i) Phase 1 – This is termed 'consistently negative'. It indicates that the SOI values for the two previous months are both negative.

ii) Phase 2 is termed 'consistently positive', indicates that the SOI values for the two previous months are both positive.

iii) Phase 3 is termed 'rapidly falling', indicates a marked decrease in the SOI from the previous month to the current month.

iv) Phase 4 is termed 'rapidly rising', indicates a marked increase in the SOI from the previous month to the current month.

v) Phase 5 is termed 'consistently near zero', indicates that both the SOI values for the previous two months are close to zero

Recent trends in the Southern Oscillation Index (SOI) can be used to calculate more accurately the probabilities of receiving particular amounts of rainfall at a particular location; over the next few months. The phases of the SOI were defined by Roger Stone then of QDPI, who used a statistical technique (cluster analysis) to group all sequential two-month pairs of the SOI (from 1882 to 1991) into five clusters (see legend below & help on use of trends in the SOI) [Stone and Auliciems 1992]

Droughts of 2004 and 1984 and 1960 occurred during the consistently near zero (Phase V) while the floods, 1962, 1977/78, 1981and 1997-98 occurred during rapidly changing (IV or V) phases.

SOI Phases were found to be more useful as factors than SOI values. In this study, modeling monthly rainfall amounts has been done using SOI-phases as a factor (Chapter 4).

2.3 Quasi-Biennial Oscillation (QBO)

The quasi-biennial oscillation (QBO) is a quasi-periodic oscillation of the equatorial zonal wind between easterlies and westerlies in the tropical stratosphere with a mean period of 28 to 29 months. The alternating wind regimes develop at the top of the lower stratosphere and propagate downwards at about 1km per month until they are dissipated at the tropical tropopause. Downward motion of the easterlies is usually more irregular than that of the westerlies. The amplitude of the easterly phase is about twice as strong as that of the westerly phase. At the top of the vertical QBO domain, easterlies dominate, while at the bottom, westerlies are more likely to be found.

The QBO was discovered in the 1950s, but its cause remained unclear for some time. Rawinsonde soundings showed that its phase was not related to the annual cycle, as is the case for all other stratospheric circulation patterns. In the 1970s it was recognized by James Holton and Richard Lindzen [Holton and Lindzen, 1972] that the periodic wind reversal was driven by atmospheric waves emanating from the tropical troposphere that travel upwards and are dissipated in the stratosphere by radiative cooling [Andrews, Horton and Leovy, 1987], [Baldwin, 2001]. The precise nature of the waves responsible for this effect was heavily debated. In recent years, however, gravity waves including the ones generated by solar and lunar gravitational force variations have come to be seen as a major contributor. There is a positive correlation of 0.02984 found in this study between daily rainfall and QBO values for the period under investigation meaning that other factors play more significant roles in determining rainfall amounts in the region.

Effects of the QBO include mixing of stratospheric ozone by the secondary circulation caused by the QBO, modification of monsoon precipitation, and an influence on stratospheric circulation in northern hemisphere winter (the sudden stratospheric warmings).

Equatorial waves in the lower stratosphere drive the quasi-biennial oscillation (QBO) and the semi-annual oscillation, which are the primary modes of variability of the equatorial stratosphere and also influence the variability of the polar vortex. They may also be important in stratosphere troposphere interaction. Two well-known equatorial waves, the Kelvin wave and westward-moving mixed Rossby gravity wave (WMRG) have been extensively investigated by theoretical, numerical modeling and observational studies. However, there is relatively less observational knowledge of vertical propagation characteristics of the waves and how the propagation is influenced by the basic ambient flows. ERA-40 data for two different years (1992 and 1993) have been used to investigate the behaviour of the equatorial waves under opposite phases of the QBO. Studies have provided an unprecedented and detailed view of vertical propagation of equatorial waves in different QBO phases. In the easterly-QBO phase there is more upward propagation of the Kelvin wave compared with the westerly QBO-phase, but less of the westward-moving mixed Rossby-gravity wave and Rossby wave. In general, equatorial waves in the lower stratosphere have higher frequency (and phase speed) than those in the upper troposphere.

2.4 SOLAR –LUNAR GEOMETRY

In this section we discuss how eight predicting factors; *sdec, ldec, atide, etide, synod, mld, perigee* and *apogee* may be obtained from solar-lunar geometry. The factors are chosen because they primarily influence the gravitational excitation potential of the moon and that of the sun on the atmosphere. We consider Figure 2.12 where O, C and S denote the centers of the earth, moon and sun respectively and P is the point of gravitational excitation in the atmosphere close to the earth surface. OE is along the Equator. Solar declination (*sdec*) and lunar declination (*ldec*) are the angles EOS and EOC respectively. The angle, as measured from the equator and is positive when the target (sun or moon) is in the Northern (+) and negative in the southern (-) hemisphere.

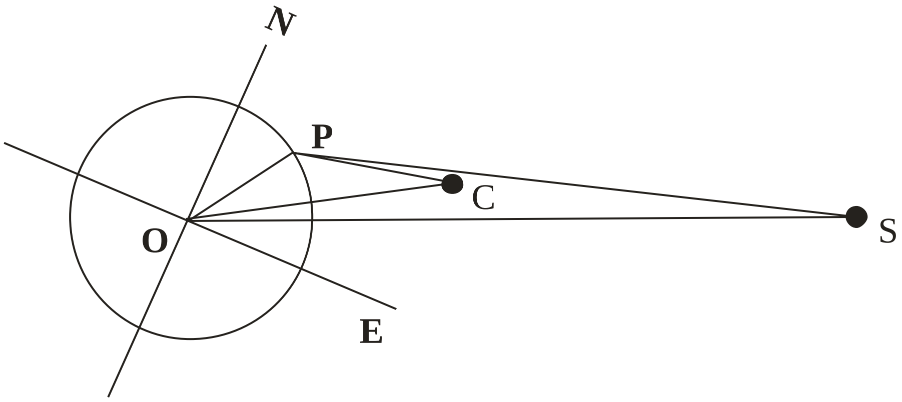


Figure 2.12
Geometry for calculation of tidal potentials.

When we consider the tidal potential due the moon, P is a point near the earth's surface. *N* denotes the North Pole. The potential of the attraction of *C* at point *P* is -*γM/PC,* where *M* denotes the mass of *C* and *γ* the gravitation constant [Lindzen and Chapman 1969], [Lamb, 1932]. If we put OC=D, OP=a, and denote the moon's (geocentric) zenith-distance at P, viz. the angle POC by Θ, this potential is equal to the local excitation, Ω_{tidal} and may be written;

$$\Omega_{tidal} \ = \ -\frac{\gamma M}{(D^2 - 2aDcos\Theta + a^2)^{1/2}}$$

(2,2)

We require, however, not the absolute accelerative effect on P, but the acceleration relative to the earth. Now the moon produces in the whole mass of the earth an acceleration equal to; γM/D^2 parallel to OC, and the potential of a uniform field of force of this intensity is evidently ; -γMcosΘ/D^2 It is the acceleration at *P* relative to the earth that produces tides. The potential associated with the acceleration of the earth as a whole is. Subtracting this from (2,3) above we get

$$\Omega_{tidal} = \ -\frac{\gamma M}{(D^2 - 2aDcos\Theta + a^2)^{1/2}} + \frac{\gamma M}{D^2} \ acos\Theta.$$

(2,3)

Expanding (2,3) in powers of *(a/D*), which is in our case a small quantity, and retaining only the most important term, first term, we get;

$$\Omega_{tidal} = \frac{3}{2}\frac{\gamma M a^2}{D^3}\left(\frac{1}{3} - cos^2\Theta\right).$$

(2,4)

Considered as a function of the position of P, this is a zonal harmonic of the second degree, with OC as axis so that the excitation is maximized when Θ = 0 and that P is on OC. An equivalent equation for the solar gravitation excitation is obtained by replacing the value of M by the mass of the sun. Thus the excitation is inversely proportional to the cube of the lunar or solar distance and is dependent on the lunar declination, Θ.

In this study an atmospheric tide state, *atide* occurs whenever O, C and S are co-linear or nearly collinear. O C and S were taken to be nearly collinear if the magnitude of the difference between *sdec* and *ldec* is less than 2 degrees. At that time the atmospheric tide is present somewhere in the tropics and not necessarily at P.

An enhanced tide (*etide*) is taken to occur when points OPCS are co-linear. During that time, the enhanced tide is now located at P and *sdec*=latitude at P (overhead moon and sun at P). *etide* occurs only during the new moon phase. We note that a solar eclipse event condition at P is satisfied whenever PCS are co-linear but that will not necessarily satisfy either the *atide* or the *etide* state at P. Thus the solar eclipse will always have tidal effects at some location where the declinations coincide with the latitude as seen in Figure 2.13.

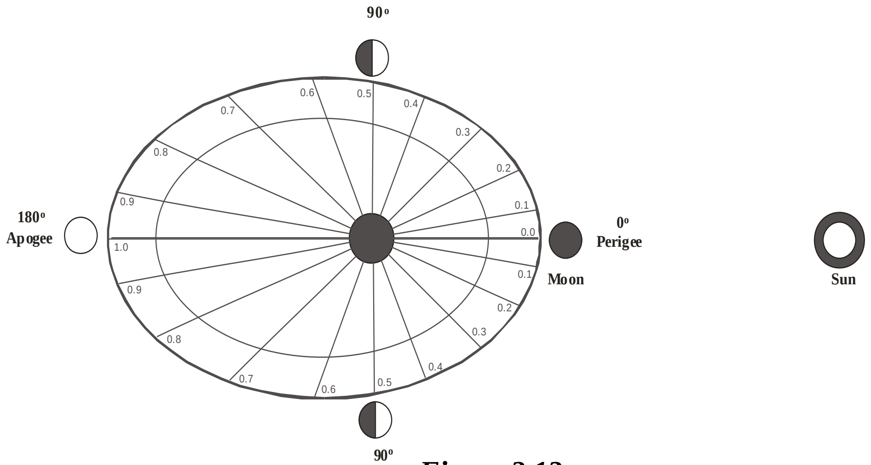


Figure 2.13

Lunar orbit showing New moon at perigee, the condition for the greatest tidal forces.

Due to the elliptic nature of the lunar orbit the relative strength of the tidal force within a lunation is determined by the earth-moon distance denoted by a synodic decimal value between 0.0 and 1.0. Figure 2.13 shows the earth-moon system with the earth at a lunar elliptical orbit focus. The magnitude of the tidal forces are symmetrical for the two halves of the lunation. The factor representing the tidal strength in any one month was taken to be the value of the synodic decimal at mid-month and referred to as the *synod*. *synod* has a value of 1.0 at apogee and 0 at perigee.

The moon describes an orbit round the earth in a plane inclined at 5.15 ° to the ecliptic; the pole of the orbit revolves about that of the ecliptic once in 18.60 years, so that the inclination of the plane of the moon's orbit to the earth's equator varies between 23.45°± 5.15° or 18.30° and 28.60°.The moon's declination consequently changes during each passage round its orbit between maximum northern and southern values which may vary from 18.5° to 28.5°. The change in maximum lunar declination (*mld*) influences lunar angular velocity relative to a terrestrial observer.

The value of the maximum lunar declination is the numerical value of the factor *mld* for the month. Values of *mld* used in this study for the period 1901-2050 are shown in Figure 2.14. MLDs have a 18.6 year cycle in agreement with Yndestad et al [Yndestad, William and Vladimir 2008].

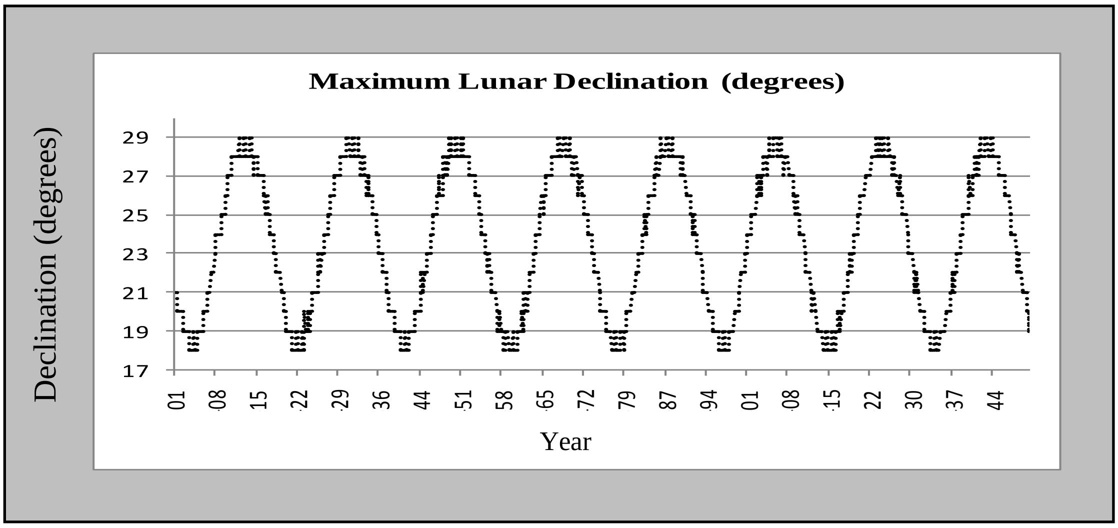


Figure 2.14(a)
Maximum lunar declination monthly values for the period 1901-2050

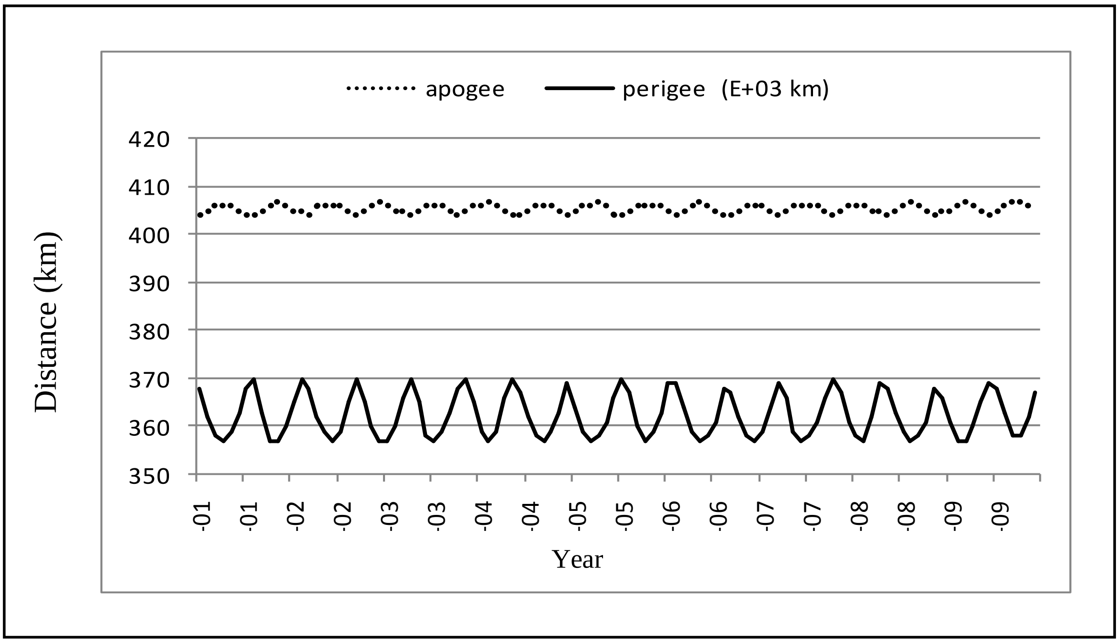


Figure 2.14 (b)
Apogee and Perigee distances for the period 1901-1910.

Perigee and apogee distances vary along the lunar orbit. Mean distance of the moon from the earth is 384405 km, or 60.335 times the earth's radius (6371.2 km) while the eccentricity of the orbit is considerable, and slightly variable; the mean ratio of the maximum distance, at apogee, to the minimum value, at perigee, is 1.1162, and the maximum ratio is 1.1411. The period from one apogee to the next is called the anomalistic month and the apogee revolves round the lunar orbit once in 8.8 years as shown in Figure 2.14(b). For each month, the average perigee and apogee distance is calculated. Numerical values represent the factor perigee (*prg*) and apogee (*apg*) as calculated by means of a tides calculator obtained from Dcsymbols [Dcsymbols, 2013]. Figure 2.14(b) shows apogee and perigee distances for the period 1901-1910. We observe from equation (2,10) that perigee variation can have more significant influence on tidal variation than apogee given that tidal potential is inversely proportional to the cube of the distance. During a perigee, the moon is 40,000 km closer than during an apogee and this distance varies by about 10,000 meters twice each year [Horizons, 2013]. The lunar phase (*lunaph*) is the integral value representing any of the four lunar phases, phase one (New Moon) being represented by integer 1.

2.5 ASTRONOMICAL BASIS FOR ENHANCED ATMOSPHERIC TIDES

Maximal tide raising forces occur only when the Sun and Moon are in direct mutual alignment. This occurs at *syzygy* (either full Moon or new Moon), provided also that the Moon or Sun be in eclipse with the Earth. The former two bodies must also be at the closest approach to the Earth, i.e., the Moon at *perigee* and the Sun at *perihelion*. Repetitions of syzygy, perigee, and eclipse are defined, respectively, by three lunar months *[*Wood 1986*]*. The *synodic* (29.5 days) representing every second recurrence of syzygy, the *anomalistic* (27.6 days) representing the recurrence of perigee, and the *nodical* (27.2 days) representing every second recurrence of the Moon positioned at its node, lying on the plane of the ecliptic, a requirement for an eclipse. The Earth and Sun attain closest approach (perihelion) once every anomalistic year. The anomalistic year is only slightly longer than the mean calendar year because perihelion advances very slowly, 1 day every 57 years [Neumann and Pierson,1966*]*. Perihelion presently occurs on January 2 in the Christian calendar.

For any point on the earth surface the occurrence of enhanced atmospheric tide may also be determined by the difference between the lunar and the solar declination at that point. A difference of zero means that the sun and the moon are in mutual alignment indicating the eclipse of the sun condition. In this study, enhanced atmospheric tidal events are taken to occur in Kenya so long as the said difference between solar and lunar declinations is less than one degree. Enhanced atmospheric tides are strongest at the equator where Kenya is located. The number of days during which the one degree requirement is met indicate the prevalence of the enhanced tide in any single month of the year.

Atmospheric tides have been detected as surface barometric pressure, averaging 1013.25 mb at sea level, fluctuates slowly by a maximum of about 50mb because of weather conditions. There is also a regular daily variation of up to a few mb showing diurnal and semidiurnal components, like the ocean tide. This atmospheric tide is completely solar, the lunar component being too small to observe. The effect is greatest at the equator and at continental locations, with a diurnal component of 0.3 to 0.5 mb, and a semidiurnal component of 1 - 2 mb, for a total range of 3 to 4 mb daily. This range is least at the solstices, and greatest at the equinoxes. In polar regions, the amplitude of the variation is only about 0.3 mb [Berry, 1945]

2.6 SUNSPOTS NUMBERS

The number of sunspots appearing on the solar surface has been recorded each month through observations and calculation for a long time. Currently, sunspot numbers are clearly headed towards a minimum given the trends and the near symmetry of the current maximum, typically referred to as Modern Maximum, which comprises Cycles 17 to 23 in Figure 2.15. The current cycle, Cycle 24, will probably mark the end of the Modern Maximum, with the sun switching to a state of less strong activity. While there are three main groups of prediction methods [Kristof, 2010] – precursor methods, extrapolation methods and model-based predictions – the National Aeronautics and Space Administration (NASA) and the Solar Influences Data

Analysis Center (SIDC) have finally used the precursor method and made their predictions for Cycle 24 [SIDC, 2013]. The smoothed sunspot numbers and predictions are shown in Figure 2.15.

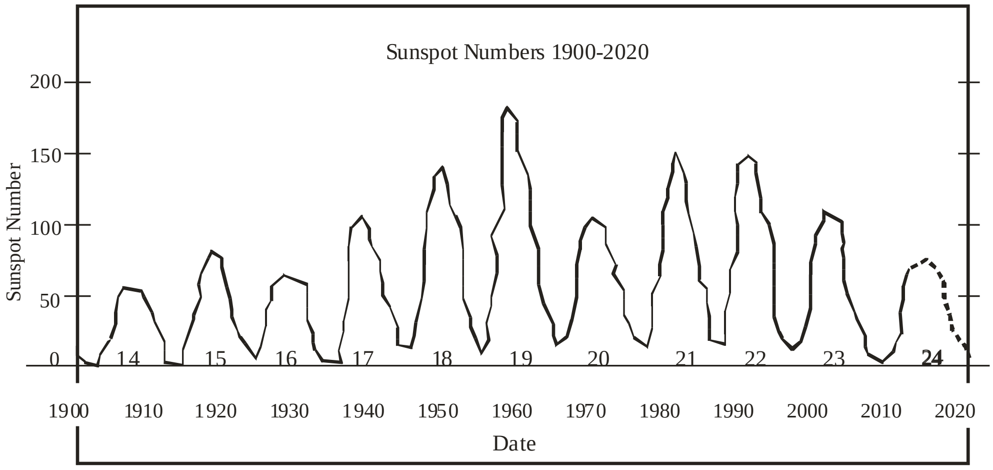


Figure 2.15
Smoothed sunspot numbers of Cycles 14 to 24 showing the predicted (dotted) segment.

Sunspot numbers have been associated with a change in climate, including severe climatic conditions during the Maunder Minimum – the period 1640–1705 which was characterised by a conspicuous lack of sunspots [Lassen and Christensen 1995]. Total solar irradiance increases when the number of sunspots increases. Total solar irradiance is higher at solar maximum, even though sunspots are darker (cooler) than the average photosphere. Meehl and Arblaster [Meehl and Arblaster, 2009] analysed sea surface temperatures from 1890 to 2006. They then used two computer models from the US National Center for Atmospheric Research to simulate the response of the oceans to changes in solar output. They found that as the sun's output reaches a peak, the small amount of extra sunshine over several years causes a slight increase in local atmospheric heating, especially across parts of the tropical and subtropical Pacific where sun-blocking clouds are normally scarce. The small amount of extra heat leads to more evaporation, producing extra water vapour. In turn, moisture is carried by trade winds to the normally rainy areas of the western tropical Pacific, fuelling heavier rains.

In 2008, White and Liu provided evidence that the 11-year solar cycle may be the trigger for El Niño and La Niña events by using harmonic analysis on observed and model data [White and Liu, 2008]. The dotted line in Cycle 24 represents NASA's predicted sunspot numbers for 2012–2020. A model such as the one developed in this study captures inter-annual rainfall variability by involving sunspot numbers as predictors.

Sunspot Cycle 24 is the last cycle of the current maximum while the dotted line shows the sunspot numbers that NASA have predicted for 2013–2020. The current prediction for Sunspot Cycle 24 gives a smoothed sunspot number maximum of about 69 in 2013 [Hathaway, Wilson and Reichmann, 1999]'s method of predicting the behaviour of a sunspot cycle is fairly reliable once a cycle has reached about 3 years after the minimum sunspot number occurs [Hathaway, Wilson and Reichmann, 1999]

2.7 Rainfall Model Formulation with Generalized Linear Models (GLMs)

2.7.1 Introduction.

Rainfall as a variable seems unpredictable for three reasons:

i) It has both continuous and discrete values,

ii) The variable has exact zero values visible in Figures 2.16 and 2.17. Deciding how to handle the numerous zero values occurring in a rainfall distribution presents enormous challenge.

iii) Kenya rainfall is a highly skewed variable -most analysis procedures are designed to handle normal distributions – therefore rainfall exhibits non-independent characteristics.

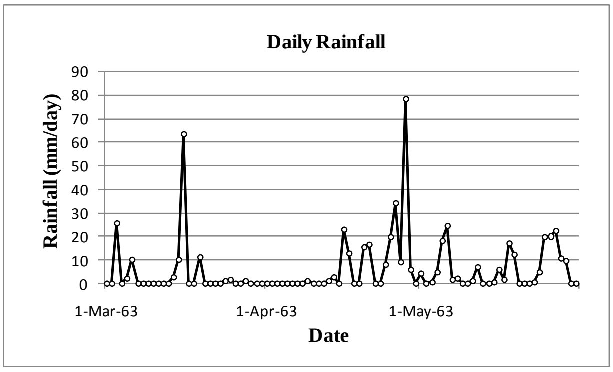


Figure 2.16
Curve Fit and rain Amount. The distribution has numerous zero values.

The approach we have used here, which has been used to handle environmental data is often composed of two separate components: a discrete element at zero and a continuous element recorded above zero. Figure 2.16 shows the time distribution of the ONDJ rainfall in 1963-4. In this approach, Rainfall is typically modeled using the two-components;

i) Modeling the occurrence,

ii) Modeling the amount.

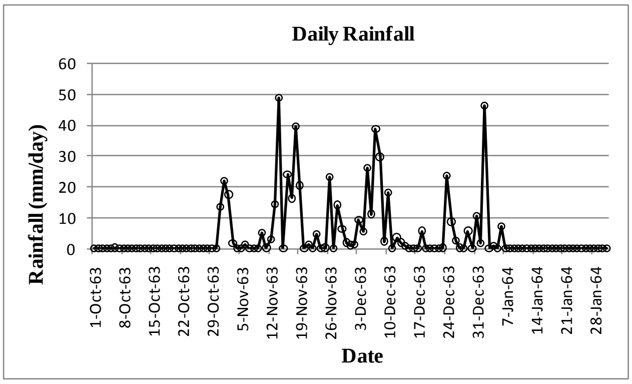


Figure 2.17
Distribution with less number of exact zero values

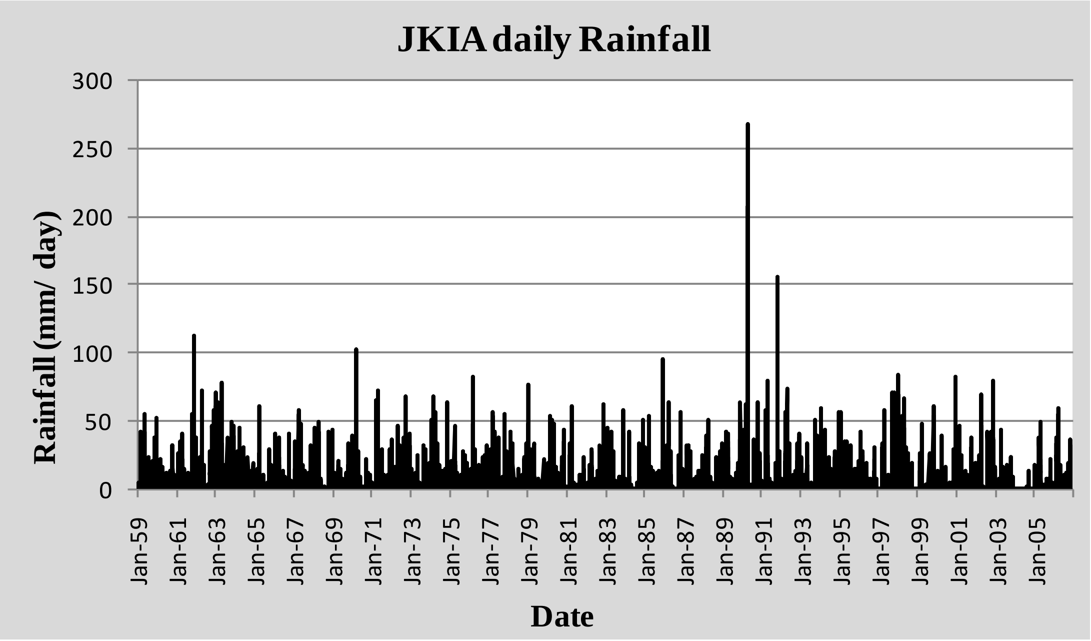


Figure2.18
JKIA daily Rainfall 1959-2005 showing an outlier value.

Models will often not be able to capture outliers as this one in 1990 because of the low probability of occurrence of such an event. From Figure 2.18, the probability of obtaining one value beyond 250mm is 1/564 = 0.0018 (there are 564 values in the dataset)

2.7.2 Modeling the occurrence of rainfall

Two methods have been commonly used to model occurrence:

i) Markov Chains.

Rainfall occurrence can be viewed as a sequence of random variables

$$\mathbf{y(t),t= t_1, t_2 \ldots\ldots..t_T}$$

(2,5)

where,

$$y(t) = \begin{cases} 1 - \text{Rainfall has occured} \\ 0 - \text{No rainfall has occured} \end{cases}$$

(2,6)

ii) A Renewal Process.

This process considers a sequence of alternating wet and dry spells of varying length, with each spell having an assumed distribution and that all intervals are independent

2.7.3 The Gamma Distribution

Since rainfall amounts are skewed as to the right in Figure 2.19 (rainfall amounts are exact zeros for most of the days) a function to use for a particular rain season is the gamma distribution function.

$$f(x) = \begin{cases} kx^{\alpha-1}e^{-x/\beta}, & x > 0, \\ 0 & \text{otherwise.} \end{cases}$$

2,7)

Where $\boldsymbol{\alpha} > 0$ the shape parameter and $\boldsymbol{\beta} > \mathbf{0}$ the scale parameter.

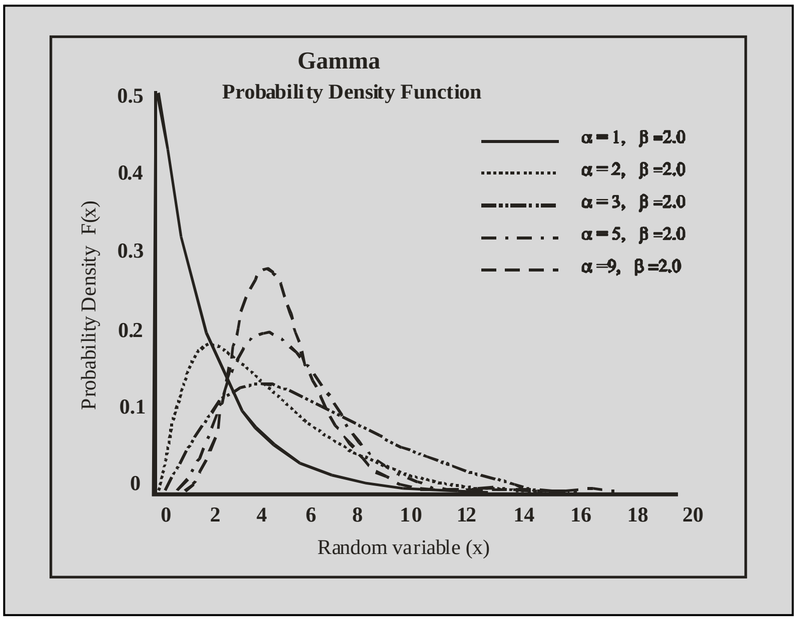


Figure 2.19

Gamma Distribution Functions.

The gamma distribution may therefore be appropriate only when modeling rainfall for a season.

2.7.4 Exponential Dispersion Models (EDMs)

The Normal, Binomial, Poisson, Inverse Gaussian, Exponential, Gamma, and Tweedie distributions all have distributions that form part of the exponential dispersion model family. The Binomial and Poisson distributions are both discrete distributions, with the Poisson distribution being used when the data involves counts. The Binomial distribution is used when the data deal with proportions and the outcome is either a ‘success’ or ‘failure’. The Normal, Inverse Normal, Exponential and Gamma distributions are all continuous distributions. The Gamma distribution is used when the response variable is skewed and the variance is not constant. The Exponential distribution is a special case of the Gamma distribution used when the shape parameter (**α**) is equal to one. Finally, the Tweedie distribution is a mixed distribution, which means that it can model data with both discrete and continuous components, such as the Poisson-Gamma distribution. The Tweedie distribution is especially useful in modeling rainfall, as illustrated in Section 3.5. Table 3.1 provides information about several distributions that come from the EDM family, including their variance functions (See Section 2.7.5). These seven distributions demonstrate that EDMs can consist of discrete, continuous, or mixed distributions.

We begin by assuming the Nairobi monthly rainfall follows one of the standard exponential dispersion family of distributions and we will therefore be an Exponential Dispersion Model (EDM). EDMs have a probability density function or a probability mass function, which can be written in the following form [Gill 2001];

$$p(y,\theta, ø,) = a(y,ø) \exp\left\{\frac{1}{\theta}[y\theta - k(\theta)]\right\} \quad . \qquad (2.8)$$

where $ø > 1$ is the dispersion parameter; μ is the position parameter and $\mu = k'(ø)$; y is the Nairobi monthly rainfall amount and θ is the canonical parameter. y does not depend on the parameters θ, and ø.

The notation $Y \sim ED(\mu, ø)$ indicates that a random variable Y comes from the EDM family, with location parameter μ and dispersion parameter ø, as in equation (2,8).

2.7.5 The GLM

A GLM, such as the Tweedie used in this study, satisfies two conditions:

1. *It is an EDM ; ie* $y_i \sim ED(\mu_i, ø/w_i)$*. The value of prior* weights w_i is 1
2. The expected values of the y_i, say μ_i, are related to the covariates *xi* through a monotonic differentiable link function, $g(\cdot)$

Table 2.1 provides information about several distributions that come from the EDM family, including their variance functions. These seven distributions demonstrate that EDMs can consist of discrete, continuous, or mixed distributions.

Table 2.1

The characteristics of some of the distributions. [McCullagh & Nelder]

Distribution	$k(\theta)$	$\mu = E(Y)$	Variance Function
Normal	$\theta^2/2$	θ	1
Poisson	e^{θ}	e^{θ}	μ
Binomial	$\ln(1+e^{\theta})$	$e^{\theta}(1+e^{\theta})$	$\mu(1-\mu)$
Gamma	$-\ln(-\theta)$	$-1/\theta$	μ^2
Inverse Gaussian	$-(-2\theta)^{1/2}$	-2θ	μ^2
Tweedie	$\frac{\theta(1-p)^{(2-p)/(1-p)}}{(2-p)}$ for $p \neq (1,2)$	$-k'(\theta)$	μ^2 for $p \neq (0,1)$

Examples of the predictors *(x_i)* are : SOI phase, Solar and lunar declination, , Sunspot numbers , Month (1,2,3,…,12), Solar Declination and Lunar Declination . They may also be referred to as explanatory variables or covariates.

2.7.6 The link Function

The link function, *g(.)* . This function is the one to be determined by fitting so that

$$g(\mu_i)= x_i^T \beta \qquad (2,9)$$

and

$$y_i = \beta\, x_i^T + e_i \qquad (2,10)$$

or

$$y_i = g(\mu_i)\, x_i^T + e_i \qquad (2,11)$$

is a linear function, hence the name –generalised linear model. e_i are the random residuals (errors in estimating y_i)

Table 2.2
Commonly used link functions. μ and p are the mean and power. ϕ is the Normal cumulative.

Distribution	Canonical Link	Form	Other Links	Form
Binomial	logit	log[p/(1 − p)]	probit	ϕ^{-1}(p)
			c-log-log	log[-log(1 − p)]
Poisson	log	log μ	identity	μ
			square root	log (μ)
Gamma	inverse	1/μ	log	log (μ)
			identity	μ

In this study we use the Tweedie distribution which follows a gamma distribution and log link function, *log (μ)* due to the skewed nature of the rainfall distribution. The linear component, $x_i\beta$, is called the linear predictor and is given the symbol η_i, so that,

$$\eta_i = x_i^T \beta. \qquad (2,12)$$

The linear function *g(.)* is differentiable, so that β can be estimated and monotonic and that

x^T_i has only one value corresponding to each μ_i .

$$g(\mu_i)= x_i^T \beta \qquad (2,13)$$

The most commonly used form is g(.) is $\eta_i = \theta = g(\mu)$. So that the linear predictor,

$$\eta_i = x_i^T \beta. \qquad (2,14)$$

When we define x_{it} as the covariate vector for unit i at time t, the link function becomes;

$$g(t_{it})= \eta_{it} \qquad (2,15)$$

and also

$$y_{it} = g(\mu_{it})x_{it}+e_{it} \qquad (2,16)$$

the variance as a function of the mean, and consequently the distribution of the response variable becomes;

$$Var[y_{it}] = \phi V\,(\mu_{it}) \qquad (2,17)$$

2.7.7 The Mean and The Variance

Members of the EDM, family written in the form of Equation (2,8), have a mean and variance defined as follows, where $\kappa(\theta)$ and ϕ are determined from equation (2.1) [McCullagh and Nelder].

$$\text{Mean of } Y: \quad E[Y] = \mu = \kappa'(\theta) \qquad (2,18)$$

$$\text{Variance of } Y\,(var[Y]): \quad var[Y] = \kappa''(\theta) \qquad (2,19)$$

The variable θ is related to the mean μ through Equation (2,18). The relationship between μ and θ is often written as $\tau(\theta) = \kappa'(\theta) = \mu$ and $\theta = \tau^{-1}(\mu)$. The function $\tau\,(\theta)$ is referred to as the mean-value mapping and gives the functional relationship between μ and θ.

2.7.8 Variance Functions

Although not described in the original setup of a GLM, the variance function is important as it uniquely identifies a distribution within the class of EDMs. Equation 2.10 shows that $\kappa'(\theta)$ is a function of the mean and thus $\kappa''(\theta)$ is also dependent on the mean. For this reason $\kappa''(\theta)$ is often replaced by the variance function $V(\mu)$ so that,

$$V(\mu) = \kappa''(\theta). \qquad (2,20)$$

The role of the variance function is to describe the mean-variance relationship of a distribution when the dispersion parameter is held constant. If Y follows an EDM with mean μ, variance function $V(\mu)$, and dispersion parameter $\emptyset$, then the variance of Y can be written as,

$$var(Y) = \emptyset V(\mu). \qquad (2,21)$$

The variance function that uniquely identifies the Normal, binomial, Poisson, inverse Gaussian, gamma, and Tweedie distributions is illustrated in Table 2.2. The Tweedie distributions, are classified by a special form of the variance function ($V(\mu) = \mu^p$).

2.7.9 Deviance

One method to measure the appropriateness of a fitted model is to examine the difference between the fitted values $\hat{\mu}$ and the observed values y. In standard Normal distribution based regression, this measure is equivalent to the residual sum-of-squares (Hardin and Hilbe, 2001]. In the framework of GLM, this measure of difference is called the deviance, $D(y; \mu)$, and can be

calculated as follows,

$$D(y; \mu) = \emptyset D^*(y; \mu) = 2\emptyset[\ell(y; y) - \ell(\hat{\mu}; y)], \qquad (2,22)$$

where D^* is called the scaled deviance and has only an approximate χ^2 distribution, and ℓ is the log-likelihood function. The deviance can be used to compare models [Hardin and Hilbe, 2001] and [Nelder and Wedderburn, 1972].

2.7.10 Estimation of parameters

Using an iterative procedure [Dobson, 2002] to obtain the maximum likelihood estimators of the parameters β, it is possible to fit a model to a data set. GLMs are estimated numerically with these parameters. The likelihood function is generally defined as follows;

$$L(\xi; y) = \prod_{i=1}^{n} f(y;\ \xi), \tag{2,23}$$

Where n is the sample size of the data set, and ξ is the parameter of interest. It is easier to work with log-likelihood function, which is defined as

$$\ell(\xi; y) = \log L(\ \xi; y) \tag{2,24}$$

$$L(\xi; y) = \log \prod_{i=1}^{n} f(y;\ \xi), \tag{2,25}$$

$$= \sum_{i=1}^{n} \log f(y;\ \xi). \tag{2,26}$$

In order to use this method, the log-likelihood, $l(\theta, \emptyset; y)$ needs to be determined as follows;

$$l(\theta, \phi; y) = \sum_{i=1}^{n} a(\phi, y) + \frac{1}{\phi}\ [y\theta - \kappa(\theta)] \tag{2,27}$$

The maximum likelihood estimates for β_j can now be found by taking the derivative of the equation (2,27) above with respect to β_j

Now,

$$\frac{\partial \ell}{\partial \beta j} = \frac{\partial \ell}{\partial \theta_i} \times \frac{d\theta_i}{d\mu_i} \times \frac{d\mu_i}{d\eta_i} \times \frac{\partial \eta_i}{\partial \beta j} \tag{2,28}$$

Each of the four derivatives on the RHS can be obtained separately by differentiating the log-likelihood function as follows:

$$\frac{\partial \ell}{\partial \theta_i} = \sum_{i=1}^{n} \frac{1}{\phi}\ [y_i - \kappa'(\theta_i)] \tag{2,29}$$

$$= \sum_{1=1}^{n} \frac{1}{\phi} [y_i - \mu_i], \tag{2,30}$$

since $$\mu_i = \mathrm{E[Y]} = \kappa'(\theta_i) \tag{2,31}$$

The second component uses the relationship $\mu_i = E[Y] = \kappa'(\theta_i)$ as well;

$$\mu_i = \kappa'(\theta_i) \tag{2,32}$$

$$\frac{d\mu_i}{d\theta_i} = \frac{d\kappa\prime(\theta_i)}{d\theta_i} \tag{2,33}$$

$$= \kappa''(\theta_i) \tag{2,34}$$

$$=V(\mu_i) \tag{2,35}$$

Inverting this final equation gives

$$\frac{d\theta_i}{d\mu_i} = \frac{1}{V(\mu i)} \tag{2,36}$$

The third component differentiates the link function g(μ_i) = η_i,

$$\eta_i = \mathrm{g}(\mu_i) \tag{2,37}$$

$$\frac{d\eta_i}{d\mu_i} = \frac{\mathrm{g}(\mu_i)}{d\mu_i} \tag{2,38}$$

$$= \mathrm{g}'(\mu_i). \tag{2,39}$$

Inverting equation (2,39) gives

$$\frac{d\mu_i}{d\eta_i} = \frac{1}{\mathrm{g}\prime(\mu_i)} \tag{2,40}$$

The final expression uses $\eta_i = \beta_0 x_{i0} + \beta_1 x_{i1} + \ldots\ldots\ldots\ldots + \beta_j x_{ij} + \ldots\ldots\ldots\ldots\ldots \beta_p x_{ir}$,

Where r is the rank of β. Thus, the derivative of η_i with respect to β_j is x_{ij} .

Combining these four expressions shows that the equation 2.21 can be written as the "score equation" for GLMs,

$$\frac{\partial \ell}{\partial \beta_i} = \frac{1}{\phi}\sum_{i=1}^{n}\frac{(y_i-\mu_i)}{V(\mu_i)}\frac{x_{ij}}{g'(\mu_i)} \tag{2,41}$$

The maximum likelihood estimator is found by setting Equation (2,41) equal to 0 and solving for $j = 1, 2, . . . , r$. When $\partial\ell/\partial\beta_i = 0$, the value of ϕ does not need to be known. This is an important concept of GLMs because an estimate of β can be found without knowing ϕ.

Equation (2,41) can only be solved through numerical techniques involving iteration, such as the Newton-Raphson method or the method of scoring [Dobson, 2002] and [Hardin and Hilbe, 2001]

2.7.11 Quasi-likelihood methods

In many situations, some details of the distribution governing the data is known, however the distribution may not be specified exactly. In addition, there are some cases for which the distribution is known, however it difficult to evaluate, such as the Tweedie distributions. This precludes the use of maximum likelihood, which requires exact specification of the distribution in order to construct the likelihood. The idea of quasi-likelihood addresses this concern [McCulloch and Searle, 2001].

Quasi-likelihood methods were first proposed by [Wedderburn, 1974], and are a methodology for regression that requires few assumptions about the distribution of the dependent variable. Hence they can be used with a variety of outcomes [Zeger and Liang, 1986]. In likelihood analysis, the actual form of the distribution must be specified. However, in quasi-likelihood, only the relationship between the outcome mean and covariates, and the mean and variance, needs to be specified [Zeger and Liang, 1986]. The focus of quasi-likelihood is on methods for inference about β, and hence ϕ can be treated as a nuisance parameter. A quasi-likelihood can be used if the researcher does not know the density function of the distribution, but knows its mean and variance. It is defined for one observation, Q, as,

$$Q(y;\, \mu) = \int \frac{(y - \mu)}{V\,(\mu)}\, du\,. \tag{2,42}$$

This quasi-likelihood has the same properties as a true log-likelihood with regards to the derivatives of β, enabling GLMs to be fitted for any distribution using a quasi-distribution. To define a quasi-likelihood function, only the relationship between the mean and variance needs to be specified through the variance function [Wedderburn, 1974].

2.7.12 Power-variance (Tweedie) distributions

Of special interest within EDMs is a class of distributions with power mean-variance relationships $V(\mu) = \mu^p$. Any distribution whose variance function like this belongs to the class of distributions known as the Tweedie family of distributions, named by Jørgensen [Jørgensen, 1987] after Tweedie [Dunn, 2004]. This section describes Tweedie distributions, and demonstrates how these distributions can be used to model rainfall.

Most of the important distributions commonly associated with GLMs are contained within the Tweedie distribution framework, including the Normal ($p=0$), Poisson ($p=1$ *and* $\phi=1$), gamma ($p=2$), and inverse Gaussian distributions ($p=3$). Tweedie models exist for all values of p outside the interval (0, 1), however only the four distributions already mentioned have density functions which have explicit analytic forms [Dunn & Smyth, 1996].

Tweedie distributions with *p>1* have strictly positive means, with *p>2* being continuous for positive Y, and a shape similar to the gamma, but more right skewed. Distributions with *p<0* are continuous on the entire real axis. Finally, for *1<p<2* the distributions are supported on non-negative real numbers, and the distributions are mixtures of the Poisson and gamma distributions, with a mass at zero [Dunn & Smyth]. These distributions have been called ‘compound Poisson’, ‘compound gamma’, and ‘Poisson-gamma’ distributions. Due to the characteristic of being able to model both discrete and continuous combinations simultaneously, these distributions have a special use in being able to model both the occurrence and amount of rainfall.

The mean, μ, and canonical parameter, θ can be found for a Tweedie distribution by noting that $\kappa''(\theta) = d\mu/d\theta = \mu^p$ and the mean is given by $\mu = \kappa'(\theta)$. This allows the density function for a Tweedie distribution to be specified. Hence,

$$\mu^p = \frac{\partial^2 \kappa}{\partial \, \theta^2}. \qquad (2,43)$$

$$= \frac{\partial}{\partial \theta} \left(\frac{\partial \kappa}{\partial \theta}\right). \qquad (2,44)$$

$$= \frac{\partial \mu}{\partial \theta}. \qquad (2,45)$$

Taking the reciprocals of both sides and integrating with respect to μ gives,

$$\theta = \begin{cases} \frac{\mu^{1-p}}{1-p} & p \neq 1, \\ \log \mu & \text{p=1.} \end{cases}. \qquad (2,46)$$

By setting the arbitrary constant of integration to 0, and noting that $\mu = \kappa'(\theta)$ gives,

$$\kappa(\theta) \;=\; \begin{cases} \frac{\mu^{2-p}}{2-p} & p \neq 2, \\ \log \mu & \text{p=2.} \end{cases} \quad (2,47)$$

The Tweedie densities can thus be written as,

$$f_p(y; \mu, \phi) = a_p(y, \phi) \exp\left\{\frac{1}{\phi}\left[y\ \frac{\mu^{1-p}}{1-p} - \frac{\mu^{2-p}}{2-p}\right]\right\} \mathit{for}\ p \neq (1{,}2)\,. \quad (2,48)$$

2.7.13 Tweedie distribution and the Quasi-likelihood

Following Equation (2.36) the Tweedie distribution has the following quasi-likelihood distribution (when setting the arbitrary constant of integration to 0

$$\mathrm{Q}(\mu;\ \mathrm{y}) = \int \frac{y-\mu}{V(\mu)}\, d\mu \quad (2,49)$$

$$= \int \frac{y-\mu}{\mu^p}\, d\mu. \quad (2,50)$$

$$= \int\ \frac{y}{\mu^p} - \mu^{1-p} d\mu. \quad (2,51)$$

$$= \int (y\mu^{-p} - \mu^{1-p}) d\mu. \quad (2,52)$$

$$= \frac{y\mu^{1-p}}{1-p} - \frac{\mu^{2-p}}{2-p}.. \quad (2,53)$$

This equation has the same likelihood function as equation (2,42), except now there is no need to estimate *a(y, ϕ)*. This is extremely helpful as often the *a(y, ϕ)* term cannot be written in closed form, or is of a form which is extremely difficult to calculate.

2.7.14 Software for Fitting a Tweedie Function

The model design involves fitting a GLM of the Tweedie family to the Nairobi Rainfall. The R Statistical Software was used to calculate the model parameters while a MS Excel spreadsheet was used in the calculation of the estimated values. R is free software, and this copy was obtained from NCAR, Boulder, Colorado during the Colloquium on Africa Climate held in August of 2011.

To fit a Tweedie GLM the tweedie library is needed program [Dunn, 2004]. The two functions loaded were the *tweedie.profile*; and the *tweedie* family model. The *tweedie.profile* function is the most suitable and works only if $p \geq 1$ which gives maximum likelihood value of p and *ø* at 95% confidence.

Once the variance power has been calculated using *tweed.profile*, and the link function chosen (default link is canonical –the *log* option was used), the R command was used:

family=tweedie(var.power=p,link.power=1-var.power)

The program R is used extensively in this study and is the program that was used to create the rainfall models in generated in Chapters 4.

2.7.15 Tweedie Distributions and Rainfall

To model rainfall using the Tweedie model, one vital assumption needs to be made: the amount of rainfall that occurs during any rain event follows a gamma distribution. Let i be a rainfall event, and R_i be the amount of rainfall that occurs during this event. It is assumed that each R_i follows a gamma distribution, with mean $-\alpha\gamma$ and variance $-\alpha\gamma^2$ (Gam($-\alpha,\gamma$)). It is also assumed that the number of rainfall events during the time period (usually month or day), called N, follows a Poisson distribution with mean, λ. Thus when no rainfall has occurred on that particular event, $N = 0$. Finally Y represents the total daily or monthly rainfall, and is represented as the Poisson sum of gamma random variables, such that $Y = R1 + R2 + . . . + RN$. This same setup can be applied to differing timescales. For example, if Ri represents the amount of rainfall per day, then Y is the total monthly rainfall. The resulting distribution of Y is called a Poisson-gamma distribution [Dunn, 2004], and belongs to the class of Tweedie distributions when $1 < p < 2$.

A Poisson-gamma distribution has probability function, however Jørgensen shows that it takes the following form [Jørgensen B, 1987]:

$$\log\ f_p\ (y;\mu,\phi) = \begin{cases} -\lambda, & for\ \ y = 0 \\ -\frac{y}{\gamma} - \lambda - \log y + \log W\ (y,\phi,p), & for\ y > 0, \end{cases} \cdot \tag{2,54}$$

Where $\gamma = \phi(p\text{-}1)\ \mu^{p\text{-}1}$, $\lambda = \mu\ 2\text{-}p/[\ \phi(2\text{-}p)]$, and W is an example of Wright's generalized Bessel function. It can be written as,

$$W(y,\phi,p) = \sum_{j=1}^{\infty} \frac{y^{-j\alpha}(p-1)^{\alpha j}}{\emptyset^{j(1-\alpha)}(2-p)^{j}!\ \Gamma(-j\alpha)}\ \ ,\ . \tag{2,55}$$

Where $\alpha = (2\text{-}p/1\text{-}p)$. The mean of the Poisson-gamma distribution is μ, and its variance, as with all Tweedie distributions, is

$$var(Y) = \phi\mu^p \tag{2,56}$$

The probability of obtaining no rain on any particular event is given by [Dunn, 2004];

$$Pr(Y = 0) = exp(-\lambda) = exp\ \left[-\frac{\mu^{2-p}}{\phi(2-p)}\right] \qquad (2,57)$$

2.7.16 Diagnostic Testing

The purpose of creating a model is to adequately summarize the important characteristics of the data by finding a parsimonious model that explains what is happening in the data without using meaningless, or too many parameters. In the creation of a model, often this model may show departures from the given data and thus not fit the data sufficiently. Diagnostic testing is used to determine whether the model adequately fits the data. There are a number of diagnostic tests that are available for GLM, and these include: a Q-Q plot; scatter plots of residuals and covariates; comparison of residual sizes; and residual deviances. These techniques allow the suitability of the link function and assumed distribution to be tested, as well as testing of the data for influential values, outliers, or pattern. There are four main reasons why a fitted GLM may not adequately represent the data and these include,

- The model fits well for most observations; however a few isolated cases do not. These isolated cases are called outliers;
- The link function is incorrectly specified;
- The response variable, Y is incorrectly specified; and/or
- The linear predictor *(η)* may not be correctly specified, or is missing some terms.

2.7.17 Residuals

A general tool used in diagnostic analysis is residuals. Residuals are a measure of how different expected values of the responses emerge from the observed responses. In simple regression models, the raw residuals $(y - \hat{y})$ are used, however these are generally inadequate when using a generalized linear model. The two most common residuals to use for GLMs are the Pearson residuals and deviance residuals. The Pearson residuals, which are also used in GEE models, have an approximate Normal distribution $N(0, \phi)$. Deviance residuals are related to the concepts of deviance $D(y; \mu)$, and also have an approximate Normal distribution. Quantile residuals have also been recently proposed by [Dunn and Smyth, 1996] to be used with GLMs, and have an exact Normal distribution when μ and $\phi$are known exactly.

2.7.18 Definition of Quantile Residuals

In continuous responses, the quantile residual is defined as,

$$r_{Q,i} = \phi^{-1}F(y_i; \mu_i, \phi), \tag{2,58}$$

where $F(y_i; \mu_i, \phi)$ is continuous and is the distribution function of a random variable Y , and $\phi(\cdot)$ is the cumulative distribution function of the standard Normal distribution.

In the discrete case, if $a_i = \lim_{y \uparrow yi} F(y_i; \mu_i, \phi)$ and $bi = F(y_i; \mu_i, \phi)$, then the quantile residuals are defined as,

$$r_{Q,i} = \phi^{-1}(\mu_i), \tag{2,59}$$

where μ_i is a uniform random variable on the interval $(a_i, b_i]$.

2.7.19 Residual Plots

Any of the residuals discussed in section 2.6.17 can be plotted against a variety of statistics and other indices. Each provide different information about departures from the fitted model. Since residuals should ideally be random, any pattern observed in the plots indicate problems with the fitted model. These residual plots can therefore help the researcher determine if there are any isolated departures. Furthermore, by plotting the residuals against the fitted values, as well as against the covariates, systematic departures can also be determined [Chandler, 2003].

2.7.20 Correct Distribution

One of the most important components of a GLM is that the correct distribution is chosen for the response variable. To check that the chosen distribution is adequate for the data, a normal probability or Q-Q (quantile) plot can be produced. If the model fits well, this plot should yield a straight line at 45 degrees. While quantile residuals are the ideal choice for GLMs, other residuals can be used.

2.8 GENERALIZED ESTIMATING EQUATIONS (GEES).

The class of generalized linear models (GLMs) introduced in Section 2.7.1 play a central role in regression problems which have discrete or continuous response variables. However they are based on the classical assumption that observations within a data set are independent. GLMs were extended by Liang and Zeger [Liang and Zeger, 1986] so that longitudinal or correlated data could be analyzed, and this approach is known as the Generalized Estimating Equation (GEE) method. This method has received wide use in medical and biological applications such as epidemiology, gerontology, and biology [Ballinger, 2004], and is becoming increasingly popular in other disciplines such as organizational and psychological research. Much of the appeal of GEEs is due to their broad capabilities, including: modeling correlated responses; allowing for time-varying covariates; and facilitating regression analysis on dependent variables that are not normally distributed [Ballinger, 2004].

2.8.1 Introduction

GEEs were introduced as a method of estimating the regression model parameters when the response variable is dependent. The GEE approach differs in a fundamental conceptual way from the techniques included under the rubric of ‘random-effects’, ‘multilevel’, and ‘hierarchical’ models which have previously been used to model correlated data. The techniques used in these models explicitly model and estimate the variations seen between observations, and incorporate these estimates and the residual variance into standard errors. The GEE method does not explicitly model the variation. Instead it focuses on, and estimates its counterpart: the similarity of the observations [Hanley, Edwardes, Negassa and Forrester. 2003]. GEEs develop a population average or marginal model. In marginal models, the primary interest of the analysis is to model the marginal expectation of the response variable given the covariates. In other words, for every one unit increase in a covariate across the population, the GEE tells the user how much the average response would change [Zorn, 2001].

The correlation, or more generally, the association between the response variables is modeled separately and is regarded as a nuisance parameter [Ziegler, Gromping, Kastner and Blettner, 1996]. Thus, a basic premise of the GEE approach is that the researcher is primarily interested in the regression parameters β and is not interested in the variance-covariance matrix. GEEs are not meant to be used in situations in which scientific interest centres on the variance parameters. This section focuses on the class of GEE models originally developed by [Liang and Zeger, 1986]. This GEE approach is now commonly referred to as the GEE1 approach. Further developments are currently being made into different types of GEEs. While the focus of the chapter is on GEE1 models, the other types of GEEs are discussed briefly.

2.8.2 Longitudinal and correlated studies

GEEs are traditionally used to model correlated data from longitudinal or repeated measures units, as well as from clustered or multilevel studies. Longitudinal studies are defined by the characteristic that subjects are measured repeatedly throughout time. These studies require special statistical methods because the set of observations taken on one unit are usually inter-correlated, [Diggle, Liang and Zegler, 1994]. The issue of accounting for correlation also arises when analyzing a single time series of measurements, such as rainfall. Although similar techniques can be applied to this type of data, inferences are usually less robust. The correlation must therefore be taken into account in order for valid scientific inferences to be made , [Diggle, Liang and Zeger, 1994]. The data examined in this study is a single time series measurement, rainfall, which is measured over time. The site of the rainfall is considered as one unit, and the rainfall measured at each site would be the repeated measures over time. The prime advantage of studying rainfall in this manner is that multiple sites can be examined simultaneously and it is an effective way to study change. However, if more than one site is examined simultaneously in one model, it can be thought of as an example of a longitudinal study. Correlated data has been examined through a variety of different approaches. The statistical methods for modeling longitudinal data are well developed when the response variable is approximately Normal [Liang and Zeger, 1986]. Statistical models for non-Normal outcomes however, are not as developed. Where analyzing longitudinal data there are two classical approaches which have been used in the past: the first is univariate mixed model, split-plot, or repeated measures anova; and the second is based on a multivariate anova called manova. Two other extensions to the classical approaches for modeling correlated data include multivariate modeling and mixed models. The former treats all measurements on the same unit as dependent variables, and models these simultaneously. The latter focuses on fixed and random effects within the model, with the correlation between the observations being a consequence of random effects [Dunlop, 1994].

2.8.3 Notation

The following notation is used: let y_{it} be a vector of responses with a set of corresponding r covariates or factors, X_{it}, where i indexes the K units of analysis $i = 1, 2, . . . ,K$; and t indexes the time points. $t = 1, 2, . . . , n_i$ for each unit.

Thus the number of clusters observed is K. Also,

$$N= \sum n_i,$$

(2,60)

and is the total number of observations across all units. The first element of x_{it} is set to 1 to allow the inclusion of an intercept.

Furthermore, let $y_i = [y_{i1}, y_{i2}, . . . , y_{ini}]$ denote the corresponding column vector of observations on the response variable for unit *i*, and $X_i = [X_{i1},X_{i2}, . . . ,X_{ini}]$ indicate the n_i × *r* matrix of covariates for unit *i*.

In the case of rainfall data, to correspond with the notation described above, the following notation is applied:

Each site forms one unit or cluster. Therefore if only one site is examined, $K=1$. If two sites are examined, then $K=2$. Thus, $i=1$ for one site and $i=1, 2$ for two sites.

The response variable, y_{it}, is the amount of rainfall recorded. Thus, if one site is examined, the response variable becomes y_{1t}. If two sites are examined, there are two response vectors of y_{1t} and y_{2t}.

The observed time, *t*, corresponds with the time values at which the rainfall is measured. For example $t = 1, 2, 3$ would correspond with measurements taken at time point 1, time point 2, and time point 3. The number of time points is n_1 for site one and n_2 for site two.

2.8.4 Assumptions

Before explaining the concept of GEEs, there are four assumptions about the use of GEEs to model correlated data that need to be articulated. The most crucial assumption is that the following conditional expectation needs to be specified correctly,

$$\mu_{it} = E[y_{it}|x_{it}] = E[y_{it}|\boldsymbol{X}_i] \qquad (2,61)$$

Equation (2,61) implies the conditional mean μ_{it} of y_{it}, given the explanatory variable X_i, measured at all possible time points n_i, is equal to a set of the same point specific explanatory variables x_{it} [Dahmen and Ziegler, 2003].

The second assumption is that the response variable y_{it} should have a mean and variance which are characterised by a GLM. It is further assumed that a true conditional $n_i \times n_i$ covariance matrix exists [Dahmen and Ziegler, 2003]. Finally, it is imperative that any missing data is missing completely at random (mcar), otherwise results become inconsistent [Dobson, Puride, and Williams, 2002].

2.8.5 GEEs and rainfall

There is a general consensus that rainfall is correlated. For monthly rainfall this means that the rainfall observed during any particular month, depends on a number of previous months' conditions. Studies have shown that this is the case, and thus the correlated structure of rainfall data should not be ignored when creating a model [Chandler and Wheater, 1998]; [Beersma and Buishand, 2003]. Even though researchers have realized that rainfall data is correlated, introducing these dependencies into a model leads to difficulties. For example, parameter identification becomes difficult and models have an increased number of parameters. Thus, researchers typically assume that rainfall is independent. However [Lall, Rajagopallan and Tarboton, 1996] state that if this independent assumption is violated, then the precision of any results obtained are over or underestimated and this leads to incorrect conclusions about the significance of parameters [Dahmen and

Ziegler, 2003]. Past research thus shows that it is important to take the correlated structure of rainfall into account when creating a rainfall model. Generalized estimating equations are especially designed to handle correlated data and past reviews indicate that utilizing this powerful estimating technique may be beneficial to rainfall modeling.

2.8.6 Specification of GEEs

A basic feature of GEE models is that the joint distribution of a unit's response vector y_i does not need to be specified. Instead, only the marginal distribution of y_{it} at each time point needs specification. To clarify, assume there are two time points and the outcome variable is approximately Normal. GEEs only assume that the distribution of y_{i1} and y_{i2} are two univariate Normal distributions, rather than assuming that y_{i1} and y_{i2} form a (joint) bivariate Gaussian distribution. Thus, GEEs avoid the need for multivariate distributions by only assuming a functional form for the marginal distribution at each time point [Hedeker, 2005]. Since the GEE model can be thought of as an extension of GLMs for correlated data, the GEE specifications involve those of GLM, with one addition. Thus, GEE models require the user to specify the following,

- The linear predictor,

$$\eta_{it} = x'_{it}\, \beta \qquad (2,62)$$

where x_{it} is the covariate vector for unit i at time t.

- The link function, used to relate the response variable to the linear combination of the covariates,

$$g(\mu_{it}) = \eta_{it} \qquad (2,63)$$

- The variance as a function of the mean, and consequently the distribution of the response variable,

$$Var[Y_{it}] = \phi V\,(\mu_{it}) \qquad (2,64)$$

- The correlation structure of the response variable.

The fourth condition is what differentiates a GEE model from a GLM [Liang and Zeger, 1986] introduced a 'working' correlation structure to obtain consistent and efficient estimators for regression parameters when observations were correlated.

2.8.7 Working correlation matrix

It is assumed a true correlation between units exists, however it is very rare that this true correlation is actually known. Thus, a working correlation matrix, ***R***, is produced to obtain an estimate of the covariance matrix [Zorn, 2001]. This working correlation is of size $t \times t$ because one assumes that there are a fixed number of time points t at which units are measured. A given unit does not have to be measured at all t time points; each individual's correlation matrix $\boldsymbol{R}_i$ is of size $n_i \times n_i$, with the appropriate rows and columns removed if $n_i < t$.

It is further assumed that the correlation matrix ***R***, and thus $\boldsymbol{R}_{\mathrm{i}}$, depend on a vector of association parameters, denoted by α. That is, the working correlation matrix, now fully defined as $\boldsymbol{R}_i(\alpha)$, is completely specified by the vector of unknown parameters, α. This unknown vector of parameters has a structure which is determined by the investigator and is assumed to be the same for all units. It represents the average dependence among the observations. Although $\boldsymbol{R}_i(\alpha)$ is chosen at own discretion, it is best to try to choose $\boldsymbol{R}_i(\alpha)$ to be consistent with empirical correlations and on the basis of theoretical considerations [Dobson, Puride and Williams, 2003]. This is because accurately representing the correlation matrix improves the efficiency of the GEE estimates. Despite this, there is little information available about how to choose the best correlation structure [Dahmen and Ziegler, 2003], and often it is difficult to determine. As long as μi is correctly specified however, and the covariance matrix converges to some fixed matrix, then consistent results can still be obtained, even if the incorrect $\boldsymbol{R}_i(\alpha)$ structure is identified [Dahmen and Ziegler, 2003]. Finally, any loss of efficiency is reduced as the number of units increases [Dobson, Puride and Williams, 2003]. The most common structures used to model the working correlation matrix are the independent, exchangeable, autoregressive, stationary, non-stationary, unstructured, and fixed correlation structures. The broad range of options available for specifying the correlation structure is another advantage for using the GEE approach. Some of these structures are examined in more detail below.

2.8.7.1 Independent Structure

The independent structure is the simplest form that the working correlation matrix can take, as it assumes that no correlation actually exists and observations within the series are independent. Because users assume that the responses within each unit are independent of each other, this approach sacrifices one of the benefits of GEE in that it does not account for within-subject correlation [Ballinger, 2004]. In general, this structure does not make logical sense for longitudinal data, since such data is usually highly correlated.

[Fitzmaurice, 1995] shows that using an independent structure for correlated data can lead to large efficiency loss of time-varying covariates. Thus, this structure would not be recommended for variables such as rainfall. With this structure, the working correlation matrix becomes the identity matrix, $\boldsymbol{R}_i(\alpha)$ = I, and the resulting GEE is then called the Independent Estimating Equation [Dahmen and Ziegler, 2003]. No estimation of α is required, since no correlation is assumed to exist. This structure does not simply produce the algorithm used for a GLM, as it still involves the 'working' correlation matrix, which a GLM does not. For the independent structure, $\boldsymbol{R}_i(\alpha)$

is defined as,

$$\boldsymbol{R}_{u,v} = \begin{cases} 1, & u = v, \\ 0, & otherwise \end{cases} \tag{2,65}$$

In matrix notation this becomes,

$$\boldsymbol{R}_i = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} \tag{2,66}$$

2.8.7.2 Exchangeable Structure

The exchangeable structure assumes that there is a common correlation within observations. Thus, all of the correlations in $\boldsymbol{R}_i(\alpha)$are equal [Hedeker, 2005]. An exchangeable correlation may be used when each pair of observations within a time frame has approximately the same correlation. For the exchangeable structure, $\boldsymbol{R}_i(\alpha)$ is defined as,

$$\boldsymbol{R}_{u,v} = \begin{cases} 1, & u = v, \\ \alpha, & otherwise \end{cases} \boldsymbol{V} \tag{2,67}$$

In matrix notation this becomes,

$$\boldsymbol{R}_i = \begin{bmatrix} 1 & \alpha & \cdots & \alpha \\ \alpha & 1 & & \alpha \\ \vdots & & \ddots & \vdots \\ \alpha & \alpha & \cdots & 1 \end{bmatrix} \qquad (2,68)$$

2.8.7.3 Autoregressive Structure

For data that are correlated within cluster over time, an autoregressive correlation structure is specified to set the within-subject correlations as an exponential function of this lag period, which is determined by the user [Ballinger, 2004]. The autoregressive structure assumes time dependence for the association between observations and considers each time series to be an AR(m) process. The most difficult task for this structure is determining the correct order of the autoregressive process [Hardin and Hilbe, 2001]. It is common to choose an AR(1) structure, which is defined as $\propto^{|u-v|}$

$$\boldsymbol{R}_{u,v} = \begin{cases} 1, & u = v, \\ \propto^{|u-v|}, & otherwise \end{cases} \qquad (2,69)$$

In matrix notation this becomes,

$$\boldsymbol{R}_i = \begin{bmatrix} 1 & \alpha & \alpha^2 & & \alpha^{n-1} \\ \alpha & 1 & \alpha & \cdots & \alpha^{n-2} \\ \alpha^2 & \alpha & 1 & & \vdots \\ & \vdots & & \ddots & \vdots \\ \alpha^{n-1} & \ldots & & \alpha & 1 \end{bmatrix} \qquad (2,70)$$

2.8.7.4 Unstructured structure

The unstructured form of the working correlation matrix is the most general of all of the correlations discussed in this dissertation as no structure is imposed on the correlation matrix. This form requires all $n_i(n_i - 1)/K$ correlations of $\boldsymbol{R}_i(\alpha)$ to be estimated, and thus when there are many time points this structure becomes very computationally burdensome. An unstructured correlation matrix is used when there is no logical ordering for the observations in the cluster, and is recommended if the number of observations is small in a balanced and complete design [Horton and Lipsitz, 1999]. This correlation matrix is the most efficient structure, but is only useful when there are relatively few observations as its estimate is not guaranteed to be a positive number and there is often a problem with inverting $\boldsymbol{R}_i(\alpha)$ [Hedeker, 2005]. For the unstructured structure, $\boldsymbol{R}_i(\alpha)$ is defined as,

$$\boldsymbol{R}_{u,v} = \begin{cases} 1, & u = v, \\ \propto_{uv} & otherwise. \end{cases} \tag{2,71}$$

In matrix notation this becomes,

$$\boldsymbol{R}_i = \begin{bmatrix} 1 & \alpha_{12} & \cdots & \alpha_{1\mathrm{n_i}} \\ \alpha_{21} & 1 & \cdots & \alpha_{2\mathrm{n_i}} \\ \vdots & & \ddots & \vdots \\ \alpha_{ni1} & \alpha_{ni2} & \cdots & 1 \end{bmatrix} \tag{2,72}$$

2.8.7.5 Fixed Correlation

A fixed correlation structure is fixed at some user-defined value and can be imposed if there is some knowledge of the structure of the correlation matrix from another source [Hardin and Hilbe, 2001]. With this structure, the working correlation is not estimated at each step, but instead takes the correlation as fixed throughout the entire process.

2.8.8 GEE Estimation

As GEEs can be thought of as a moderation in the GLM to incorporate correlated data, it makes sense that they involve a moderation to the estimating or score equation, U_j , used in GLMs (Section 3.3, Equation (3.8)). GEEs are modified by using the ‘working’ correlation matrix in the score equations to account for the correlations in the data [Hardin and Hilbe, 2001]. To begin, the following terms need to be defined in order to setup the score equations for GEE models:

The working correlation matrix, $\boldsymbol{R}_i(\alpha)$ was already defined in section 2.7.7, with α fully characterizing $\boldsymbol{R}_i(\alpha)$. Note that $\boldsymbol{R}_i(\alpha)$ is a $n_i \times n_i$ working correlation matrix for the i unit.

A_i is defined as a $t \times t$ diagonal matrix, with the variance function $V(\mu_{it})$, as the tth diagonal element.

Finally, a working variance-covariance matrix for y_i, which incorporates the ‘working’ correlation matrix and thus the correlations of the data is defined as,

$$V_i(\alpha) = \phi A_i^{1/2} \boldsymbol{R}_i(\alpha) A_i^{1/2} \tag{2,73}$$

This ‘working’ covariance matrix will be equal to $cov(Y_i)$ if $\boldsymbol{R}_i(\alpha)$ is indeed the true correlation matrix for the response variable. It is a transformation of the variance $V(\mu_i)$ term into a matrix form to account for the correlation between observations.

2.8.9 Generalized Estimating Equations Estimator

The generalized estimating equation estimator can now be defined as:

$$U_k(\beta) = \sum_{i=1}^{K} D_i{}^T [V_i]^{-1} (y_i - \mu_i) = 0, \tag{2,74}$$

where D_i is a matrix of partial derivatives of μ_i and β_i (where $D_{it} = \partial\mu_i / \partial\beta_t$), and V_i is the working variance-covariance matrix of y_i. This score equation for estimating β is the solution to a set of k ‘quasi-score’ differential equations [Zorn, 2001], as equation (2.74) only depends on the mean and variance of y_i.

2.8.10 Estimation of $\boldsymbol{\beta}$

The ultimate aim of a GEE is to find the most adequate model to represent a given data set by finding values for the unknown β parameters. To estimate β, the GEE estimator equation (2,74) is rearranged to obtain the following (for the derivation of this formula see Appendix A.1),

$$\hat{\beta} = \sum_{i-1}^{K} \left(D_i{}^T \hat{V}_i{}^{-1} D_i{}^T \right)^{-1} \sum_{i=1}^{K} \left(D_i{}^T \hat{V}_i{}^{-1} y_i \right). \tag{2,75}$$

As GEEs are not a likelihood-based method of estimation, computations based on likelihoods are not possible. Thus, in order to find a solution for Equation (2.67), estimation may be accomplished either via generalized weighted least-squares, or through an iterative process [Zorn, 2001]. Essentially, solving the GEE involves the following steps:

1. Specifying the model parameters of interest and in particular the variable that indicates that the data is correlated, the link function which will ‘linearize’ the regression equation; the distribution of the dependent variable and the structure of the ‘working’ correlation.

2. Computing an initial estimate of β using GLM methodology; thus assuming that observations are independent, with no correlation existing. This is done using GLM estimation techniques.

3. Given the initial estimates of β, computing the Pearson's residuals,

$$e_{it} = \frac{y_{it} - \mu_{it}}{\sqrt{V(\mu_{it})}}$$

(2,76)

4. An estimation of α, to be used in the working correlation matrix, is then computed using the Pearson's residuals and the assumed structure of $\boldsymbol{R}_i$ specified in step 2. It should be noted that the number of nuisance parameters and the estimator of α vary depending on the correlation structure chosen. [Liang and Zeger, 1986] introduced several formulas to calculate α. In addition, even though ϕ appears in all of the following formulas for α, it is not needed to obtain a consistent estimate of β. Different texts use differing methods of calculating α, although most produce very similar values.

5. The working correlation matrix, $\boldsymbol{R}_i$ can now be specified using the α value calculated in step 4 and the assumed structure of $\boldsymbol{R}_i$.

6. Using $\boldsymbol{A}_i$, defined in Section 4.4 and $\boldsymbol{R}_i(\alpha)$, defined in step 5, compute an estimate of the covariance V_i for the K units examined,

$$V_i = A_i^{\frac{1}{2}} \hat{R}_i\,(\alpha) A_i^{\frac{1}{2}}$$

(2,77)

7. Finally, update $\hat{\beta}$ using the following iteratively formula,

$$\beta_{r+1} \;=\; \beta_r + \left\{ \sum_{i=1}^{n} \frac{\partial \mu_i{}'}{\partial \beta} \, V_i{}^{-1} \; \left\{ \frac{\partial \mu_i}{\partial \beta} \right\} \right\}^{-1} \; \left\{ \sum_{i=1}^{n} \frac{\partial \mu_i{}'}{\partial \beta} \, V_i{}^{-1} \, (y_i - \mu_i \right\}$$

(2,78)

8. Complete steps 3 to 6 until convergence.

2.8.11 Calculation of *α*

2.8.11.1 Independent Structure

When no correlation is assumed to exist, and an independence structure

R*(α) = I* is chosen, *α = 1*. Thus no calculation of *α* is required.

2.8.11.2 Autocorrelation Structure

If an autocorrelation structure is chosen as the appropriate ‘working’ correlation matrix, then *α*=(α_1, ..., α_{ni-1}). An estimator of α_t can then be given as,

$$\hat{\alpha}_t \;=\; \phi \sum_{i=1}^{K} \frac{\hat{e}_{it} \; \hat{e}_{i,t+1}}{N - r}.$$

(2,79)

If the structure is specified specifically as an AR(1), then a common *α* is estimated as,

$$\hat{\alpha} = \sum_{t=1}^{n_i - 1} \frac{\hat{\alpha}_t}{n_i - 1}.$$

(2,80)

For the AR(1) structure, all R_i will be identical as this is equivalent to a one-dependent model. Other m-dependent structures can be specified [Hardin and Hilbe, 2001].

2.8.11.3 Exchangeable Structure

When an exchangeable correlation structure is chosen for ***R****(α)*, then *α* can be estimated as,

$$\hat{\alpha} \ = \phi \sum_{i=1}^{K} \ \sum_{t>t'} \hat{e}_{it}\hat{e}_{it'} \ / \left\{ \sum_{i=1}^{K} \frac{1}{2} \, n_i(n_i - 1) - r \right\}.$$

(2,81)

2.8.12 Properties of GEEs

2.8.12.1 Dispersion Parameter, ϕ

The dispersion parameter for a GEE can be estimated by,

$$\hat{\phi} = \frac{1}{N - r} \sum_{i=1}^{K} \sum_{t=1}^{n_i} {e_{it}}^2$$

(2,82)

where $N = \sum n_i$ and is the total number of observations across all units, r is equal to the number of regression parameters, and e_{it} are the estimated Pearson's residuals [Hardin and Hilbe, 2001]. Although most software packages use equation (2,82), some use,

$$\hat{\phi} = \frac{1}{N} \sum_{t=1}^{K} \sum_{t=1}^{n_i} {e_{it}}^2$$

(2,83)

The advantage of equation (2,82) over equation (2,83) is that model results for independent correlation exactly match GLM results. [Liang and Zeger, 1986] state that any consistent estimate of ϕ is admissible.

2.8.12.2 Variance of $\boldsymbol{\beta}$

In order to perform hypothesis tests and construct confidence intervals, it is of interest to obtain standard errors associated with the estimated regression coefficients, β. These standard errors are obtained as the square root of the diagonal elements of the matrix $V(\hat{\beta})$. There are two different ways to calculate the variance of $\hat{\beta}$ within GEE methodology.

The first way is the naive or 'model-based' approach. This approach often underestimates the standard error of $\hat{\beta}$; however it is simple to calculate [Dobson, Puride and Williamsl, 2003]. The second approach is called the robust or 'empirical' estimate, and yields more consistent results even when, $V(Y_{ij})$ is not equal to $\phi V(\mu_{ij})$; and $\boldsymbol{R}_i(\alpha)$ is misspecified. The naive approach gives the variance of $\hat{\beta}$ as,

$$Var(\hat{\beta}) \;=\; \hat{\sigma}^2 \; \left[\sum_{i=1}^{K} D_i^{\,T}\, \hat{V}_i^{\,-1}\, D_i\right]^{-1}$$

(2,84)

The empirical or robust approach gives the variance of $\hat{\beta}$ as,

$$Var(\hat{\beta}) = M_0^{\,-1}\, M_1\;\, M_0^{\,-1}$$

(2,85)

where

$$M_0 = \sum_{i=1}^{K} D_i^T \hat{V}_i^{-1} D_i$$

(2,86)

and

$$M_1 = \sum_{i=1}^{K} D_i^T (y_i - \hat{\mu}_i)(y - \hat{\mu}_i)^T \hat{V}_i^{-1} D_i$$

(2,87)

It should be noted that if $\hat{\sigma}^2 \hat{V}_i = (y_i - \hat{\mu}_i)(y - \hat{\mu}_i)^T$, then the naive and empirical approaches are identical. This second estimator is often called the 'sandwich' estimator.

The consistency of the variance estimate of $\hat{\beta}$ depends on proper specification of the working correlation structure, unlike the actual estimates of $\hat{\beta}$ which do not. Misspecification of the working correlation structure yields estimates of *Var(*$\hat{\beta}$*)* which do not agree with the naive approach. Thus in practice, the robust estimator is nearly always used, since specification of the correct correlation matrix is difficult to achieve [Zorn, 2001]. However, if there are less than 20 units or clusters, the naive approach should be used as it gives better estimates for the variance of $\hat{\beta}$ [Horton and Lipsitz, 1999].

2.8.13 Diagnostics

The main concern is finding a model that adequately describes the data as simply as possible. However, with GEEs the process of selecting model terms and the appropriate correlation structure is complicated by the correlation within observations. As observations are not independent of each other, the residuals are not independent either, and common likelihood-based methods of model fitting either cannot be used or need to be adjusted.

Although GEEs are increasing in popularity and improved research has refined the estimation of these equations, model selection techniques and diagnostics for GEEs has lagged [Ballinger, 2004]. There is still no universally accepted test for goodness of fit for GEE models. None of the diagnostic techniques discussed in the next section are available in any of the major statistical packages, meaning that checking the adequacy of a model is quite difficult.

The next section will outline some of the techniques that can be applied to evaluate gee models. It should be noted that all of the criteria described below are meant only as a guide for when there is no scientific knowledge presented to the researcher. The main techniques discussed are the measures for evaluating the goodness of fit of the model, choosing the best correlation structure, and choosing the best subset of covariates for a given correlation structure. Section 3.6 should be read in conjunction with this. Although diagnostics for GLMs should not be used with GEE models, they are the best approach available for testing the link function and appropriateness of the assumed response variable’s distribution.

2.8.13.1 The best correlation structure

In general, decisions about which correlation structure to use should be guided initially by theory. Despite this, choosing ***R****(α)* on the basis of theoretical considerations is sometimes quite difficult to do [Hardin and Hilbe, 2001]. There is also very little information available about how to chose the best correlation structure. [Hardin and Hilbe, 2001] suggest choosing a correlation structure by initially viewing the following guidelines:

(i) If the number of observations is small, and the design is balanced and complete, use an unstructured correlation structure.

(ii) If the observations in a cluster are collected over time thereby making the clustered data longitudinal data, then the structure should be chosen to be time-dependent, that is, an autoregressive structure.

(iii) If the observations are simply clustered and not collected over time, then an exchangeable structure is advisable.

(iv) If the number of clusters is small, then the independent model may be the best to use.

(iv) If one or more of the above points applies, then use the 'quasi-likelihood under the independence model information criterion' (QIC) to determine the best structure. The QIC is explained below.

2.8.13.2 The QIC

Pan recommends using a *QIC* to select the best correlation matrix for cases in which users may be undecided between two structures [Pan, 2001] The *QIC* is an extension of Akaike's information criterion (AIC) which uses the quasi-likelihood of a model rather than the log-likelihood. The *QIC* is called the 'quasi-likelihood under the independence model information criterion', and as its name infers, no matter what R_i is chosen, this criterion assumes independence: that is, $R = I$. It works by comparing the variance and magnitude of the squared deviances for an independence model to models that assume different sorts of correlation (for example, exchangeable, unstructured and autoregressive). It uses the model coefficient estimates and the correlation in trying to calculate the most appropriate correlation structure. The *QIC* is defined as,

$$QIC = -2Q(y;\ g^{-1}(x\beta_R)) + 2trace({A_i}^{-1}\,V) \tag{2,88}$$

where:

• $Q(y;\ g^{-1}(x\beta_R)$ is the value of the quasi-likelihood, computed using the coefficients from the model with the assumed correlation structure ***R***.

• $\boldsymbol{A_i}$ is the variance matrix of the independence model.

• $\boldsymbol{V_i}$ is the sandwich estimate of the variance using the assumed correlation matrix, ***R****(α)*.

The QIC can then be used to choose between several correlation structures, with the best structure being the one which has the lowest *QIC* value.

2.8.13.3 The best set of covariates to use

There are two methods sometimes employed to find the best subset of covariates to use in a model: the QIC_u, and the marginal R-squared.

2.8.13.4 The *QIC*$_u$

A similar technique to the *QIC* can be used to determine the best covariates to use in a given model. The new measure, called the *QIC*$_u$, is defined as:

$$QIC \ = \ -2Q\big(y;\ g^{-1}(x\beta_R)\big) + \ 2r$$

(2,89)

where $Q\big(y;\ g^{-1}(x\beta_R)\big)$ is the value of the quasi-likelihood, computed in similar fashion to the *QIC* and r is the number of coefficients in the model. The best subset of covariates is then the model that has the lowest *QIC*$_u$ value.

2.8.13.5 Marginal R-squared

Another technique that can be used to determine which subset of covariates is appropriate is an extension of the R^2 statistic, referred to as 'marginal R-square' (R^2). [Ballinger, 2004] and [Zheng, 1988] introduced this statistic to be used with GEE models that have continuous, binary and counted responses. The test measures improvement in fit between the estimated model and the intercept-only model. It does this by comparing two different quantities. Firstly it compares the predicted values produced from the model with the observed values, and secondly, it compares the squared deviations of the observations from the mean values for the response variable. Marginal Rsquare

is defined as follows,

$$R_m{}^2 = 1 - \frac{\sum_{t=1}^{n_i}\sum_{i=1}^{K}(y_{it} - \hat{y}_{it})^2}{\sum_{t=1}^{n_i}\sum_{i=1}^{K}(y_{it} - \bar{y}_{it})^2}$$

(2,90)

$$where, \qquad \bar{y} = \frac{1}{Kn_i} \sum\nolimits_{t=1}^{n_i} \sum\nolimits_{i=1}^{K} y_{it}$$

(2,91)

is the marginal mean across all time periods.

The marginal R^2 is interpreted as the amount of variance in the response variable explained by the fitted model [Hardin and Hilbe, 2001]. It has similar properties as the statistic R^2, with the exception that it can take a negative value when the model gives a less accurate prediction than the intercept-only model [Ballinger, 2004].

2.8.13.6 Analysis of residuals

Residuals are extremely important as a final check to see if the selected model adequately fits the data. However, there are limited techniques available to use with GEEs for checking the adequacy of a model using residuals. The raw residuals and Pearsons residuals are the only residuals that have currently been used to uncover any significant departures in the data. The raw residuals (*rr*) can be found via the simple formula of the observed values minus the predicted values,

$$rr_{it} = y_{\ it} - \hat{y}_{it}.$$

(2,92)

Visual inspection of the residuals and a nonparametric test of the randomness of residuals are the two main methods of determining if the model produced adequately represents the given data. Model assessment is predominantly based on graphical visualizations for GEE models.

One method of checking the adequacy of the model is to use the raw residuals and a nonparametric test to check the randomness of residuals. [Chang, Ji and Li, 1997] suggests using the Wald-Wolfowitz run test to attempt to uncover possible patterns of non-randomness within the raw residuals. The test begins by coding the raw residuals as '1' if the residual is positive, and a '-1' if the residual is negative. This test then assumes a null hypothesis that the signs of the residuals are distributed in a random sequence. It works by examining the sequence of codes produced and the count of the total number of runs of the two codes.

If n_p is the total number of positive residuals, n_n is the total number of negative residuals, and T indicates the number of observed runs in the sequence, then the expected value and variance of T are,

$$E(T) = \frac{2n_p\, n_n}{n_p + n_m} + 1$$

(2,93)

$$V(T) = \frac{2n_p n_n (2n_p n_n - n\,_p n_n\ )}{\left(n_p + n_n\right)(n_p + n_n - 1)}$$

(2,94)

The test statistic for the hypothesis that the signs of the residuals are randomly distributed is,

$$W_z\ = \frac{T - E(T)}{\sqrt{V(T)}},$$

(2,95)

which has an approximately standard normal distribution, and thus the corresponding p-value can be determined using z-tables. Extreme values of W_Z indicate that the model does not adequately reflect the underlying structure of the data and may indicate one of many situations, such as,

(i) the underlying correlation structure has been misspecified;

(ii) the covariates do not adequately represent the data;

(iii) the incorrect distribution has been chosen to represent the response
variable.

2.8.13.7 Graphical Assessment

The first step in the graphical assessment of residuals is to include a graph of the raw residuals and then check for the presence of outlier values that may seriously affect the results [Diggle, Liang and Zeger, 1994]. The model can also be checked to ensure that the raw residuals follow a random pattern and do not form clusters around certain values; this can be further verified by using the Wald-Wolfowitz test described in Section 4.5.3 [Hardin and Hilbe, 2001]. The Pearson residuals can be plotted against the linear predictor and the logarithm of the variance function to further assess model adequacy [Hardin and Hilbe, 2001]. Finally it should be ensured that the raw residuals do not show changes in patterns across the time periods as this could indicate that a different correlation structure is needed.

2.8.13.8 Summary of diagnostics

Overall there are limited diagnostics available to test the adequacy of GEE models. Most tests that can be performed have to be programmed by the analyst as most standard software do not perform the diagnostic tests described in this section. Also, no methods to assess whether the distribution chosen to describe the response variable is adequate or which link function is appropriate, have been described in this literature.

The literature only provides a few model criterion measures to assess overall model goodness of fit. The *QIC* however, is particularly useful for choosing the best correlation structure for a GEE model. Similarly, the QIC_u measure is used for model selection. Standard model criterion measures, such as R^2, are available for GEE models, however it can be difficult to interpret for nonlinear models and experience may be the only method of correctly interpreting the magnitude of R^2 in particular situations. Finally, plots of the raw residuals and Pearson's residuals verse the fitted values, the linear predictor of the variance, can be used to assess a given models adequacy.

2.8.14 Fitting a GEE to a data set

When fitting a GEE model, a user should specify the requirements specified in Section (4.4.1). Details on how to make decisions required to accurately specify these conditions are discussed in turn below. Note that the first two steps are the same as for GLM; see Section (2.6.5).

2.8.14.1 Step 1 & 2: Linear predictor and best link function

To model the expected value of the marginal response for the population $\mu_i = E(y_i)$ to a linear combination of the covariates, the user must specify a link transformation function that will allow the response variable to be expressed as a vector of parameter estimates (β) in the form of an additive model [McCullagh and Nelder, 1989]. The choices available for the link function depend primarily on the distribution specified, and a list of these available with GEE models can be seen in Table 2.3. This table gives the distributions and corresponding link functions currently available with GEE models in most statistical packages. Note that the Tweedie distribution does not appear here; it is not yet available with GEEs in any statistical packages.

Table 2.3
Link function currently available with GEE models

Distribution	Link Functions	Brief Description
Normal	Identity Link	This fits the same model as the GLM
	Power Link	Any power transformation
	Reciprocal Link	Links using reciprocal of response variable
Binomial	Logit Link	Fits logistic regression models
	Probit Link	Fits cumulative probability functions
	Power Link	Any power transformation
	Reciprocal Link	Links using reciprocal of response variable
Poisson	Log Link	
	Power Link	Any power transformation
	Reciprocal Link	Links using reciprocal of response variable
Negative Binomial	Power Link	Any power transformation
Gamma	Power Link	Any power transformation
	Reciprocal Link	Links using reciprocal of response
Multinomial	Cumulative Logit Link	

2.8.14.2 Step 3: Distribution of the response variable

The next step involves specifying the distribution of the outcome variable so that the variance might be calculated as a function of the mean response calculated in step 1 and 2 [Hardin and Hilbe, 2001]. GEEs, like GLMs, permit the specification of distributions from the exponential family of distributions, including the Normal, inverse Normal, binomial, Poisson, negative binomial, and gamma distributions.

Misspecifications of the variance function, and thus the response distribution, can have important consequences and lead to incorrect statistical conclusions [Ballinger, 2004].

In fitting a GEE (or any GLM), the user should make every reasonable effort to correctly specify the distribution of the response variable so that the variance can be efficiently calculated as a function of the mean and the regression coefficients can be properly interpreted [Ballinger, 2004].

2.8.14.3 Step 4: Form of the correlation

The final step involves the specification of the form of the correlation of responses within units or nested within a group in the sample. Even though GEE models are generally robust to misspecification of the correlation structure, it is still important that the user takes precautions in specifying this structure. This is because a structure that does not incorporate all of the information on the correlation of measurements within the cluster may result in inefficient estimators [Ballinger, 2004]. The form of the correlation structure should be chosen from one of the structures described in Section 2.8.7.

2.8.14.4 Step 5: Fitting the model and diagnostics

A GEE model can now be fitted to the data; however this usually takes considerable time and effort. Finally, and often most importantly, the model should be checked to see if it is adequate and justifiable using numerous diagnostic techniques (Section 2.8.13).

2.8.15 Cautions regarding GEE

There are a few cautions that users should be aware of when fitting a GEE model. Firstly, users should be cautioned that using the robust approach to estimate the variance of β could be highly biased when the number of units

or clusters examined is small. [Horton and Lipsitz, 1999] suggest that the GEE robust variance estimate should only be used when there are more than 20 units or clusters, that is, K should be greater than 20. If a data set contains fewer than 20 units, the naive approach to estimating the variance should be used, as it gives better estimates for the variance of β.

Secondly, although some researchers use the Wald chi-square statistic for model comparisons [Hedeker, 2005] and many current statistical packages produce a deviance or chi-square statistic for a GEE model using this technique, such a statistic is only interpretable under certain unrealistic conditions. Thus, it is not recommended for use to test whether all of the variables in the estimate are different from one another and different from zero [Ballinger, 2004]. It is not interpretable when a user wants to model correlations using the auto-regression correlation structure. Furthermore, this statistic is sensitive to large differences in the scale of different independent variables [Ballinger, 2004].Thus this type of statistic is not suitable for this study.

2.8.16 Advantages of GEEs

The major, and most obvious advantage of GEEs is they can be used to model non-Normal, correlated longitudinal data. This makes GEEs an invaluable tool when analysing data that was previously modelled using uncorrelated models. This advantage is further strengthened by the broad range of options available that help specify the correlation between observations through the working correlation matrix. The incorporation of explicit knowledge about within-unit interdependence makes GEEs even more attractive [Zorn, 2001]. As well as the production of more efficient estimates of regression parameters due to the inclusion of the correlation, GEEs also produce reasonably accurate standard errors and hence, reasonably accurate confidence intervals with the correct coverage rates [Hanley, Edwardes, Negassa and Forrester, 2003]. Another advantage is that even if an incorrect working correlation matrix is specified, it is still possible to obtain consistent parameter estimates for $\hat{\beta}$ that are asymptotically Normally distributed, provided the mean μ_i has been correctly specified as a function of all possible explanatory variables x_i [Dahmen and Ziegler, 2003]. This is a clear advantage, as understanding the relationship of the correlation is often quite difficult [Zorn, 2001]. Also the GEE approach has some built-in robustness as it requires no specification of the full likelihood of the response variable's distribution. As GEEs are an extension of GLMs, they allow the outcome variable to taken on numerous different forms, such as continuous, dichotomous, polychotomous, ordinal, or even count data. This makes their practicality even greater [Zorn, 2001]. Finally, as GEEs are becoming increasingly popular, more readily available packages have incorporated GEEs into their programs making the computations much easier.

2.8.17 Limitations of GEEs

GEEs are gaining popularity, however there is some evidence that the use of an incorrect dependence structure within the GEE approach can produce worse results than if using an independent structure to model correlated data [Sutradhar and Das, 1999]; [Crowder, 1995]. It has been further commented that solutions for $\hat{\alpha}$ may not exist for various reasons, leading to the complete breakdown of the estimation of the regression parameters. [Cologne, Fujita, Carter and Ban, 1993] also found that when the true correlation structure was quite simple (for example exchangeable), then GEEs were quite efficient.

However, when the structure is more difficult, the efficient results are often not obtained if the correlation structure is wrongly specified. In the case when the correlation structure is complicated, then every effort should be made to approximate the true correlation structure correctly, as consistent results are not obtained when the correlation structure is wrongly specified.

2.8.18 Handling Missing Data

One limitation with using GEEs to estimate parameters is that incomplete data sets can complicate the analysis. Often data sets have missing data, such as when rainfall is not recorded on a particular day. If data is missing completely at random (mcar), consistent results can still be obtained; however the notation and calculations used become more complicated [Horton and Lipsitz, 1999]. In particular, the estimation of the working correlation matrix becomes quite tedious.

A series of approaches, when data is missing in the dependent variable, has been proposed recently. However, these methods are rarely used as they are extremely difficult and they are not available in accessible form with standard software [Dahmen and Ziegler, 2003]. Also, the analysis of a data set that contains missing observations produce differing results between differing packages [Horton and Lipsitz, 1999]. The three data sets that will be used in this dissertation do not have any missing data and thus this limitation is avoided.

2.8.19 GEE Software

The GEE algorithm has been incorporated into many major statistical software packages, including SAS, STATA, HLM, LINDEP, GAUSS, SUDANN,R, and S-Plus. However most of the packages are restricted to only modeling a limited number of response outcome distributions (Table 4.1). Further advancements in the area of GEE software is continuously occurring, and existing software is being constantly revised and updated to include new research. For an overview of software packages offering GEE methodology. See [Zorn, 2001] and [Horton and Lipstiz, 1999].

2.8.20 Summary

This section has described the GEE approach for modeling longitudinal and correlated data. This approach has several features which makes it particularly useful and popular. Because it is a generalization of GEE, many types of dependent variables can be accommodated within the GEE family of models. Also, the selection of the variance-covariance matrix is not as critical as with other models because GEEs provide standard errors that are robust to misspecification of the variance-covariance matrix. This is an attractive feature, especially for situations where the scientific interest is in estimation and inference of the regression parameters and not of the variance-covariance structure. The converse of this is that if there is scientific interest in the variance-covariance structure of the longitudinal data; then GEEs are not appropriate (at least in its GEE1 implementation). [Liang and Zeger, 1986] applied the name ‘GEE´ to emphasize the nature of the generalization of the original estimating equation due to the focus on the marginal distribution. These models do not start with a probability-based model, or likelihood. There is an implied quasi-likelihood form to the GEE model which may or may not coincide to a probability-based model. The GEE model was extended assuming a correlation structure that was estimated by combining information across panels. The ancillary parameter (α) was estimated to get a working correlation matrix. By applying the correlation matrix to each unit, the β regression coefficients can be estimated. Thus, the focus is on the marginal distribution, where the units are summed together after taking into account the correlation.

CHAPTER 3

3.0 MATERIALS AND METHODS

3.1 Description of Data

There are four major rainfall data sets used in this study: Dagoretti, JKIA, Nairobi mean and the CRU Kenya rainfall.

3.1.1 Dagoretti data

Dagoretti data consists of daily gauge measurements taken at Dagorreti station for the period beginning in February 2nd 1959 to December 2005.

3.1.2 Jomo Kenyatta International Airport (JKIA) data

JKIA data is the gauge readings taken daily at JKIA for the period 1959 to 2005.

3.1.3 Nairobi Monthly Mean

Nairobi Monthly mean rainfall values for the period 1959-2005 was calculated as the mean of Dagoretti and JKIA data for the same period. The mean is taken as the representative Nairobi rainfall values for the period. This data consists of 563 values. In this study, Nairobi mean is referred to as KenMet rainfall.

Figure 3.1 shows a boxplot showing the relative distribution of mean Nairobi monthly rainfall.

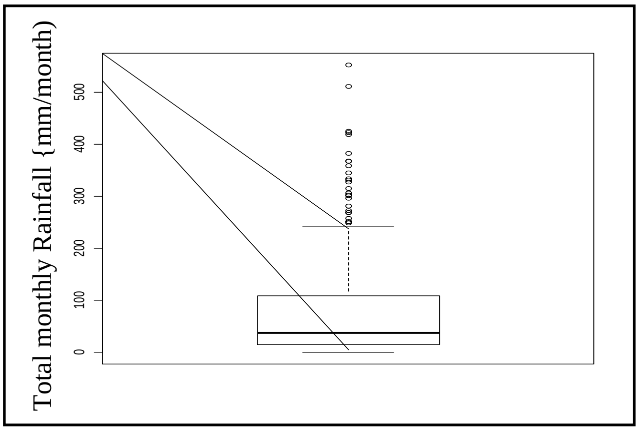


Figure 3.1
Nairobi mean-monthly Rainfall.

Summary statistics for the three stations are as follows:

Table 3.1

JKIA, Dagoretti and Nairobi mean monthly rainfall.

Summary Statistics.

	mean	sd	skewness	kurtosis	0%	25%	50%	75%	100%	n
DAGO	84.03	97.67	1.97	4.84	0	15.80	45.0	116.35	622.6	564
JKA	61.64	82.35	3.13	17.97	0	7.95	32.0	89.65	841.0	564
NAIROBI	73.64	85.41	1.96	4.77	0	14.51	37.6	108.06	553.3	564

From Table 3.1 it is evident that mean Nairobi rainfall as well as with the other two datasets are skewed to the right.

A histogram in Figure 3.2 shows the distribution.

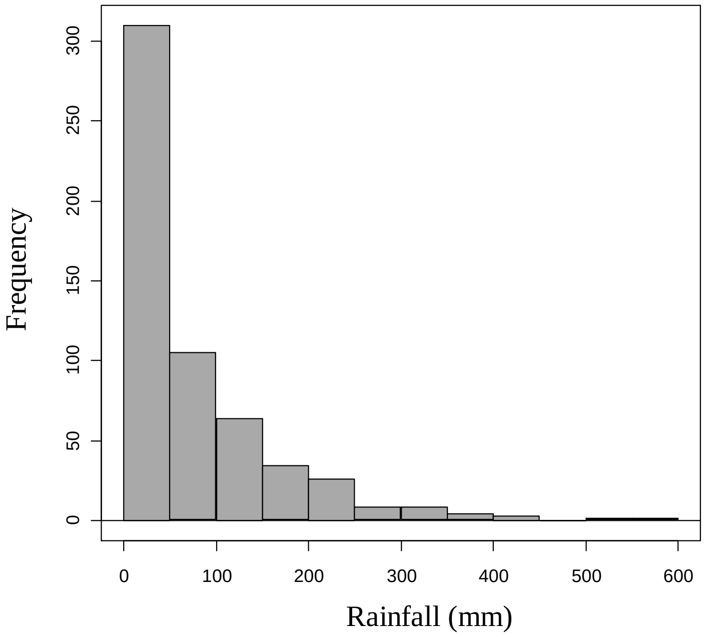


Figure 3.2
Histogram of Nairobi monthly Rainfall

Most months (311 out of 564) receive between 0mm and 50mm of rainfall monthly with 18 of them receiving exact zero mm. As there were some months where the rainfall was recorded as zero or that the amounts were not recordable, a model that combines the monthly rainfall occurrence and rainfall amounts is necessary. A histogram of Nairobi monthly rainfall amounts in Figure 3.2 shows that the majority of months experiencing less than 100mm of rain. No obvious outliers can be seen.

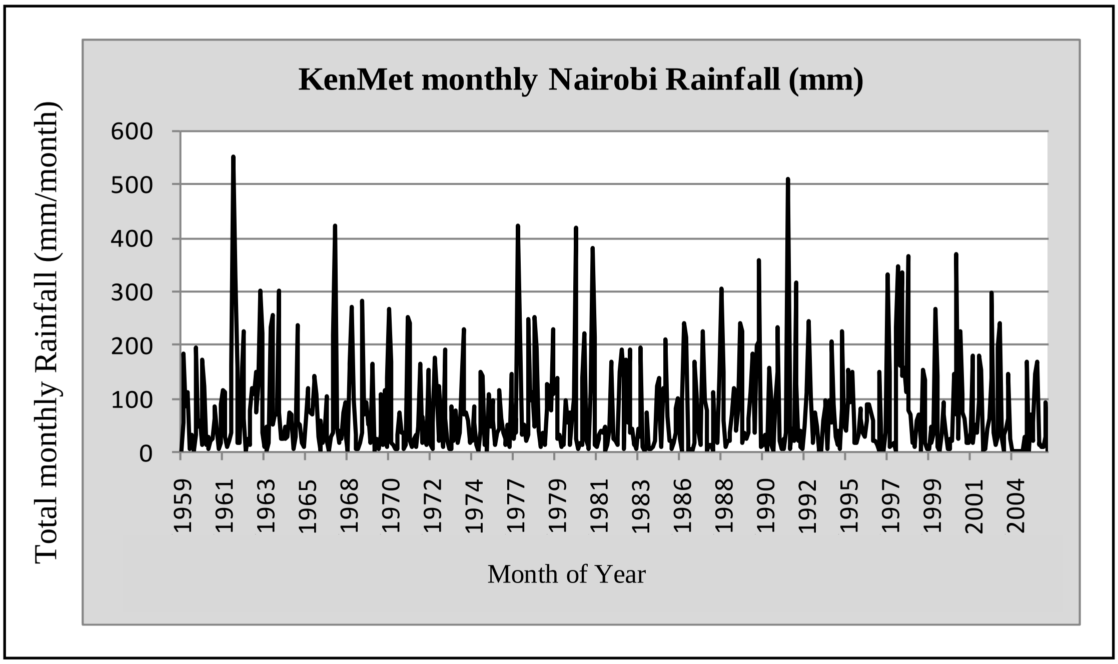


Figure 3.3
KenMet monthly Rainfall 1959-2005

The monthly data has a mean of 73.6 mm and the median amount was 37.65 mm. This supports the conclusion that Nairobi rainfall data is skewed (the mean and the median are very different). With the removal of the high values, the mean and standard deviation do not change significantly. It changed to 72.01 and the standard deviation to 81.03. The three months were therefore not excluded as outliers in the analysis.

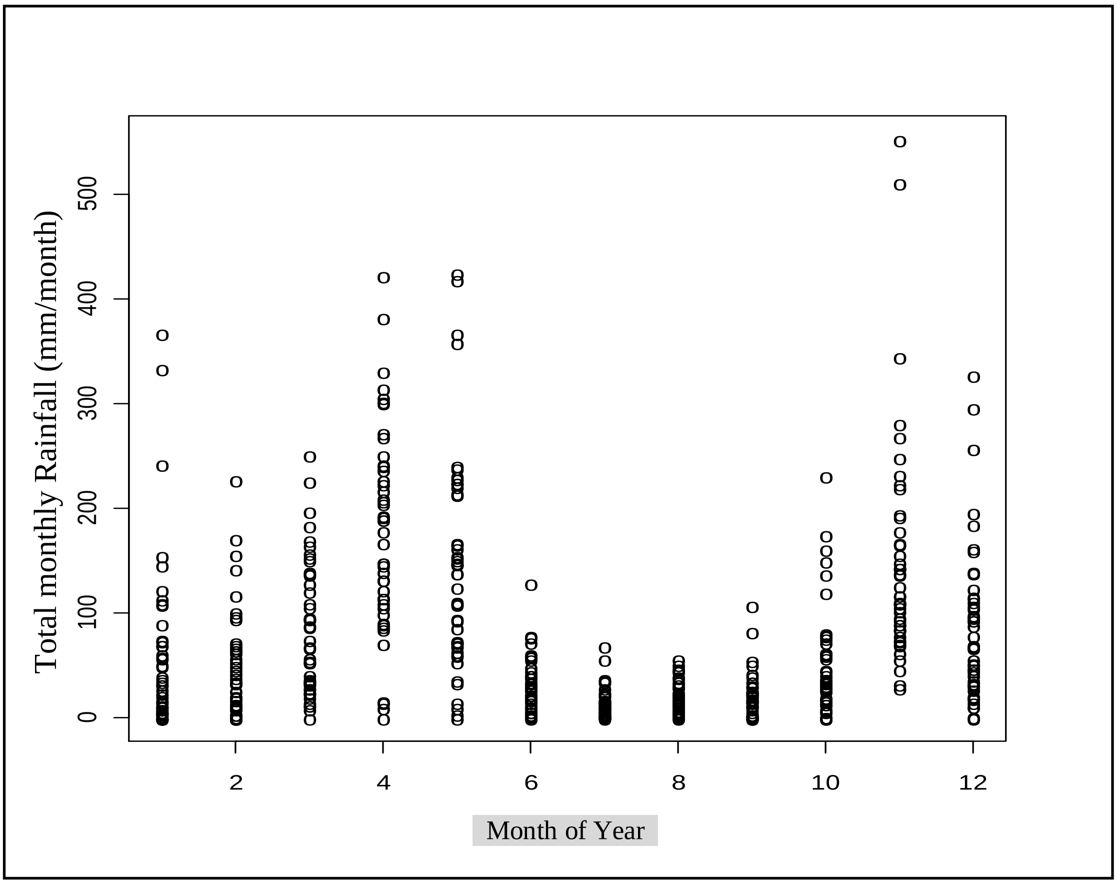


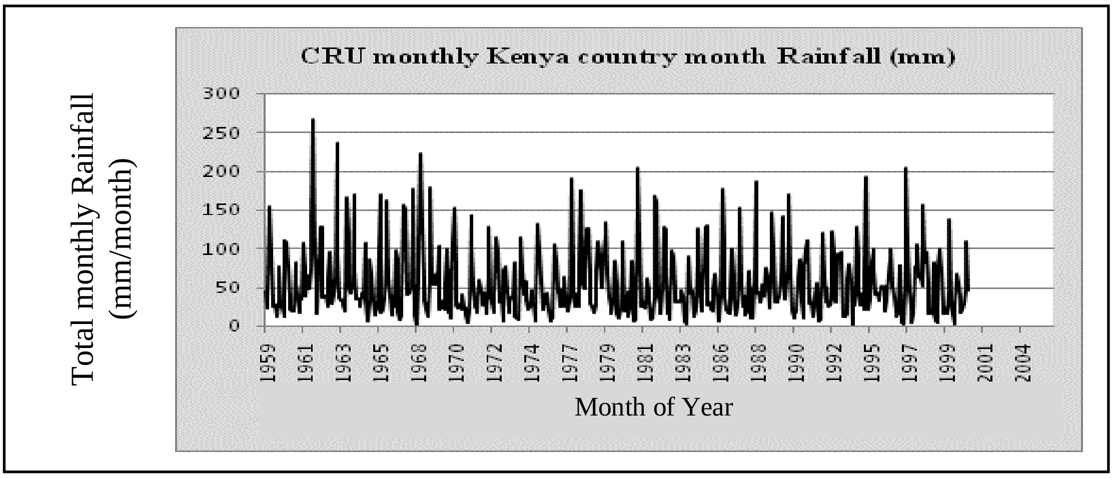
Figure 3.4
Nairobi monthly 1959-2005 bimodal variation.
The bimodal cycle begins in February and ends in January of the following year.

The monthly precipitation shows outstanding peaks; one in Nov.1961(553.3mm) and Nov 1991 (511.9). We seek to find the cause(s) of these peaks. Kenya rainfall shows a bimodal annual distribution with two peaks, as shown in Figure 3.4. The peaks are the two rain seasons; the short rains (October, November, December) and the long rains (March, April, May).

3.1.4 CRU Kenya dataset

The University of East Anglia, UK, provided research data. Datasets are managed by a variety of people and projects within CRU. Some were available on-line; others were requested from the person responsible for them. Files ending in ".gz" were decompressed using *gzip* (most platforms) or *jZip* (Windows). The various datasets on the CRU website were provided for all to use, provided the sources were acknowledged. Acknowledgement was done by citing the papers referenced on the references page. The website was also be acknowledged as deemed necessary. CRU endeavours to update the majority of the data pages at timely intervals although this cannot be guaranteed by specific dates. In this work we have used data supplied and maintained by and owned by Dr. Tim Mitchell. The data provides monthly, seasonal, and annual climate observations averaged for political units of the world. Kenya monthly means (1901-2000) was used in the analysis.

The country aggregation is based on the CRU TS 2.0 gridded data-set. The gridded data were aggregated into countries using political boundaries according to [Mitchell Hulme, and New, 2002]: Climate data for political areas. Area 34:109. This data has been analysed and the results compared with KenMet data [Mitchell, Hulme, and New, 2002, 2002]. This data is referred to here as CRU Kenya. Figure 3.5 shows CRU Kenya averaged monthly rainfall. This dataset has the advantage as modeling data because of its long time series of 100 years.

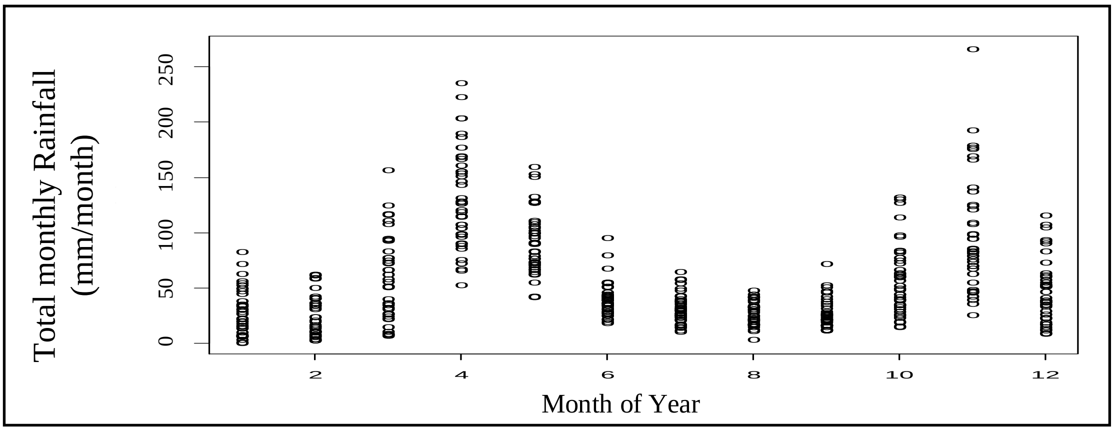


<u>Figure 3.5</u>

CRU Kenya country averaged rainfall 1901- 2000.

N	Mean	Std.Dev	Min	Q1	Median	Q3	Max
504	56	44	1.3	25	41	76	267

The means and the standard deviations are very different between CRU Kenya and KenMet rainfall. They measure different quantities. CRU Kenya is averaged for the whole region while KenMet is for a single location. CRU Kenya is more representative for the country.



<u>Figure 3.6</u>

CRU Kenya country averaged rainfall 1901- 2000 showing bimodal variation

CRU Kenya data also exhibit the bimodal annual characteristics. The only difference being that the minimum values are higher than KenMet while the maximum values are lower than those of KenMet as seen Figure 3.6.

3.1.5 NASA

NASA has provided invaluable information on eclipse data at the eclipse page;

http://eclipse.gsfc.nasa.gov/eclipse.html

3.1.5.1 Solar and lunar Declinations and Distances

Solar and Lunar declination can be obtained from NASA [NASA Solar].

Figure 3.7 shows the variation of lunar distance from the earth for the period 1995to2006.

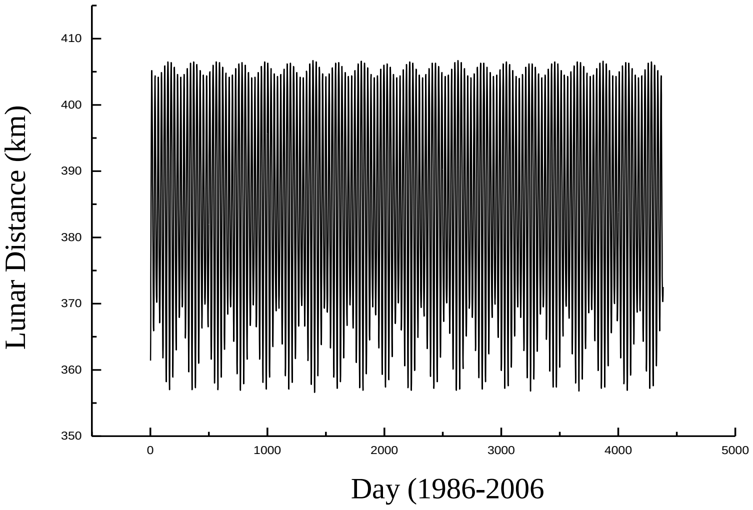


Figure 3.7
Lunar distance.

Lunar distance oscillates with a period of one synodic month of 29.25 solar days causing atmospheric ocean tides of similar frequency.

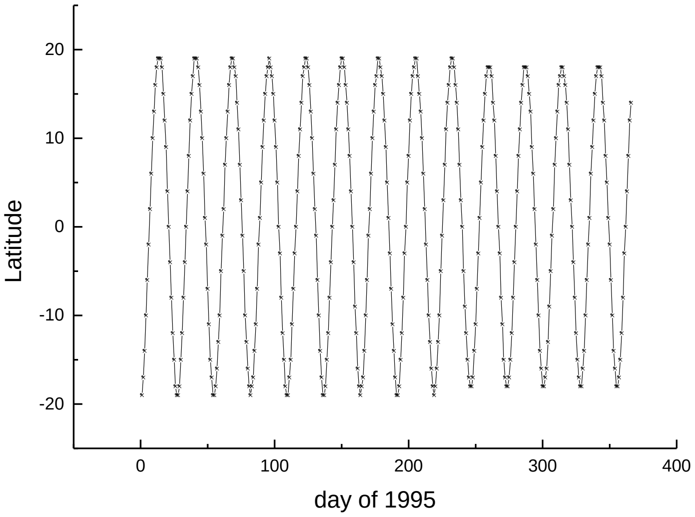


Figure 3.8
Lunar position in 1995.

Figure 3.6 shows the lunar position on each day for the first days of 1995.The wavelength of the oscillation is still 29.25 days while the amplitude varies with an all-time maximum of 28degrees. The maximum in Figure is 18 degrees

3.1.6 The Nyahururu Experiment.

Data for the astronomical aspects of this project have been obtained from NASA except one – the one taken at Nyahururu on December 4th, 1983.

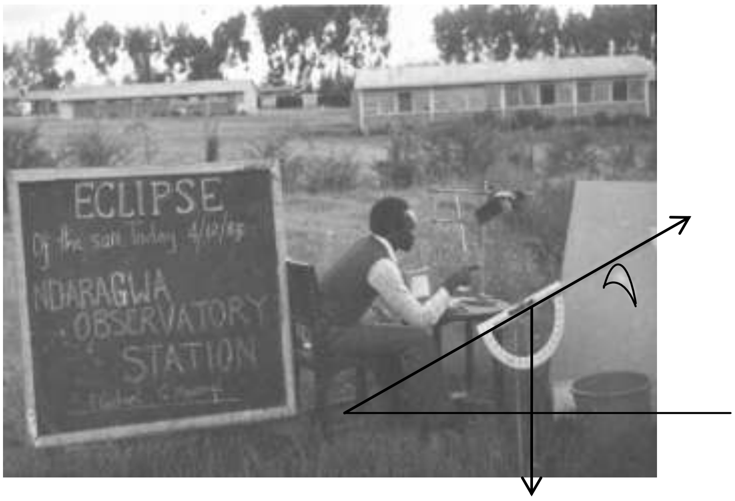


Figure 3.9
Taking measurements during the 1983 solar eclipse event. (Photo by Gachari F.)

The photograph in Figure 3.9 was taken by the author and shows an observer taking some measurements. In this experiment, several measurements can still be confirmed from the picture) e.g, Calendar Date (Gregorian), type of eclipse, time of eclipse; etc. The Nyahururu station is located directly at the equator (0N, 36E) and the high altitude together with the leeward side of the Aberdare Mountains ensured that visibility was not obscured by prevailing clouds. The tools used on that day and their readings are shown on the Table 3.2.

Table 3.2

Validation of NASA values

Instrument	Measurement	Value	NASA value
Clock	Time	evening	(RA)16h41m26.2s
Clinometers	Azimuth angle	29^{o}	4.41pm. Local Time
Camera/Mirror	Sun's Image on screen	40% visible	(Gamma=0.4015)
Date	signboard	December 4, 1983	December 4, 1983
White Screen	Type of eclipse	partial	partial

As a case study, these values were found to agree with NASA values used in this study.

3.2 MODELING NAIROBI RAINFALL

The model developed in this study is of the form;

Response variable ~ predictor(s).

Where the response variable is the monthly rainfall and the predictors are a sum of factors and covariates described below.

3.2.1 Fitting a GEE to Nairobi Rainfall data

The advantage of modeling rainfall using a GEE instead of a GLM were discussed in section 2.7.; that rainfall data is correlated.

The steps involved are summarized as follows:

1) Linear predictors were chosen and the best link function identified.

2) The distribution of the response variable was chosen and a GLM was fitted which gave the initial estimates of the fit parameters. The choice was the Tweedie distribution. To specify the Tweedie, The mean, μ the dispersion parameter, ϕ and the variance power, p were required. Standard algorithms in R-software calculate μ and MLE was used to work out ϕ and p.

3) A GLM was fitted; on the response variable to obtain the initial estimates for the fit parameter α, ϕ and β.

4) A GEE was then fitted by using the initial fit parameters to obtain new updated values of fit parameters β by means of a variance-covariance matrix.

5) Diagnostics tests were perform to assess the appropriateness of the fit.

3.2.2 The Predictors

Predictors used initially were:

i) Solar declination

ii) Lunar declination.

iii) Month of the year.

iv) Sunspot numbers

The following variables were then added to the set of four predictors above. They are defined in this study:

v) Lunar phases.

vi) Atmospheric Tide phases.

vii) Maximum Lunar declination values.

viii) Gravity atmospheric tide state, atide.

ix) Enhanced gravity atmospheric tide state, etide.

3.2.3 Calculating the Predictors

3.2.3.1 The Month

The month represents any one of the twelve months of the year. It indicates the month during which the rainfall measurement was taken.

3.2.3.2 SOI Phase

The SOI Phase represents the state of the SOI values as described in section. It takes one of I to V values. The five Phases of the SOI are as follows (see section 2.3.4):

I - consistently negative, **II** - consistently positive **III** - rapidly falling **IV** - rapidly rising **V** - consistently near zero

3.2.3.4 Sunspot Phase

Sunspot phases are categorised in this study according to the number of sunspots recorded in the month. The categories are Cold (C),Warm(W), Hot(H) and very Hot(vH) where

0<C<=50, 50<W<=100, 100<H<=150, 150<vH.

3.2.3.5 Lunar Phase (*lunaph*)

Lunar phases are obtained from lunar ephemeris available from NASA [NASA]. The phase of the moon is determined by the fraction (represented by the angle in degrees between 0 and 180) lit by the sun and visible from the earth. The lunar phases were classified in three ways:

a) New-moon(NM), Half-moon(HM) and Full-moon(FM). where 0<NM<=45, 45<HM<=135, 135<FM<=180.

b) First-quarter(Q),second-quarter(2Q), third-quarter(3Q) and fourth-quarter(4Q). where 0<1Q<=45, 45<2Q<=90, 90<3Q<=135 and 135<4Q<=180.

c) Denoted by digits 1 to 4 depending on the phase of the moon beginning with first quarter represented by digit 1.

3.2.3.6 Lunar Declination (*ldec*)

The declination of the moon as observed from Nairobi. The value of the lunar declination at mid-month. Values are from -29 to 29.

3.2.3.7 Solar Declination (*sdec*)

The mean solar declination angle between -23^{o} and 23^{o} during the month. The variable represents the season.

3.2.3.8 Atmospheric Tide (*atide*)

The frequency in terms of the number of days in any other month when the magnitude of the angular difference between *sdec* and *ldec* is less than three degrees. This is the condition for atmospheric tide occurring anywhere within latitudes -28.5and 28.5 any time of the year. Atide values range from 0 to 8.

3.2.3.9 Enhanced Tide (*etide*) Enhanced Atmospheric tide is the frequency in terms of the number of days during which the air tide state occurs at a reduced magnitude of the angular difference of one degree or less. (0 to 2).

3.2.3.10 Atide Phase (*atideph*)

Atide phase. This is the measure of the prevalence of the tidal state. Measured by the number of days in the same month during which the tidal state prevails.(L-low-0to3 days, M-medium-3-6 days, H-high-more than 6days. Atideph may neither covary or cofactor with atide. The two are not mutually independent

3.2.3.11 Maximum Lunar Declination (*mld*)

The maximum amplitude in degrees of the lunar declination during the month. Mld ranges from 18 to 29.

3.2.3.12 The synodic decimal *(synod)*

The decimal value between 0 1 and 1 representing the percentage of the lunar surface lit by the sun as seen from a terrestrial observer at mid-month.

3.2.3.13 Perigee distance *(prg)*

The integral value of the turning point lunar distance in kilometers during the perigee of the month.

3.2.3.14 Apogee distance *(apg)*

The integral value of the turning point lunar distance in kilometers during the apogee of the month.

3.2.3.15 Sunspot Number *(ssn)*

The number of sunspots visible on the solar surface in the month.

3.3 SOFTWARE USED IN THIS STUDY

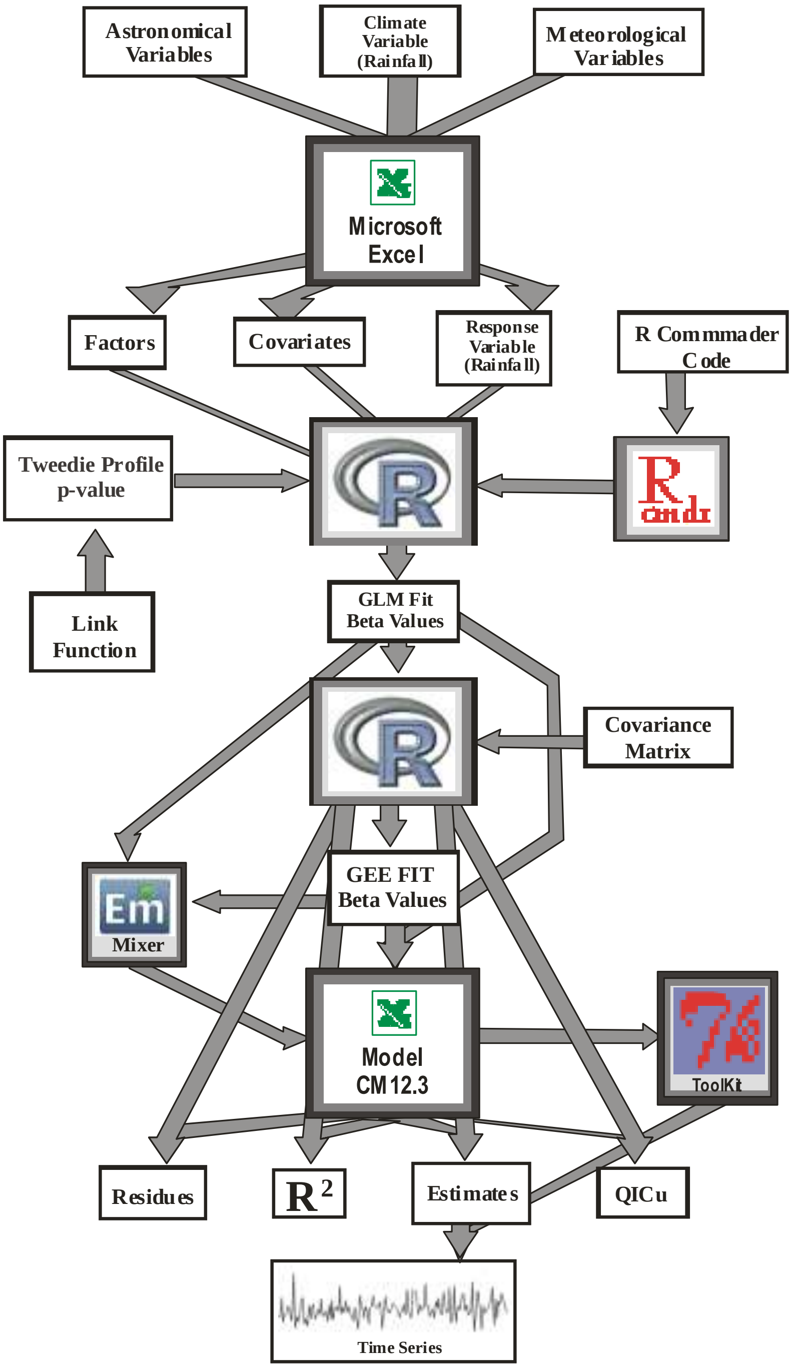


Figure 3.10
Software used in this study

In a Microsoft Excel worksheet vectors representing the astronomical variables – solar and lunar declinations, and their derivatives; atide and etide as well as the standstill states were entered. The spreadsheet may also contain any number of other explanatory variables that may be desired. The explanatory variables variables are defined in section 3.3.2.

Once the vectors have been entered in the spreadsheet, they were categorized as either factor or covariate. The selected response variable was also included as one of the vectors.

We fitted the response variable-rainfall with a GLM of the Tweedie family. This was achieved first by running the *tweedie.profile* routine available in the R distribution and obtaining the p-value necessary for performing the GLM. The R commands necessary to carry out these steps were written down and ran as a batch in the Rcommander program. They may be run directly on the R console one by one. A link function was specified for the GLM as the log link.

Upon fitting the response variable with a GLM determined by the selected explanatory variables the fit output includes among others, the beta values which represent the coefficients of the linear equation of the fit together with the intercept. These beta values were used to calculate the estimates and perform diagnostics which determine the quality of the. In this study, the quality of the fit involved plotting the fit values along with the measurements for a visual assessment as well as calculating the marginal R squared value for the fit. Raw residual plots as well as plots of the Pearson residuals were obtained to further assess the goodness of the fit.

A QQ plot was also performed and plotted. To assess the appropriateness of the explanatory variables the QIC_u value for the fit was obtained.

Rainfall data is often correlated. It is advisable to perform a GLM to satisfaction and then take a step further and fit a GEE on the response variable. To perform a GEE we used the methods described in section 2.8 with our choice of covariance matrix as AR(1). That is how new beta values were obtained for the calculation of estimates. The bet values are used to construct the equation for the estimate for each month in the period 1901 to 2050. In this study the model so obtained has been named Climate Model 12.3 (CM12.3) and each estimated monthly value is calculated by means of a 100 terms equation.

A program EM mixer was obtained from ConvexDNA [Convexdna, 2014]. This program which is a slider is useful for carrying out a what-if analysis on the estimates. It may be used to improve the quality of the fit by adjusting, the values of the coefficients. A useful set of plotting software is available as a toolkit as was used in plotting and analyzing the model output [Toolkit].

3.4 THE FITTING PROCEDURE

As discussed in Section 2.7.14, this model design is based on fitting a GLM of the Tweedie family to the Nairobi rainfall distribution. R Statistical Software was used to fit a GEE and obtain the beta values of the fit. The beta values describe how the response variable relates with the predictors in the equation:

rainfall ~predictor(s)

The specific steps used in the fitting procedure were as follows:

Step 1. Initial estimate of beta using GLM were computed by;

a) Calculating *p* to be used in the variance function of the Tweedie distribution using profile likelihood function. An R-program *tweedie.profile* was used for this calculation.

b) A Tweedie GLM model was fitted to the data with the *p* value found by using the profile likelihood function and the log link function.

At this point it was possible to use the beta values and calculate rainfall estimates using the beta values obtained from the GLM fit. However, rainfall data is correlated and therefore it was necessary to fit a GEE instead. To do this, it was necessary to use a correlation matrix of choice to calculate new beta values. We chose AR(1) correlation matrix and fitted a GEE by:

a) Calculating new phi value and alpha.

b) Calculating an estimate of the covariance matrix.

c) Calculating new set of beta values.

Using the new beta values, it was then possible to work out rainfall estimates for the period 1901 to 2050. This was made possible because factors involving solar-lunar geometry e.g. *sdec* and *ldec* were available from NASA for this period. The code that was used to perform the GEE fit is available as Appendix I.

CHAPTER 4

4.0 RESULTS AND ANALYSIS

In this chapter we describe how the fitting process was used to design rainfall models and the results obtained. We begin with a simple model having a single predictor-month and then proceed to obtain models having multiple predictors. Towards the end of the chapter the models are used to obtain projected monthly and annual values of rainfall.

4.1 SINGLE PREDICTOR MODEL

We now use a simple model with only one predictor to describe the fitting process.

4.1.1 Step 1.

In this simple model which illustrates the steps followed, only one predictor is used; the month of the year. The link function is the log link.

4.1.2 Step 2

To obtain variance power, *p* we use R function- *tweedie-profile* .The initial p-values are selected so that the requirement 1<p<2 (section 2.7.5) is met for Tweedie function and specified in *p-vec* in the R command.

```
>power=tweedie.profile(rain~month,p.vec=seq(1.5,2.0,length=10),do.plot=TRUE,do
.smooth=TRUE,do.ci=TRUE,method="interpolation")
>p=power$p.max
```

The p-values are shown in the log-likelihood plot in Figure 4.1 below.

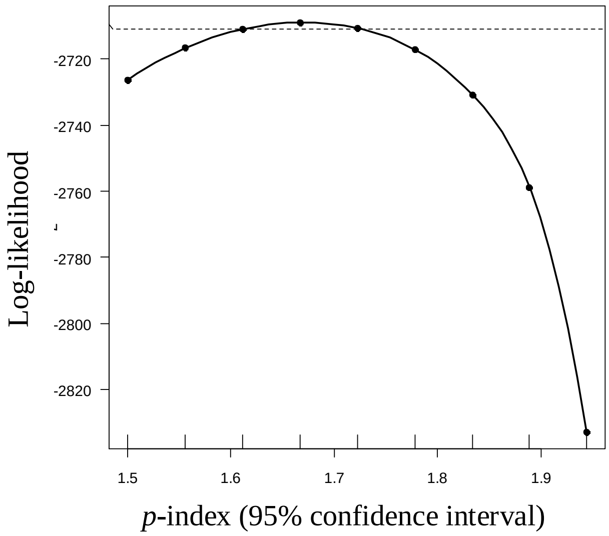


Figure 4.1
Log-likelihood plot for the p-value.
Nairobi monthly rainfall and the month predictor at 95% confidence. The p-value at maximum likelihood is 1.672336.

4.1.3 Step 3 Fitting a GLM

To fit the rainfall data to a GLM the following R command was used;

```
>glmmodel<-glm(rain~month,family=tweedie(var.power=p,
link.power=0),x=TRUE)
```

Fitted values (RED) were then plotted against the measured values (BLUE) as can be seen in Figure 4.2 below.

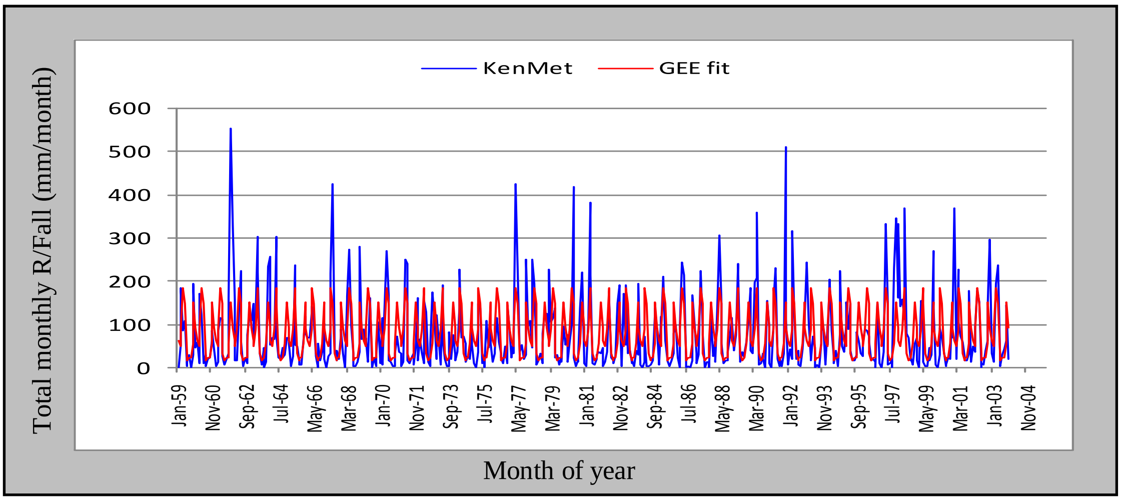


Figure 4.2
GLM Fit

GLM fit beta values were:

```
[1,]  (Intercept)     4.0861611
[2,]  month[T.2]      -0.2388545
[3,]  month[T.3]     0.3821451
[4,]  month[T.4]     1.1236528
[5,]  month[T.5]     0.9051177
[6,]  month[T.6]    -0.6513141
[7,]  month[T.7]    -1.4282353
[8,]  month[T.8]    -1.1152909
[9,]  month[T.9]    -1.0112070
[10,] month[T.10]  -0.1023758
[11,] month[T.11]   0.9284366
[12,] month[T.12]   0.4220864
```

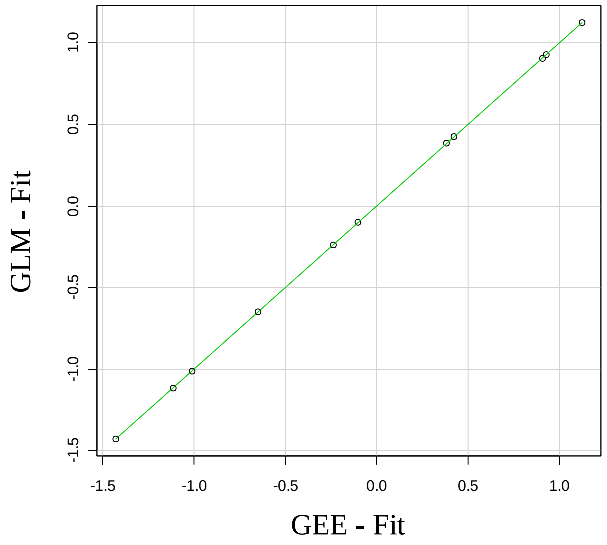
The value of phi was 2.92515. It is evident from figure 4.1 that the fit was not good enough. The model was unable to capture values greater than 200 mm/month calling for further model improvement.

4.1.4.Step 4 Fitting a GEE

Using the beta values obtained in the GLM fit, it was now possible to calculate values of alpha to be used in the variance-covariance matrix required for a GEE fit. Upon fitting a GEE fit, diagnostics were performed to assess the appropriateness of the fit. A new set of beta values are now available as follows;

```
[1,]  (Intercept)     4.0861219
[2,]  month[T.2]   -0.2388154
[3,]  month[T.3]     0.3821844
[4,]  month[T.4]     1.1236921
[5,]  month[T.5]     0.9051569
[6,]  month[T.6]    -0.6512749
[7,]  month[T.7]    -1.4281961
[8,]  month[T.8]    -1.1152517
[9,]  month[T.9]    -1.0111678
[10,] month[T.10    -0.1023365
[11,] month[T.11]     0.9284757
[12,] month[T.12]    0.4220810
```

As can be seen in Figure 4.3 below, the beta values for GLM and GEE when the formula is rain~month are not different.



<u>Figure 4.3</u>

Beta value for GLM and GEE fits for rain ~ month.
The GLM and GEE values are very close.

GEE fit values are plotted in Figure 4.4 below which shows how the model was able to predict historical values. The model was then improved by adding more factors and covariates. At this stage, the model equation had twelve terms only, the coefficient term and a term for each beta value.

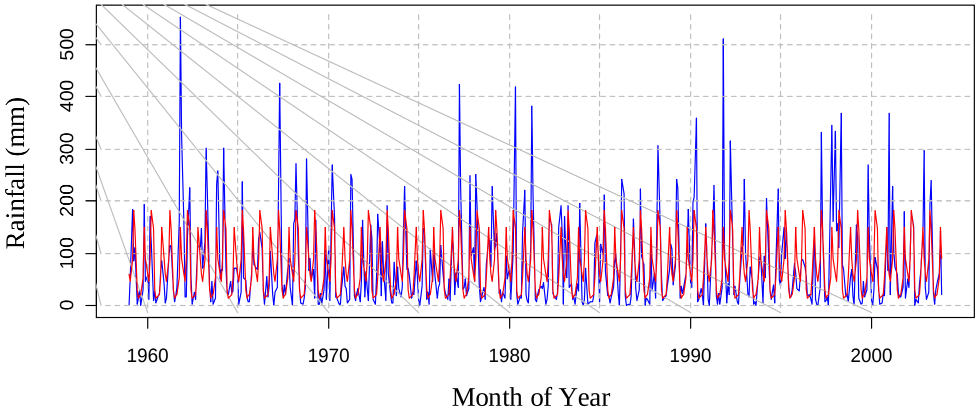


Figure 4.4
GEE fitted values for rain ~ month.

4.1.5.Step 5 Diagnostics

This fit process obtained an R^2 of 40.53606% and a QIC_u value of 18983.34

4.1.6 Residual plot for model

Figure 4.5 shows the raw residuals plotted against each observation

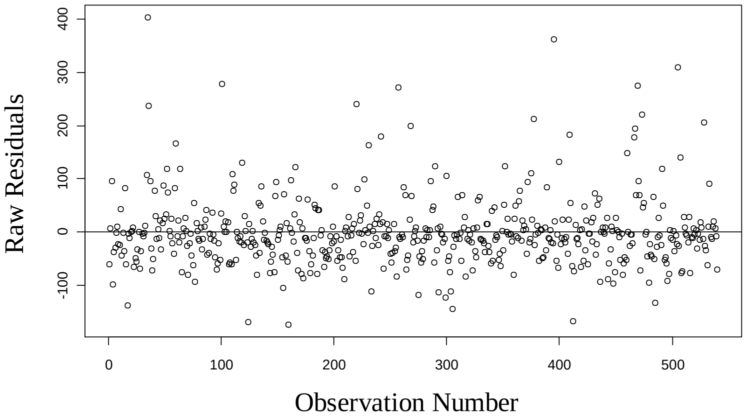


Figure 4.5
Raw Residuals for GEE fitted values for rain ~ month.

A ideal model has all the values on the y=0 line because there is no difference between the measured and the estimated values. A good model have residuals distributed close to y=0.

Another type of residuals calculated was the Pearson's residuals shown in figure 4.6.

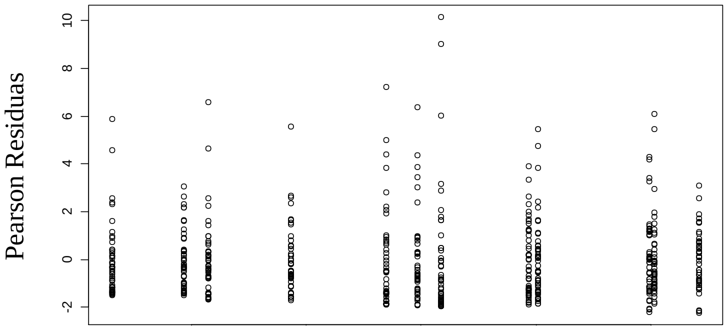


Figure 4.6

Pearson residuals for GEE fitted values for rain ~ month.

The Pearson's residuals show that linear predictors are well distributed showing a good fit for the predictor - month. A perfect fit has the Pearson residuals randomly distributed within the linear predictors. Residuals in this case are restricted to specific values of predictors showing that the fit was not good enough. The distribution was improved further by considering more than one predictors.

4.1.7 QQ Plot

A Quantile-Quantile (QQ) plot is a scatter plot comparing the fitted and empirical distributions in terms of the dimensional values of the variable (i.e., empirical quantiles). It is a graphical technique for determining if a data set come from a known population. In this plot on the y-axis are the empirical quantiles while on the x-axis are the ones obtained by the theoretical model. The R command *qqplot()* was used for drawing the QQplot and the command *abline(0,1)* for drawing a 45-degree reference line as in Figure 4.7.

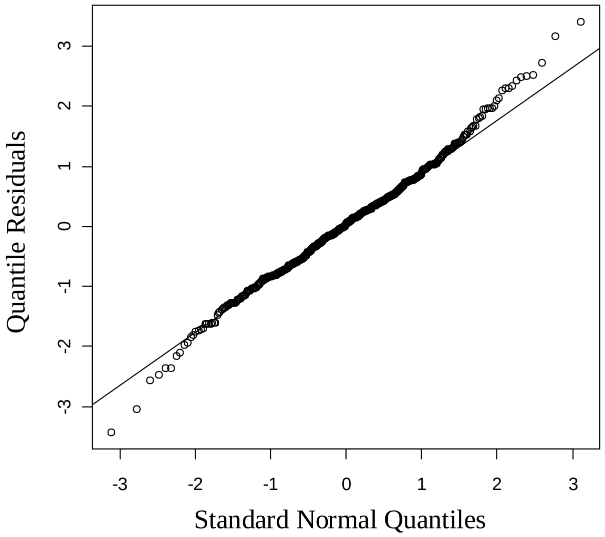


Figure 4.7
QQ plot for GEE fitted values for rain ~ month.

From the QQ plot it is evident that the tail and the head values are not captured well by a model with month as the only factor. Other factors and covariates were then added to the model to improve the performance. A model with multiple factors is discussed below.

4.2 MODEL WITH MULTIPLE FACTORS AND COVARIATES

The steps used in section 3.2.1 were then followed each time with an increasing number of predictors. In order to decide which predictors bore the heaviest cause of variability, their individual correlation with monthly rainfall was calculated. The predictors were therefore arranged in order of relevance as shown in Table 4.1

Table 4.1

Correlation between factors aon monthly rainfall

	Predictor	*Factor or Covariate*	*correlation with Monthly rainfall*
1	atide	Covariate	0.403
2	atideph	Factor	0.376
3	sdec	Covariate	-0.104
4	qboph	Factor	-0.093
	qbo	Covariate	0.083
6	soi	Covariate	-0.067
7	mld	Factor	-0.041
8	ssph	Factor	-0.028
9	month	Factor	-0.023
10	ldec	Covariate	-0.014
11	ssn	Covariate	-0.013
12	year	Factor	0.009
13	soiph	Factor	0.008

Modeling was done with each of the factors and covariates individually to deter mine their suitability. The results are shown in Table 4.2.

Table 4.2

Suitability of factors and covariates as determined by QIC_u and R^2.

Predictor	*QICu*	*R^2*	*selected(√)*
sdec (factor)	19434	41.40	√
month (factor)	18983	40.53	√
atide (factor)	16657	19.02	√
atideph (factor)	16544	14.65	√
atide (covariate)	16501	11. 82	
atideph (covariate)	16357	11.58	
soi (factor)	16562	10.80	√
ldec (factor)	16406	8.06	√
soiph (factor)	16237	2.70	√
qboph (factor)	16244	1.83	√.
sdec (covariate)	16120	1.04	
qboph (covariate)	16124	0.86	
qbo (covariate)	16126	0.63	√
soi (covariate)	16132	0.47	
mld (covariate)	16136	0.16	√
mld (factor)	16136	0.16	
ssph (factor)	16141	0.13	√
ssph (covariate)	16138	0.07	
month (covariate)	16138	0.05	

Factors qbo and ssn did not converge when considered as factors. Factors and covariates were therefore selected as ticked(√) in Table 4.2. We note that the most important factors are sdec, atide, atideph and soi emphasizing the influence of atmospheric tidal state as defined in this study. With these important factors and covariates the rainfall formula was:

rain~sdec(factor)+month(factor)+atide(factor)+atideph(factor)+soi(factor)+ldec(factor)+s oiph(factor)+qboph(factor)+qbo(covariate)+mld(covariate)+ssph(factor).

The criteria of determining the best performing model was discussed in section 2.8.13.8. Values of *QICu* and R2 were used in this model to gauge model performance. After many trials involving different combinations of factors and covariates, the best fit using all factors was of the form:

rain~month(factor)+sdec(factor)+soi(factor)+soiph(factor)+qbo(covariate)+ ldec(factor)+qboph(factor)+atide(factor)+ssph(factor)+mld(covariate)

The rainfall equation has 173 terms, a p-value of 1.5449, a QIC_u value of 28392, and a marginal R squared value of 63.7%, the highest achieved value. Beta values for this fit are shown in Appendix I. The other trials performed are shown in Table 4.3.

<u>Table 4.3</u>

Modeling using all factors and covariates

a) Modeling with Meteorological and Astronomical Factors	*p - value*	*QICu*	*R^2*
sdec(F)+atide(F)+soi(F)+ldec(F)+soiph(F)+qboph(F)+qbo(C)+mld(C)+ssph(F)	1.54	28391	63.7
sdec(F)+month(NA)+atide(F)+atideph(F)+soi(C)+ldec(F)+soiph(C)+qboph(C)+qbo(C)+stdst(C)l+ssph(C)	1.62	21680	48.9
month(F)+atide(F)+atideph(F)+soi(C)+ldec(F)+soiph(C)+qboph(C)+qbo(C)+mld(C)+ssph(C)	1.63	21025	47.8
sdec(F)+atide(F)+atideph(F)+soi(C)+ldec(F)+soiph(C)+qboph(C)+qbo(C)+mld(C)+ssph(C)	1.62	21680	48.9
sdec(F)+month(F)+atide(F)+atideph(F)+ldec(F)+mld(F)+ssph(F)	1.63	21021	47.5
sdec(F)+atide(F)+atideph(NA)+ldec(F)+mld(F)+ssph(F)	1.63	21021	47.5
sdec(F)+atide(F)+ldec(F)+mld(C)+ssph(C)	1.63	21021	47.5
month(F)+atide(F)+ldec(F)+mld(C)+ssph(C)	1.63	21043	46.9
b) Modeling with Astronomical Factors only			
sdec(F)+atide(F)+ldec(F)+mld(F)+lunaph(F)+etide(F)	1.618367	22077	52.3
sdec(F)+atide(F)+ldec(F)+mld(F)+luniph(F)+etide(F)	1.634694	21037	48.0
month(C)+sdec(F)+atide(C)+atideph(F)+ldec(F)+mld(C)	1.644898	20417	45.5
month(C)+sdec(F)+atide(C)+ldec(F)+mld(C)	1.644898	20414	45.4
sdec+atide+ldec+mld	1.644898	20414	45.4
atide+ldec+mld	1.736735	16944.	15.1

Figure 4.8 shows the log-likelihood plot for *p-value* using both meteorological and astronomical predictors.

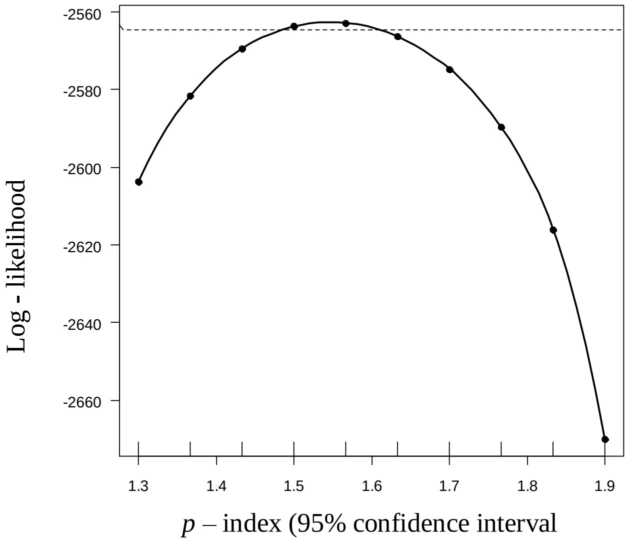


Figure 4.8
Log-likelihood plot for the p-value using multiple predictors.
The p-value at maximum likelihood is 1.544898 at 95% confidence.

A plot of predicted values overlaid on the measurements is shown in Figure 4.9

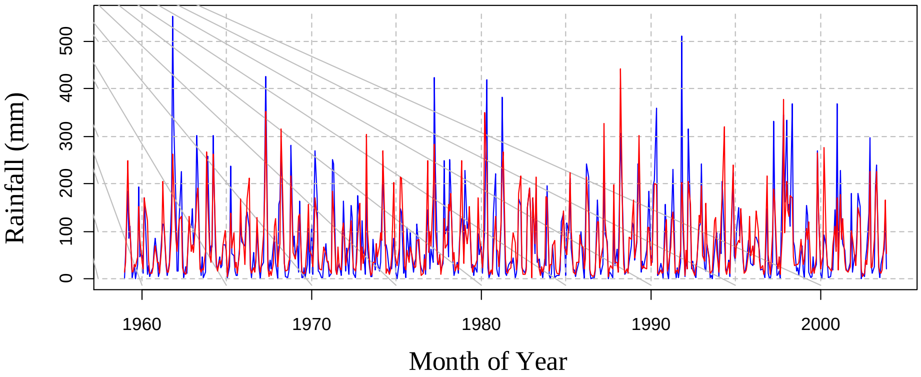


Figure 4.9
Predicted Nairobi monthly rainfall (red).
Overlaid on the measured monthly amounts (blue) obtained using multiple predictors.

The multiple predictor fit was able to capture most of the peak values not captured in the single predictor fit. The improved model was able to capture the peak values of

November 1961, May 1967, April 1977, May 1980, 1989, 1997. It however placed unknown peaks in 1968, 1973, 1974 1976 and 2000.

Raw residuals were plotted against the observation number representing each fit. The plot is shown in Figure 4.10. The plot indicates a good fit.

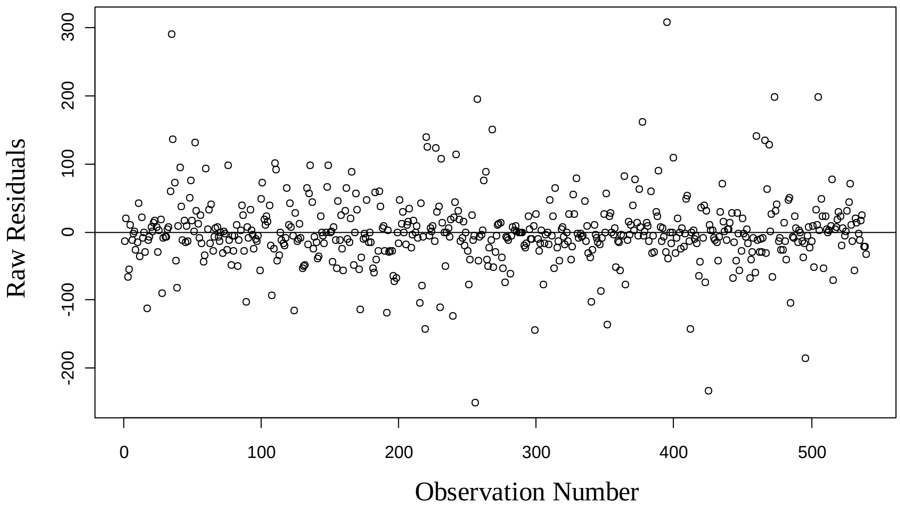


Figure 4.10
Raw residuals obtained for multiple predictors.
The residue distribution close to the zero line shows a better fit than in Figure 4.10.

Figure 4.11 shows Pearson's residuals obtained for the fit using all the factors. The distribution of the residuals show an improved fit.

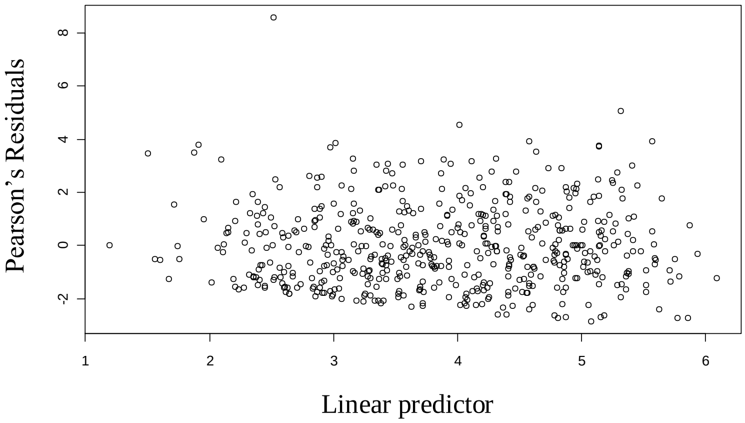


Figure 4.11
Pearson residuals for multiple predictors.
The distribution of the residuals is more random than in Figure 4.6 showing a better fit.

When a QQ-plot is done for the fit, it was observed that the inter-quartile values are well represented. However, the tail and the head values failed to get well represented in the fit seen in Figure 4.12

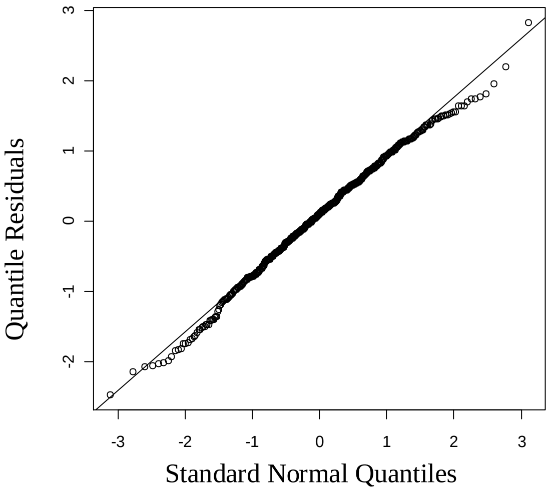


Figure 4.12
QQ plot for multiple predictors.
The inter-quartile distribution is close to the line y=x.

Although this fit provided high value of the marginal R squared, the fit values may not be used to estimate future rainfall amounts. The reason is that is that the future values of following factors are unknown: soi, soi phases, qbo and qbo phases.

4.3 MODEL DESIGN 12.3

4.3.1 Model 12.3 Fitting

Modeling of future rainfall was done with the following factors whose future values could be determined beforehand; month, sdec, ldec, atide lunaph, luniph and mld. The results obtained using different predictor combinations are shown in Table 4.4

Table 4.4
Modeling using combinations of factors(F) and covariates(C).

Formula	p-value	QICu	R2
sdec(F)+atide(F)+ldec(F)+mld(F)+lunaph(F)+etide(F)	1.618367	22077	52.3
sdec(F)+atide(F)+ldec(F)+mld(F)+luniph(F)+etide(F)	1.634694	21037	48.0
month(C)+sdec(F)+atide(C)+atideph(F)+ldec(F)+ml d(C)	1.644898	20417	45.5
month(C)+sdec(F)+atide(C)+ldec(F)+mld(C)	1.644898	20414	45.4

sdec(F)+atide(F)+ldec(F)+mld(C)	1.644898	20414	45.4
atide(F)+ldec(F)+mld(C)	1.736735	16944.	15.1

As can be seen in Table 4.4 the best results were obtained using the formula:

rain~sdec(F)+atide(F)+ldec(F)+mld(F)+lunaph(F)+etide(F).

for which the marginal R squared value of 52.3%. Beta values for this fit are available in Appendix II. Log-likelihood plot for *p* using multiple predictors is shown in Figure 4.13.

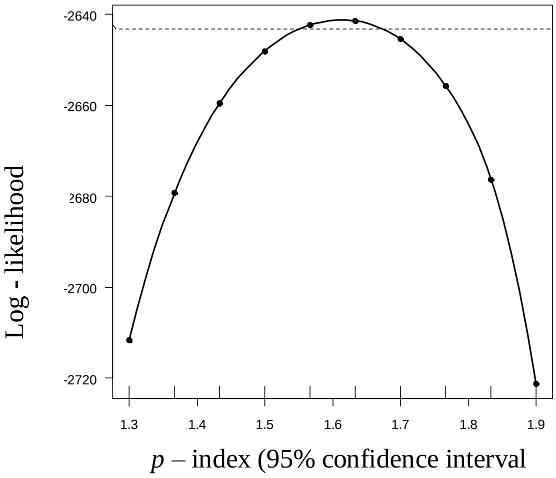


Figure 4.13
Log-likelihood plot for the p-value in Model 12.3.
The p-value at maximum likelihood is 1.6183673 at 95% confidence.

A plot of predicted values overlaid on the measurements is shown in Figure 4.14.

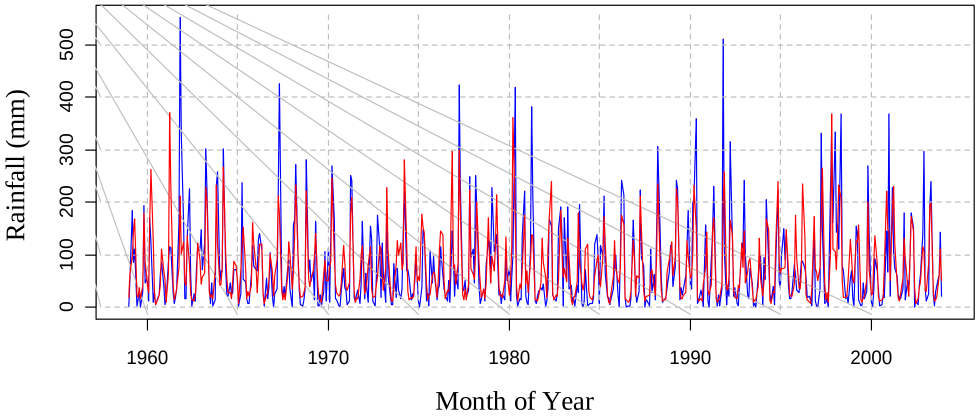


Figure 4.14
Predicted Nairobi monthly rainfall (red).
Overlaid on the measured monthly amounts (blue) obtained for Model 12.3

4.3.2 Model 12.3 Diagnostics

Model 12.3 captured most of the peak values. The model was able to capture the peak values of November 1961, May 1967, April 1977, May 1980, 1989, 1997. It however placed lower value peaks in 1961, 1976, 1994.

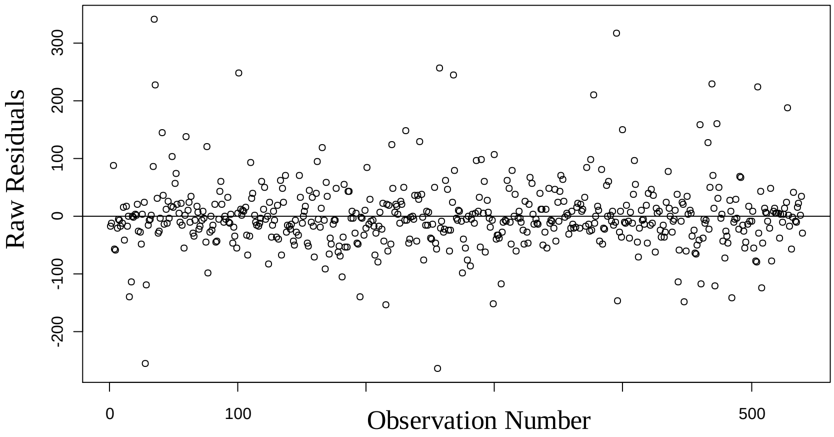


Figure 4.15
Raw residuals obtained in Model 12.3.

A look at the raw residuals in Figure 4.15 indicates that the distribution is evenly close to the zero line showing a good fit. From Figure 4.16 we find that the Pearson's residuals are even more randomly distributed than any other fit indicating that model 12.3 represents the best fit so far.

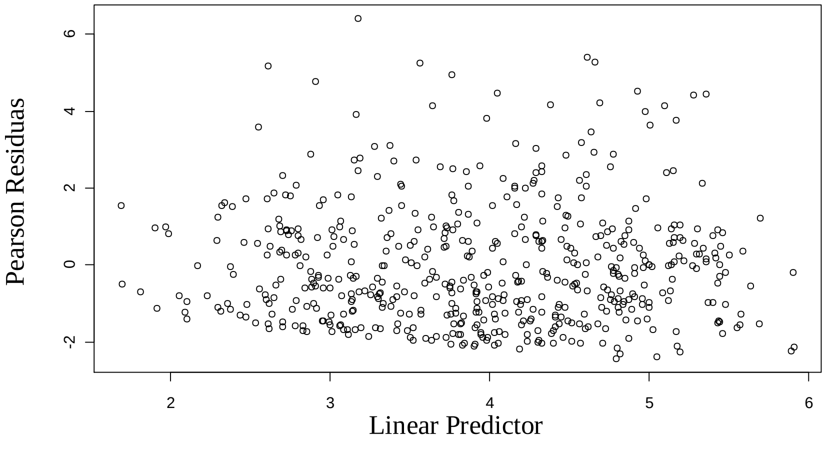


Figure 4.16
Pearson residuals for Model 12.3.

The distribution of the residuals is fairly random indicating a good fit while the QQ-plot in Figure 4.17shows the best fitting quartiles.

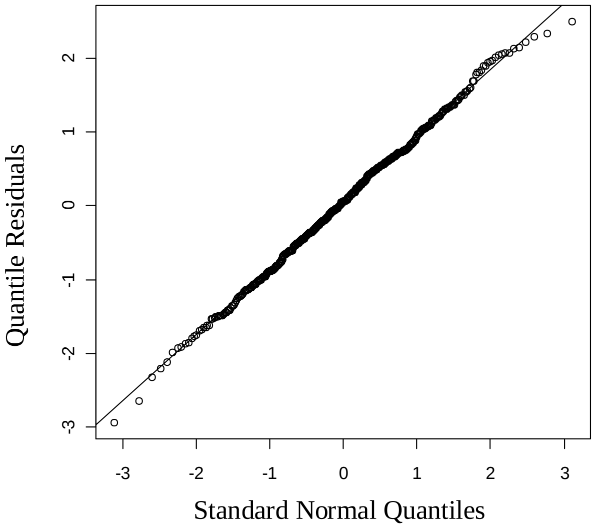


Figure 4.17
QQ plot for Model 12.3

Model 12.3 is therefore the best so far and is the one used to estimate rainfall amounts up to the year 2050.

4.3.3 Model 12.3 estimates

Monthly rainfall estimates were calculated using the formula:

rain~sdec(F)+atide(F)+ldec(F)+mld(F)+lunaph(F)+etide(F).

The equation has exactly 100 terms as shown by the beta values in Appendix II. An Excel Worksheet was used to perform the calculations so that for each month the estimate is obtained by working out the 100 term equation.

4.3.4 Model 12.3 Skill.

A plot of rainfall amounts from 1901 to 2050 is done in Figure 4.18.

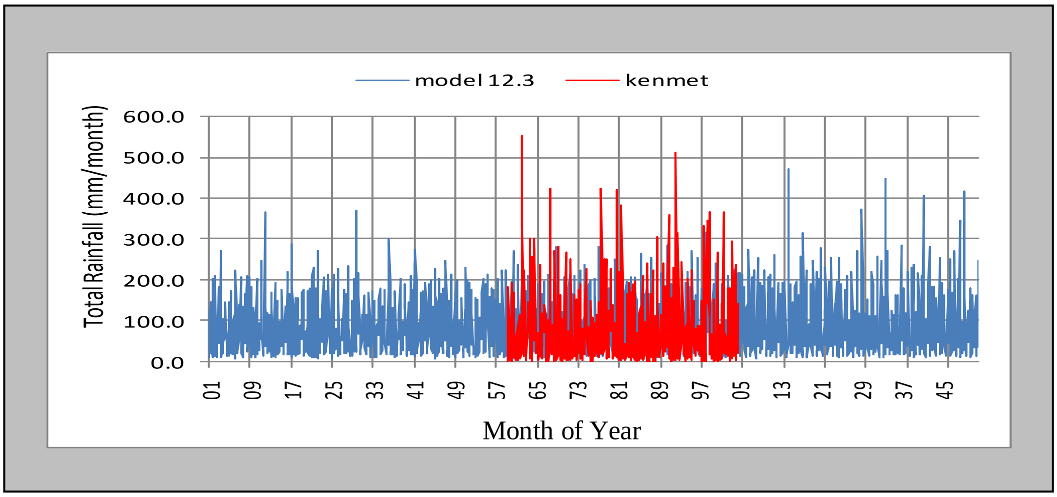


Figure 4.18
Model 12.3 estimates 1901 to 2050(blue).
Plotted together with KenMet values for the period 1959 to 2003(red).

In the figure, monthly estimates from 1901 to 2050 are plotted together with KenMet measurements.

The model overestimated the April and May of 1961 but underestimates the November from 553mm to 213mm. However, the model correctly places these values as peak values in all cases. Other peaks are also captured by the model showing that the model will correctly predict flood episodes.

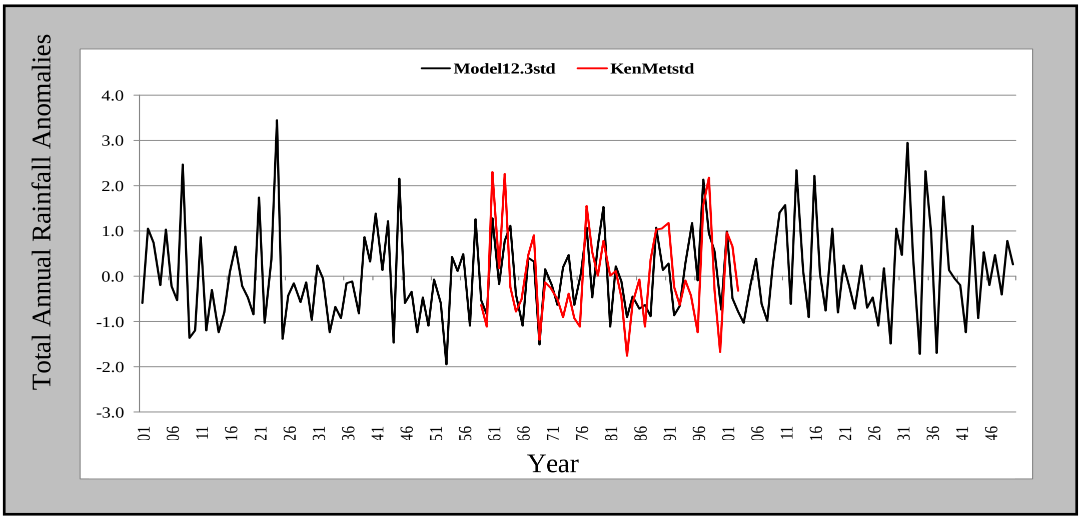


Figure 4.19
Model 12.3 annual anomaly estimates - 1901 to 2050 (black).
Plotted together with KenMet annual anomalies for the period 1959 to 2003.

One observation is that more severe floods are expected in November 2013 (698mm/month) and November 2032 (547mm/month).

When annual values were calculated the variation is shown in Figure 4.19. From the figure, it was observed that the model correctly placed the droughts and the floods in the right years. However, the amplitudes of the rainfall annual amounts were lower in the model. In order to correctly compare the variability, the two series have been standardized before plotting. The actual amounts compared are shown in Figure 4.20.

Correlation coefficient between Model 12.3 and KenMet is 0.7 which shows that the model performed well.

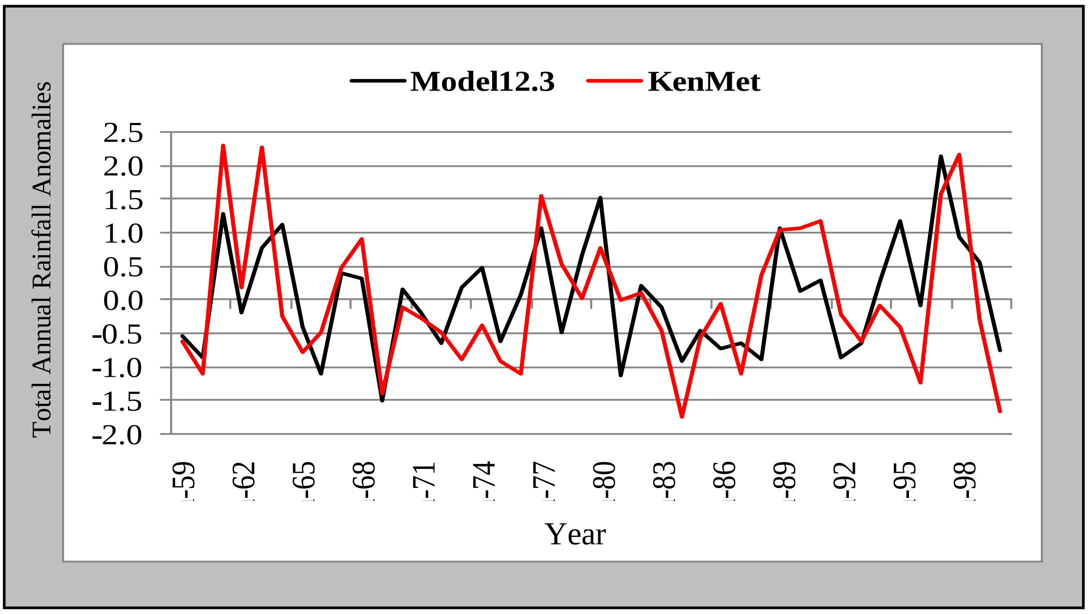


Figure 4.20
Model 12.3 annual anomalies - 1959 to 2003(black).
Plotted together with KenMet annual anomalies for the same period.

Model annual estimates were the compared with both KenMet and CRU values. The values are shown in Figure 4.21.

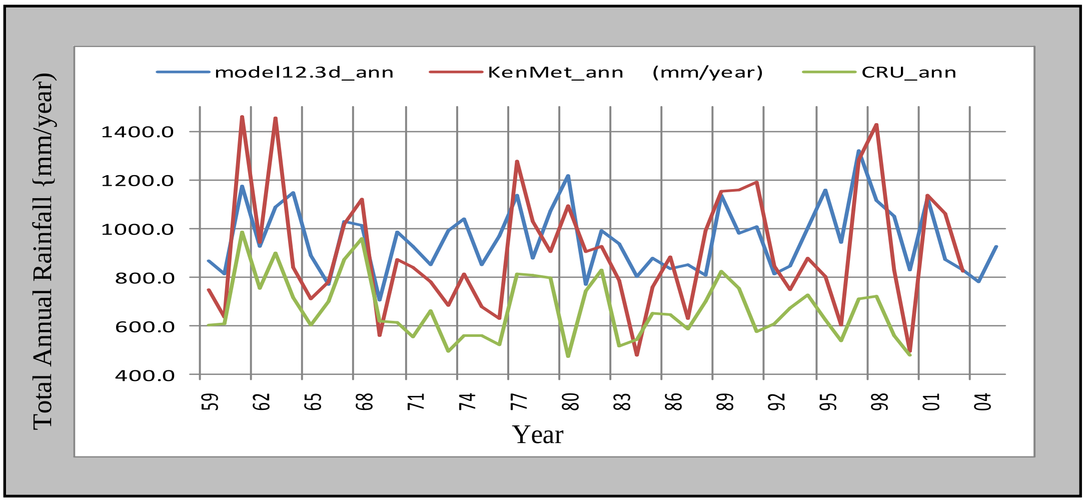


Figure 4.21
Model 12.3, KenMet and CRU.
Annual total rainfall for the period 1959 to 2000.

When compared with CRU values, both KenMet and Model 12.3 values are higher than CRU values. The correlation coefficient between Model 12.3 and KenMet is 0.7 while between the model and CRU it is 0.3.The timing of the peak values however coincide suggesting similar variability pattern. The three series were then standardized for easier visual comparison. The trend becomes more distinct as shown in Figure 4.22.

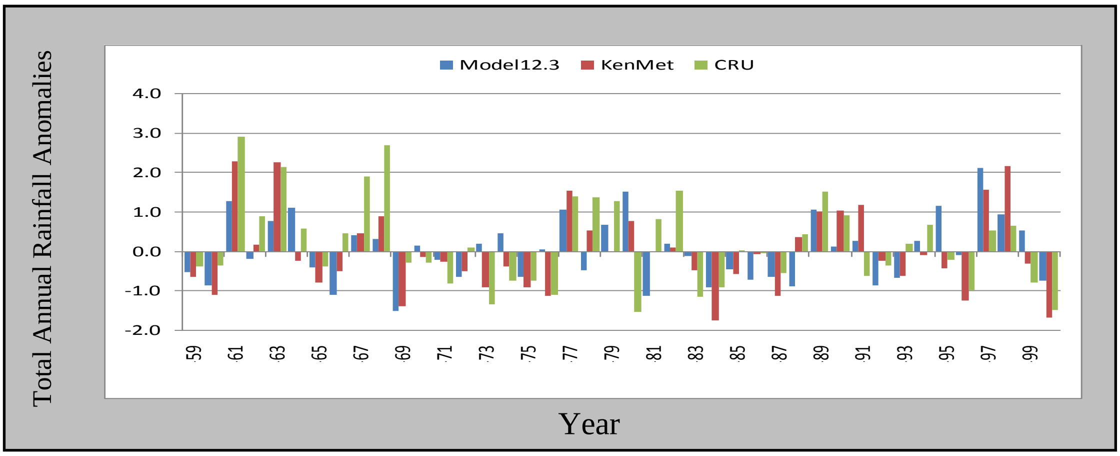


Figure 4.22
Model 12.3, KenMet and CRU Standardized values (Anomalies).
For the period 1959 to 2003.

4.3.6 Model 12.3 Reliability

Model 12.3 reliability in predicting droughts and floods was tested against known floods and droughts. Figure 4.23 shows more clearly the years in which droughts and floods are expected during the period 1901-2050. Droughts have negative anomalies (red) while floods have positive ones (black). Recorded droughts occurred in the following years

1928, 1933-34, 1937, 1939, 1942-44, 1947, 1951, 1952, 1955, 1957, 1975, 1977, 1980, 1983-85, 1991-92, 1995-96, 1999-2000, 2004 [UNDP, 2004] as indicated in Figure 4.24. Others are 1960, 1966, 1970, 1974-76, 1988, 1996 [KenMet].

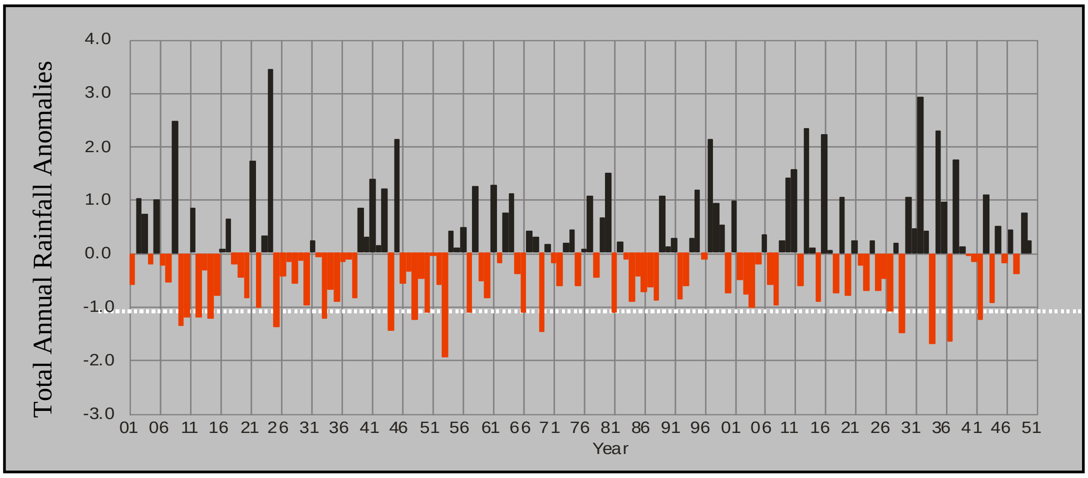


Figure 4.23
Predicted Droughts and Floods.

Recorded floods are not as numerous as the droughts years. They are: 1961, 1963, 1978, 1990-92, 1997-98, 2002 [KenMet].

Floods in Kenya are not as devastating as droughts and therefore more emphasis has been given to droughts. The severity of the specific drought or flood event depends on the geographical location in the country. The characteristics of the specific events are found in section 1.2. All the droughts and floods are placed in the right years by the model. The model indicates that future droughts will occur in the following years:

2015, 2027, 2029, 2034, 2037, 2042, 2044.

The most severe droughts will be the 2030, 2034, 2037 and 2042. Models estimates reveal that floods will occur in the following years: 2002, 2010, 2011, 2013, 2016, 2019, 2030, 2032, 2035, 2036, 2038, 2043. With the most severe being the 2013, 2016, 2032, 2035, 2038.

The floods expected in 2013 estimated at 698 mm/month will be severest since 1901. It will be heavier than the well known floods :November 1961-267mm/month, November 1963-233mm/month, December 1963-257mm/month, November 1977-345mm/month, January 1998-334mm/month. Although Nairobi monthly is not expected to vary together with the Sahel distribution exactly, we have plotted the two distributions with the Sahel precipitation obtained from several General Circulation Models GCMs. The results compare well with models results represented in Figures 4.29 and 4.30.

4.3.5 Comparison with other models

Model results were compared with Sahel rainfall distribution from CRU, GPCP and TRMM. The plot is shown in Figure 4.24.

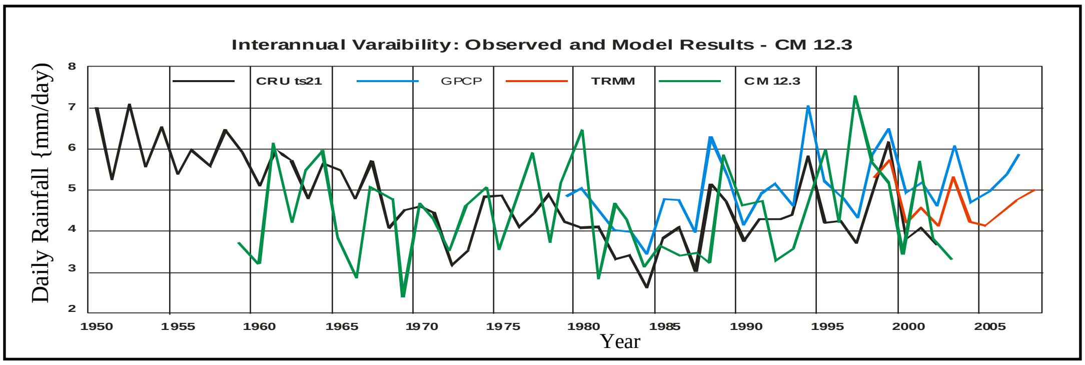


Figure 4.24

Model Performance against Sahel inter-annual variability.
Model 12.3 annual estimates for the period 1960 to 2003 compared with Sahel inter-annual variability estimates using CRU_ts21, GPCP and TRMM. [Cook, 2011].

Standardized model results from 1960 to 2050 were compared with General Climate models results for GFDL, GISSEH, MI-100, MP-ECHAMS, MRI, NCAR CV1, NCAR NCSV and UKMO and HADCM3 on Sahel rainfall plotted in Figure 4.25 below.

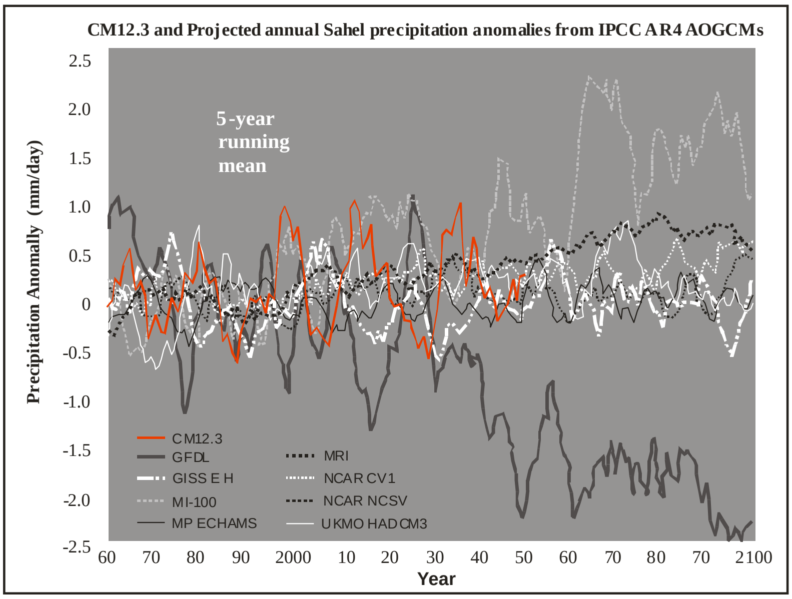


Figure 4.25

Model Performance against other models.
Model 12.3 annual estimates for the period 1960 to 2050 compared with projected Sahel annual precipitation from IPCCC AR4 AOGCMs [Cook, 2011].

IPCC models provided very divergent estimates after 2030. However, Model CM12.3 estimates agree with the values generated by Global Circulation models before 2030 and performed well in capturing the Sahel drought of the 1980s and the one after year 2000.

4.4 MODEL DESIGN SMS12.12

4.4.1 Model SMS12.12 fitting

Model SMS12.12 was trained with 30-year data from 1951 to 1980 and tested with two segments of data; 1901-1950 and 1981-2000. The predictors are solar declination (sdec), maximum lunar declinations (mld) and sunspot numbers (ssn). Figure 4.26 shows how the model demonstrates prediction stability with time.

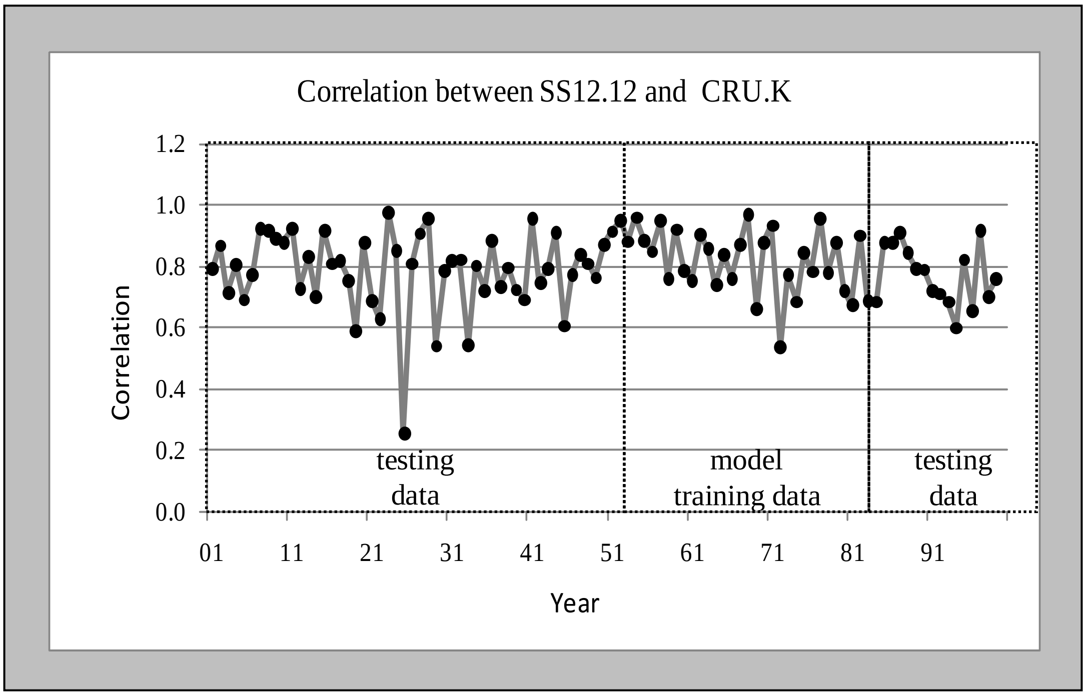


Figure 4.26
Correlation between SMS12.12 and CRUKenya monthly, for each year showing SMS12.12 stability.

Methods used for avoiding artificial prediction skill included using independent training and test data sets, cross-validation and hindcasting. Forecast skill depends on the amount of lead time, the forecast months and the strength of relationships between the predictors and rainfall. Each value represents the Pearson product moment correlation coefficient between the predicted and CRU-Kenya Country dataset for the corresponding months of the year. Correlation values remained above 0.5 throughout the period except for the 1925 value seen Figure 4.27. The chance of obtaining a value less than 0.5 in the prediction area is therefore 1month out of 1200 months (0.000833). An adjusted R^2 value of 0.62 was obtained between CRU and model estimates during the training period and reduced values of 0.52 and 0.56 in the testing datasets. SMS12.12 shows stability in estimating monthly amounts when model monthly amounts are compared with CRUKenya monthly values. The average correlation for the period 1901-2000 is 0.8.

4.4.2 SMS12.12 Rainfall projection

SMS12.12 was the used to estimate monthly rainfall totals for the period 1901-2000 after which the model was used to project monthly rainfall for the period 2001 to 2020. Monthly estimates are shown in Figure 4.27.

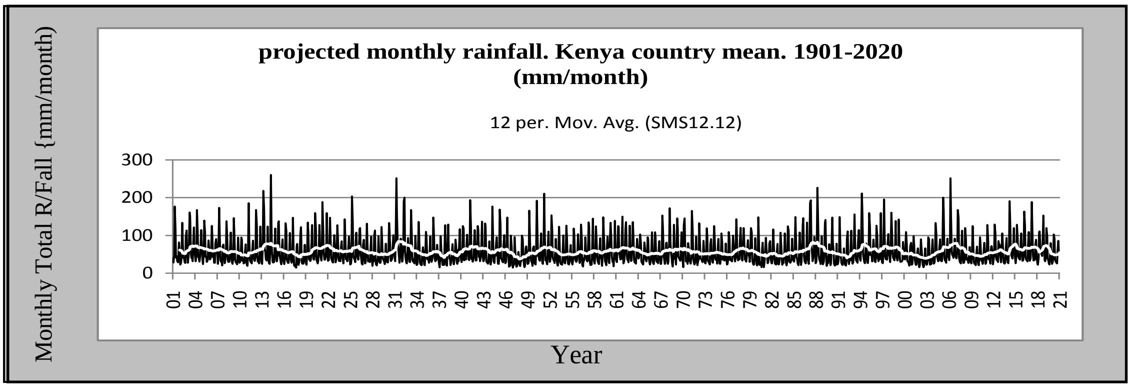


Figure 4.27
Projected monthly total rainfall by SMS12.12 for the period 1901-2020.

Model SMS12.12 estimates indicate elevated monthly totals in the periods 1912-13, 1931-32 1951-52, 1987-88, 1993-94, 1997-98, 2005-06 and depressed monthly rainfall in 1917, 1939-39, 1947-48, 1882-85,1992-93,2002-2004, 2010-11, 2019-2020. The monthly totals were then aggregated into annual values and the results standardized by the mean and standard deviation. The results are shown in Figure

4.28 in which model results have been plotted together with CRU Kenya and KenMet values for inter comparison.

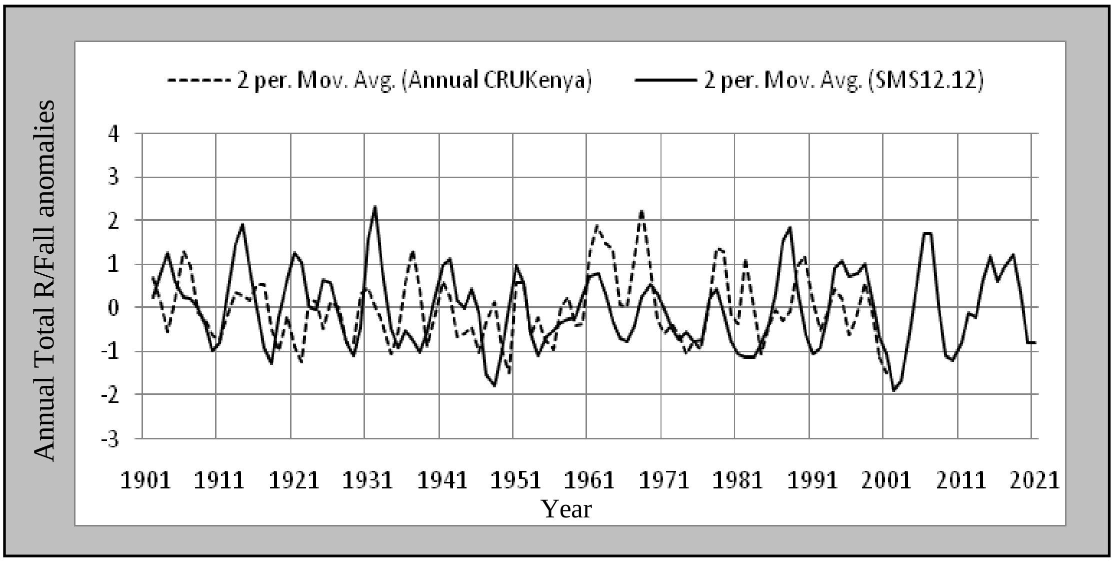


Figure 4.28
Projected annual total rainfall anomalies by SMS12.12
for the period 1901-2020. Values are plotted together with CRU.K values for comparison.

Below normal annual rainfall occurred in the following years:1928, 1933-34, 1937, 1939, 1942-44, 1947, 1952-3, 1955-57, 1975-77, 1980-85, 1991-92, 1999-2000, 2004 [UNDP, 2004]. Others are 1965, 1973-74, 1976, 1992-93, [KenMet]. Recorded floods occurred in the following years: 1961, 1963, 1977-78, 1997-98 [KenMet]. Projected model estimates indicate below normal rainfall in 2009-2011, 2015 and 2019-2020, while above normal rainfall may be expected in 2012-14, 2016,2018.

4.4.3 Model SMS12.12 Diagnostics

Probabilities of rainfall amounts were calculated in order to judge the accuracy of the model estimates. The results are shown in Figure 4.29. Estimates are comparable at all stages of model development as shown by hindcasting, training (fitting) and forecast stages as well as with the 1901-2000 climatology. Correlation values between the model and CRUKenya values are shown in Table 4.5 for the hindcast, training and forecast stages. Model estimates are therefore reliable.

Table 4.5.

Correlation coefficients within the segments

(CRU.K vs SMS12.12)

1901-1951 (hindcast)	0.90
1951-1981 (training)	0.92
1981-2000 (forecast)	0.84

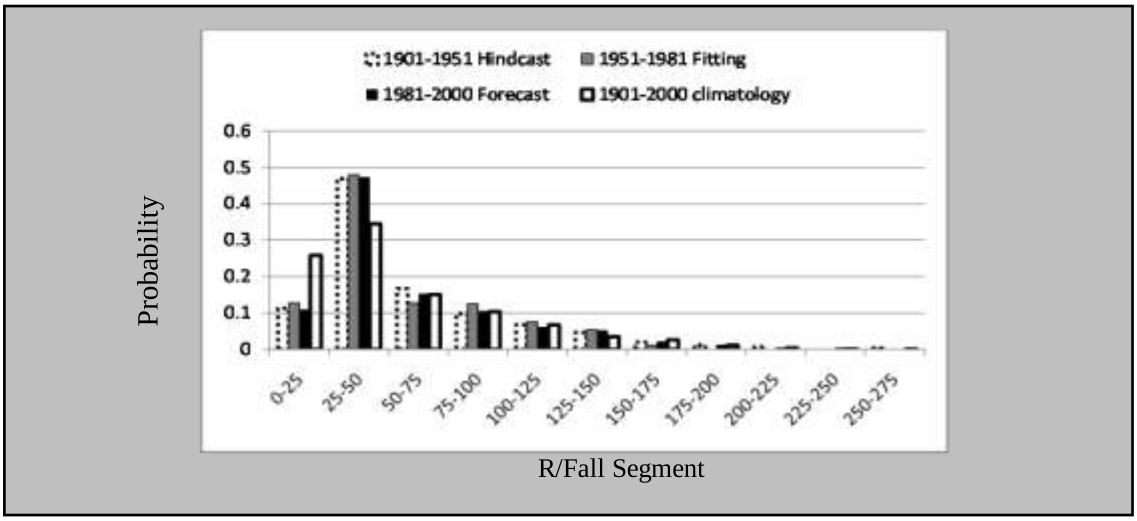


Figure 4.29
Probability of rainfall segments
for Hindcast, fitting, forecast and climatology.

4.5 MODEL CM13.1 DESIGN

5.5.1 CM13.1 fitting

Model CM13.1 is an ensemble of seven selected individual generalized linear models. Monthly total rainfall is common to all the models and they all have different combinations of explanatory variables defined in this study as described below.

4.5.2 CM13.1 Results

The first three statistical models constituting Ensemble1 are trained and tested on Nairobi monthly total rainfall for the period 1959-2003. The other four models constituting Ensemble2 are trained and tested on CRU Kenya monthly rainfall for the period 1901-2000. Each model has its own unique set of predictors as shown in Table 4.6.

Table 4.6
Characteristics of the ensemble models

Model	Factors	Training period	Annual mean(mm/yr)	R
(a) Ensemble1				
1.SLS	sdec, ldec, synod	1970-1990	1050	0.67
2.Model 12.3C	sdec, atide, ldec, mld, etide, lunaph	1959-2000	924	0.66
3. SMP	sdec, mld, prg	1970-1990	1082	0.60
(b) Ensemble2				
1. SYNODIC	sdec, atide, ldec, mld, etide, synod	1940-1970	620.0	0.78
2. SMP	sdec, mld, prg	1940-1970	654.4	0.73
3. SM	sdec, mld	1970-1990	645.4	0.73
4. SP	sdec, prg	1970-1990	643.9	0.78

Once the beta values have been obtained by fitting, the estimates were then calculated. Monthly Estimates for Nairobi rainfall were obtained by the mean of the estimates obtained by each of the three models of Ensemble1; SLS, Model 12.3C and SMP. A similar procedure was used to obtain the monthly Kenya rainfall estimates using the four model of Ensemble2; SYNODIC, SMP, SM and SP shown in Table 4.6(a). The performance of the model ensembles was assessed by working out one year moving window correlation whereby the correlation (R) between the twelve values of the year is calculated between the model estimates and the measurements. Figure 4.30 shows a plot of the time variation of R for individual models as well as the ensemble mean for Nairobi monthly total rainfall.

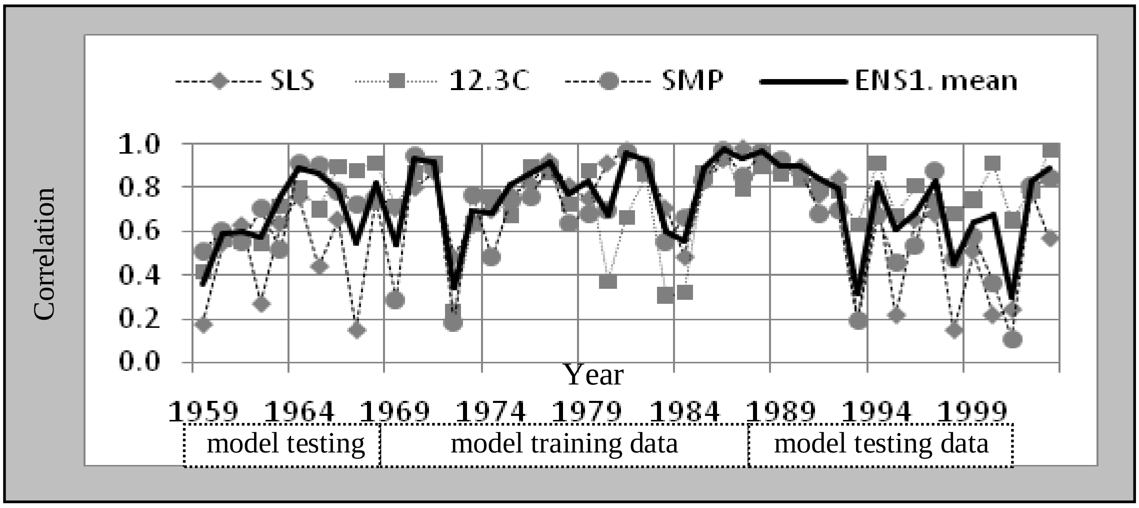


Figure 4.30
Moving window correlation between model estimates and Nairobi rainfall for the period 1959-2003 for models SLS, 12.3C, SMP as well as the enseble1 (ENS1) mean.

Except for model 12.3C which had been trained with the whole of the 1959-2003 data, the other two models; SLS and SMP had been trained on the 1970-1990 data leaving the remaining data segment for testing. A look at Figure 4.30 shows that Ensemble1 (ENS1) mean had correlations above 0.3 in both the testing and training segments of the data. From Figure 4.30 it is evident that the level of model accuracy in both the training and testing segments of the KenMet dataset is not substantially different. We found no reason to expect that correlations in the mean model will be reduced with time. ENS1 mean was therefore used to project Nairobi monthly rainfall values for the period 1901-2020.

Models fitting the CRU Kenya monthly rainfall were also tested for accuracy. As was done with the Ensemble1 models, Ensemble2 models accuracy was also tested by calculating month to month correlation for each year between model values and corresponding CRU Kenya values for the period 1901-2000. The correlation values (R), are shown in Table 4.6(b) while a plot of the one year window correlations shown in Figure 4.31.

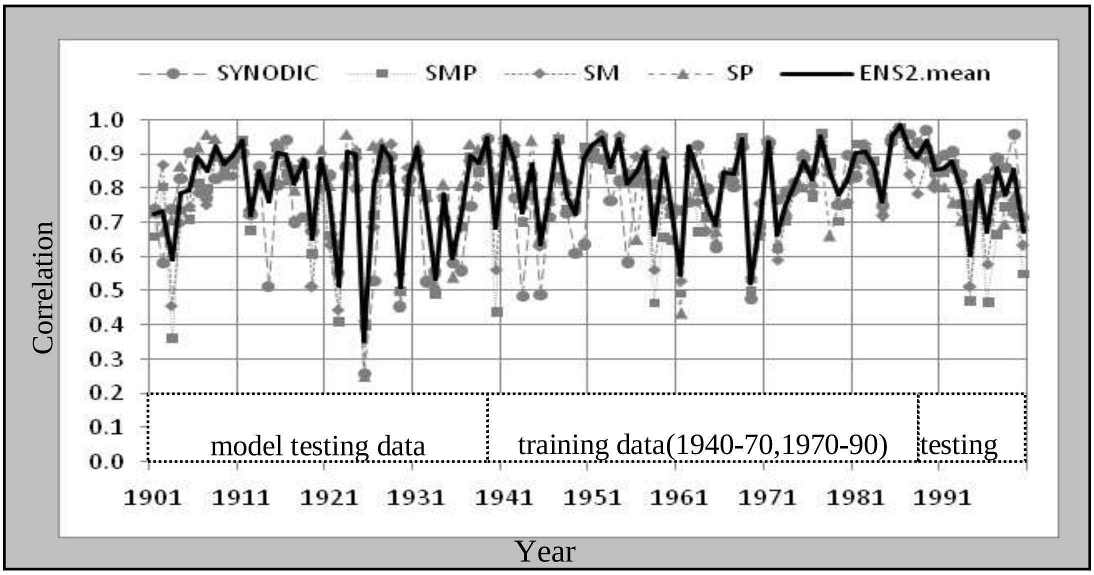


Figure 4.31

Moving window correlation between model estimates and CRUKenya rainfall for the period 1901-2000 for models SYNODIC, SMP, SM, and SP as well as the enseble2 (ENS2) mean.

We notice that Ensemble2 models fitting CRUKenya monthly totals performed better than Ensemble1 models by posting the higher values of correlations. Secondly, Ensemble1 variability is contained in Ensemble2. This means that although the ensembles are a result of fitting different models, both ensembles portray the same results for the period 1959-2003 a result confirmed by working out the correlation coefficient between corresponding monthly values in this period. The correlation between monthly values in Ensemble1 and Ensemlbe2 during the period 1959-2003 is 0.7. Therefore, Ensemble2 may be sufficient in estimating rainfall variation both in Nairobi and Kenya. Thus the two models represent the same variability as expected because Nairobi is located in Kenya. Secondly, that the two ensembles of models having different predictors can reproduce the same variability pattern increases the confidence in the model results. ENS2 mean was also used to project monthly rainfall totals for the period 1901-2020 as well.

4.5.3 Rainfall projection by CM13.1

Monthly estimates were then aggregated into annual estimates for both ENS1(Nairobi) and ENS2(Kenya). Annual values have been standardized by the long term means and standard deviations and plotted in Figure 4.32 in which both series are plotted together as anomalies so that only the annual variability is emphasized.

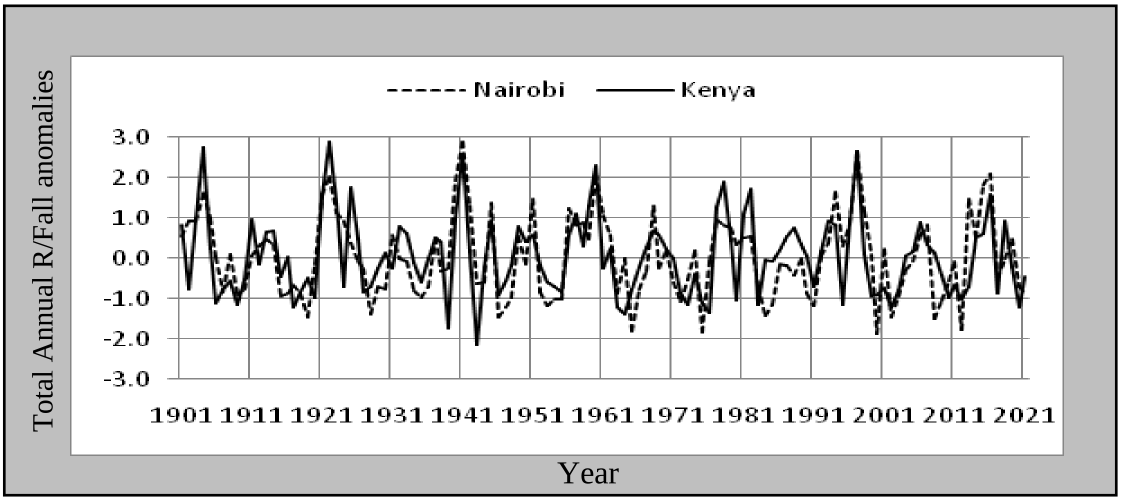


Figure 4.32
Nairobi and Kenya annual total rainfall anomalies

Model CM13.1 reliability in predicting severe hydrology events was tested against historical floods and droughts. Recorded droughts occurred in the following years:

1928, 1933-34, 1937, 1939, 1942-44, 1947, 1951, 1952, 1955, 1957, 1975, 1977, 1980, 1983-85, 1991-92, 1995-96, 1999-2000, 2004 [UNDP, 2004]. Others are 1960, 1965, 1969, 1973, 1976, 1987, 1993. Recorded floods are not as numerous as the droughts years. They are: 1961, 1963, 1978, 1997-98 [KenMet]. Floods in Kenya are less devastating than droughts and therefore more emphasis has been given to droughts. The severity of the specific hydrological event depends on the geographical location in the country. The model indicates above normal annual rainfall in 2014-16 and below normal rainfall in 2013, 2018 and 2020 as seen in Figure 4.32.

4.6 SUNSPOT NUMBERS AND ANNUAL RAINFALL.

Kenya rainfall in the modern maximum indicate a peak trend that corresponds to the that of sunspots. The trend has a peak in the 19th sunspot cycle centered around 1961. Variability of annual rainfall shows reflection symmetry in the year 1961 so that cycles 18 and 20 are object and image respectively, in Figure 4.33.

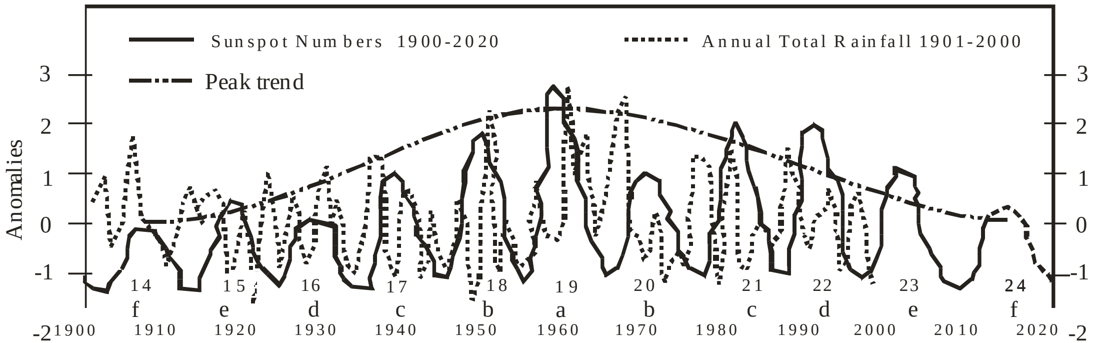


Figure 4.33
Smoothed sunspot numbers and annual total rainfall for the period 1901-2000.

We refer to events occurring prior to 1961 as objects of corresponding events after 1961 as images. Object and image pairs labeled c, d, e and f are 17 and 21, 16 and 22, 15 and 23 and 14 and 24. Object and image cycle pairs have similar rainfall peak amplitudes. Reflection symmetry demands that if sunspot turning points lead rainfall events in the objects side, the reverse will happen in the image side. While the cause(s) of the distribution symmetry is still under investigation, it is what is observed from Figure 4.33. At least two sunspot turning points are outstanding. The fist one is the heavy rainfall of the early sixties corresponding to the passing cycle the 19 maximum, the second is the drought of the mid-seventies and the passing of the minimum between cycles 20 and 21, and the third corresponds to the great Sahelian drought after the passing of the minimum between cycles 21 and 22. From Figure 4.33 one can identify a turning point for each event of severe hydrology in Kenya suggesting that sunspot numbers have a direct influence on rainfall as was found by Meehl and Julie, [Meehl and Julie, 2009]. Now that sunspots are headed for a minimum at the end of the Mordern Maximum one may expect reduced events high rainfall and perhaps a prolonged drought of the Sahelian drought of the mid-eighties. Judging from the symmetry of the Mordern Maximum a drought of the type experienced in early thirties will most likely occur in 2020 ±2 after the passage of the current cycle 24. This observation is also consistent with model SMS12.12 results as shown in Figure 4.32. Because Kenya rainfall is influenced by the Sahel climate, it is likely that the decline in rainfall amounts will be experienced in the Eastern Africa region and perhaps the Sahel region including the Greater Horn of Africa

CHAPTER 5

5.0 CONCLUSIONS AND RECOMMENDATIONS

This study was motivated by the desire to find out the physical causes of the Kenyan droughts of the early eighties and at the turn of this century. We have found that Kenya rainfall variability is seen to be largely as a result of natural causes. This is supported by the fact that the rainfall pattern could be estimated using only solar-lunar variables and their derivatives as explanatory variables in a rainfall model and obtaining a correlation coefficient as high as 0.9 between the model estimate and the measurements. It is however necessary to continue to investigate the factors which determine the unexplained variability.

Further model improvements will be possible if factors and covariates are identified which make the estimates more accurate in terms of amplitudes. The model may also be expanded to include multiple site estimates so long as reliable climate variables records are available for each site for longer lead times. Our ability to collect and store reliable data continuously will therefore always be put to test. Three statistical models: CM12.3, SMS12.12 and an ensemble of models referred CM13.1 may also be used estimate both historical and future monthly total rainfall or any other climate variable so long as appropriate factors and covariates can be identified. The statistical models successfully captured a large amount of variability in the precipitation and depicted the important relationships between the precipitation and the predictors. Correlation however does not imply cause. Furthermore, because the predictors used in the model were derived for solar-lunar geometry associated with gravity atmospheric tides, there is likelihood that the tides are key factors in rainfall variability in Kenya.

Ensemble projection indicate an increase in rainfall amounts during the period 2014-2018 with only a brief break in 2017 and the periods 2013 and 2019-2020 showing depressed rainfall amounts. Model results can be used as valuable information in planning. This information will contribute a lot in the agricultural sector especially in the flood warning system, planning and management of water resources and hydrological process. The real value of the rainfall projection will be their impact on the decision making process. It is necessary that the model results be made available to stakeholders for assessment and possible improvement.

The temporal distribution of sunspot numbers indicates that each turning point corresponds to events of severe hydrology in Kenya with time lag of 5±2 years. Therefore, such events are predictable so long as sunspots can be predicted in advance. However, the prediction of sunspots has not been easy and the current prediction of cycle 24 appears to be at the end of the Modern Maximum therefore breaking the continuity. The current maximum is fairly symmetrical increasing the confidence that sunspot activity is headed for an all time low perhaps similar to the one at the beginning of the last century with corresponding reduction in annual rainfall amounts. That is why it is possible to expect reduced rainfall amounts at the turn of cycle 24 and around the year 2020 ±2.

Model estimates indicate that before 2020 above normal rainfall may be expected in the period 2013-2018 and below normal rainfall in 2019-2020. No sunspot numbers are available to enable estimation beyond 2020 using the model and the behaviour of sunspots is uncertain after cycle 24. However, as we head towards year 2020, it is likely that that the evolution of sunspots will have taken a predictable pattern so that sunspot prediction will be possible thereby extending the range of prediction beyond 2020. However, this observation cannot yet be assumed for global data sets. Furthermore, we recommend that future studies be done on rainfall residuals so that the seasonality factor is eliminated and a better indication of the influence of sunspot numbers can be obtained. A comparison of the results with those obtained through statistical downscaling methods is also recommended.

There is a need to do further work with a view to increasing model accuracy. Parameterization of Africa land surface as a factor of variability in annual insolation absorption is recommended.

The benefits gained through research that supports agricultural production may be reduced due to the unpredictability of inter-annual variability of rainfall. A centre /unit that facilitates climate studies at this university is recommend.

Finally, the university needs to improve computation capacity to cope with the demands of running predictive models with factors and covariates of orders greater than 1.

REFERENCE

Africa Environment Outlook, 2014. Past, present and future perspectives. Available at; http://www.unep.org/dewa/Africa/publications/AEO-1/209.htm

Andrews, D.G., J.R. Holton, and C.B. Leovy, 1987: Middle Atmosphere Dynamics. Academic Press, 489pp.

Aondover Tarhule1 and Ming-Ko Woo. Towards an Interpretation of Historical Droughts in Northern Nigeria. Climatic Change, no 37, 1997. pp.601-613.

Baldwin, M.P., 2001: The Quasi-Biennial Oscillation. Rev. Geophys., 39, 179–229.

Ballinger G. Using generalized estimating equations for longitudinal data analysis. Organizational Research Methods, 7(2):127–150, April 2004.

Batterbury S. The Sahel region; assessing progress twenty-five years after the great drought. Republished paper from 1998 RGS-IBG conference. Global Environmental Change (2001) v11, no 1, 1-95.

Beersma J. and Buishand A.. Multi-site simulation of daily precipitation and temperature conditional on the atmospheric circulation. Climate Research, 25(2):121–133, 2003.

Berry F. A., Bollay E. and Beers N. R., eds., Handbook of Meteorology (New York: McGraw-Hill, 1945). pp. 483-498 and 746-749. Atmospheric tides.

Chandler R. and Wheater H.. Climate change detection using generalized linear models for rainfall — A case study from the west of Ireland II. Modelling of rainfall amounts on wet days. Research Report 195, Department of Statistical Science, University College of London, June 1998.

Chandler R. E. On the use of generalized linear models for interpreting climate variability. Research Report 232, Department of Statistical Science, University College London, February 2003.

Chang, P., L. Ji, and H. Li, 1997: A decadal climate variation in the tropical Atlantic Ocean from thermodynamic air-sea interactions. Nature, 385, 516-518.

Cologne J.B., Fujita S., Carter R. L. and Ban S.. Application of generalized estimating equations to a study of in vitro radiation sensitivity. Biometrics, 49(3):927–934, September 1993.

Convexdna. Excel Mixer Program. Available at: http://convexdna.com, 2014.

Cook K. H. Climate Variability and Prediction for Africa: Decadal time scales and Beyond. NCAR ISP Summer Colloquium 2011.Africa Weather and Climate. Jackson School of Geosciences. The University of Texas at Austin. August 2011. http://www.mmm.ucar.edu/events/ISP/presentations/Cook_Climate_Var_Prediction_4Aug_2011.pdf

Crowder. M. On the use of a working correlation matrix in using generalised linear models for repeated measures. Biometrika, 82(2):407–410, June 1995.

Dahmen G. and Ziegler A.. Generalized estimating equations in controlled clinical trials: Hypothesis testing. Biometrical Journal, 46:214– 232, 2003.

Dcsymbols. 2013: Available at: http://dcsymbols.com/tides/tides2.htm

Diggle P. Liang K. and Zeger S.. Analysis of Longitudinal Data. Oxford University Press Inc., New York, 1994.

Dobson A. J.. An introduction to generalized linear models. Chapman and Hall, London, 2nd edition, 2002.

Dobson A., Puride D. and Williams G.. Application of generalized estimating equations to longitudinal data. Longitudinal Studies, The University of Queensland, 2003.

Dunlop D. Regression for longitudinal data: A bridge from least squares regression. The American Statistican, 48:299–303, 1994.

Dunn P. and Lennox S. Simultaneous analysis of rainfall occurrence and amounts using power-variance generalized linear models. Water Resources Research, Submitted 2006.

Dunn P. K. and Smyth G. K. Randomized quantile residuals. Journal of Computational and Graphical Statistics, 5:1–10, 1996.

Dunn P., Tweedie: Tweedie exponential family models. R package version 1.02, 2004. http://www.sci.usq.edu.au/staff/dunn/twhtml/home.htm

Fitzmaurice G.. A caveat concerning independence estimating equations with multivariate binary data. Biometrics, 51:309–317, 1995.

Gachari F. Total Column Water Vapour Retrieval over Nairobi using a MAX_DOAS instrument and Validation using Satellite data. Masters Thesis, Jomo Kenyatta University of Agriculture and Technology, 2008.

Gathara 1995, Devastating Drought in Kenya: Environmental Impacts and Responses, p 18. Modified by UNEP and GOK, December 2000, Nairobi, Kenya.

Gill J. Generalized Linear Models - A Unified Approach. SAGE Publications, Inc., United States of America, 2001.

Hanley J. Edwardes M., Negassa A. and Forrester J. Statistical analysis of correlated data using generalized estimating equations: An orientation. American Journal of Epidemiology, 157(4):364–375, 2003.

Hardin J. and Hilbe J. Generalized Linear Models and Extensions. Stata Corporation, United States of America, 2001.

Hathaway D. H., Wilson M. R., Reichmann J. E., 1999: A synthesis of solar cycle prediction techniques. *Journal of Geophysical Research*, Vol. **104**, No.A10, 22375-22388.

Hedeker D. Longitudinal Data Analysis, chapter 8 : Generalized Estimating Equations, pages 193–228. University of Illinois, Chicago, United States of America, 2005.

Held, I. M., 2005. Simulation of Sahel drought in the 20th and 21st centuries. *PNAS* **102** (50): 17891–17896.

Holton, J.R. and H.C.Tan, 1980: The influence of the equatorial Quasi-Biennial Oscillation on the global atmospheric circulation at 50mb. Journal of Atmospheric Science., Vol. 37, pp. 2200-2208.

Holton, J.R. and R.S.Lindzen, 1972: An updated theory for the Quasi-Biennial cycle of the tropical stratosphere. Journal of the Atmospheric Sciences Vol. 29, pp. 1076-1080.

Horizons , 2013: Horizons Web Interface. HORIZONS Web-Interface. Available at http://ssd.jpl.nasa.gov/horizons.cgi?s_disp=1#top.

Horton N. J. and Lipstiz S. R. Review of software to fit generaized estimating equation regression models. The American Statistican, 1999.

Hyndman R. J. Nonparametric additive regression models for binary time series. Clayton, Victoria, March 1999. Department of Econometrics and Business Statistics, Monash University.

Jørgensen B. Exponential dispersion models (with discussion). Journal of the Royal Statistical Society, Series B: Statistical Methodology, 49:127–162, 1987.

KenMet. 2005: Kenya Meteorological Department Data Centre. Nairobi. Available at: http://www.meteo.go.ke/.

Kenyaweb. http://Kenyaweb.com/agriculture/ Overview of Agriculture in Kenya , 2003.

Kristof Petrovay. 2010. Solar Cycle Prediction, Living Rev. Solar Phys. 7. Cited 3rd January 2013. Available at:- http://www.livingreviews.org/lrs. 2010-6.

Lall U., Rajagopalan B., and Tarboton K.. A nonparametric wet/dry spell model for resampling daily precipitation. Water Resources Research, 32(9):2803–2823, September 1996.

Lamb, H.: 1932, Hydrodynamics, Cambridge University Press, Cambridge, England. 6th edition, 1932.

Lassen K, Friis-Christensen E. 1995. Variability of the solar cycle length during the past five centuries and the apparent association with terrestrial climate. J Atmos Terrestrial Physics. 57. 835.

Liang K. and Zeger S. L. Longitudinal data analysis for discrete and continuous outcomes. Biometrics, 42(1):121–130, March 1986.

Lindzen R. S. and Chapman S, 1969. Atmospheric Tides, Part 1. Contrasting dominant components of atmospheric tides caused by solar and lunar gravitational effects. National Center for Atmospheric Research. 133-153.

McCullagh P. and Nelder J. A.. Generalized Linear Models. Chapman and Hall, London, 2nd edition, 1989.

McCulloch C. E and Searle S. R. Generalized, Linear and Mixed Models. John Wiley and Sons Inc., Canada, 2001.

Meehl GA, Arblaster JM. 2009. A lagged warm event-like response to peaks in solar forcing in the Pacific region. J Climate. 22. 3647-3660, DOI: 10.1175/2009JCLI2619.1

Mitchell, T.D., Hulme, M. and New, M., 2002: Climate data for political areas. Area 34, p 109-112. Cited 19th January 2013. Available at http://www.cru.uea.ac.uk/cru/data/hrg/.

NASA, National Aeronautics and Space Administration. 22-09-2009. Available at: http://www.nasa.gov.

Nelder J. A. and R. Wedderburn W. M.. Generalized linear models.Journal of Royal Stastistical Society. Series A (General), 135(3):370– 384, 1972.

Neumann G., Pierson W. J. Jr *(*1966*)* Principles of Physical Oceanography *(*Prentice–Hall, Englewood Cliffs, NJ*)*.

Pan W. Akaike’s information criteria in generalized estimating equations. Biometrics, 57:120–125, 2001.

Paul Andre de la Porte. Resident Representative for the United Nations Development Programme (UNDP), Kenya Natural Disaster Profile. UNDP Enhanced Security Unit, 2004].

Republic of Kenya. National Policy for the sustainable Development of the Arid and Semi Arid Lands of Kenya,. Ministry of State for Development of Northern Kenya and Other Arid Lands, Nairobi. p 8-11, Kenya March 2004.

Shanahan , Timothy; Overpeck, JT; Anchukaitis, KJ; Beck, JW; Cole, JE; Dettman, DL; Peck, JA; Scholz, CA , 2009. Atlantic Forcing of Persistent Drought in West Africa". *Science* **324** (5925): 377–380.

SIDC, 2013. Solar Influences Data Analysis Center. Cited 13th November 2012. Available at: http://sidc.oma.be.

Solar Ephemeries - telnet://ssd.jpl.nasa.gov:6775. 2011.

Stone R. and Auliciems A. SOI Phase relationships with rainfall in eastern australia. International Journal of Climatology, 12:625–636, 1992.

Sutradhar B. C and Das K.. On the efficiency of regression estimators in generalised linear models for longitudinal data. Birometrika, 86(2):459–465, June 1999.

Tarbuck K. and Lutgens M. 1997. Earth Science. Prentice Hall, 8th Edition.

Toolkit. http://www.assessment.ucar.edu/toolkit/

Troup A. The southern oscillation. The Quarterly Journal of the Royal Meteorological Society, 91:490–506, 1965.

Tweedie M. C. K. An index which distinguishes between some important exponential families. In Statistics - Application and New Directions, pages 579–604, Indian Statistical Institute, Calcutta, 1984. Proceedings of the Indian Statistical Institute Golden Jubilee International Conference

UNDP. Kenya Natural Disaster Profile. Enhanced Security Unit.May 2004

UNEP Africa Environment Outlook. Past, present and future perspectives. United Nations Environmental Programme (2002). Retrieved 2009-02-13.

Wedderburn R. W. M.. Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. Biometrika, 61(3):439–447, December 1974.

Wendell Bechtold (Meteorologist, Forecaster), 2008, National Weather Service, Weather Forecast Office, St. Louis, MO.

White WB, Liu Z. 2008. Non-linear alignment of El Niño to the 11-yr solar cycle. Geophys Res Lett. 35: 19607.

Wikipedia. http://en.wikipedia.org/wiki/Demographics_of_Kenya#cite_not_2 , May 2012.

Wood F. J. ,1986 Tidal Dynamics Reidel, Dordrecht, The Netherlands.

Yndestad Harald, Turrel William R.. Ozhigin Vladimir, Lunar nodal tide effects on variability of sea level, temperature, and salinity in the Faroe-Shetland Channel and Barents Sea. Deep-Sea Research I 55 (2008) 1201-1217, www.elsevier.com/locate/dsri.

Yousef and Ghilly.2000. "Alert el Sahel countries; drought is approaching". *http://www.virtualacademia.com/pdf/cli209_220.pdf*

Zeger S. L. and Liang K. Longitudinal data analysis using generalized linear models. Biometrika, 73(1):13–22, April 1986.

Zhang Rong, Delworth Thomas L. 2006. "Impact of Atlantic multidecadal oscillations on India/Sahel rainfall and Atlantic hurricanes". *Geophysical Research Letters* **33** (17): L17712.

Zheng B.. Summarizing the goodness of fit on generalized linear models for longitudinal data. Statistics in Medicine, 19:1265–1275, 1988.

Ziegler A., Gromping U. Kastner C. and Blettner M.. The generalized estimation equations in the past ten year: An overview and a biomedical application. April 1996.

Zorn C. J. W.. Generalized estimating equation models for correlated data: a review with applications. American Journal of Political Science, 45(2):470–490, April 2001.

APPENDICES

APPENDIX I

Code for GEE

```
# Initial setting of data
# Load libraries that are needed to perform calculations
rm(list=ls()) # remove any previous lists
rm(list=objects()) #remove any previous objects
library(stats)
library(statmod)
library(tweedie)
# Set the directory
setwd ("D:/Documents and Settings/Administrator/My Documents/February2012/")
#rm(list=ls())
Nairobi <- read.table("D:/Documents and Settings/Administrator/My
Documents/March2012/march2003.txt",header=TRUE,sep="",na.strings="NA",dec=
".",strip.white=TRUE)
# Load the data. # Nairobi <- read.table("clipboard", header=TRUE, sep="",
na.strings="NA",   dec=".", strip.white=TRUE)
# OR file is D:/Documents and Settings/Administrator/My
Documents/February2012/allfactors_monthly
#
# Define the factors soiph and month, etc
#Nairobi$month=factor(Nairobi$month)
```

```
Nairobi$stdstl=factor(Nairobi$stdstl)

#Nairobi$ssph=factor(Nairobi$ssph)

#Nairobi$qboph=factor(Nairobi$qboph)

Nairobi$atide=factor(Nairobi$atide)

#Nairobi$atideph=factor(Nairobi$atideph)

#Nairobi$soiph=factor(Nairobi$soiph)

#Nairobi$soi=factor(Nairobi$soi)

#Nairobi$qbo=factor(Nairobi$qbo)

Nairobi$sdec=factor(Nairobi$sdec)

Nairobi$ldec=factor(Nairobi$ldec)

Nairobi$lunaph=factor(Nairobi$lunaph)

Nairobi$etide=factor(Nairobi$etide)

Nairobi$lmean=factor(Nairobi$lmean)


attach(Nairobi)


####################################################################

#FITTING THE GENERALISED ESTIMATING EQUATION

# STEP 1 - Compute an initial estimate of beta using GLM

# metholodogy. Calculate "p", to be used in the variance function

# of the Tweedie distribution using profile likelihood function.
```

```
power=tweedie.profile(rain~sdec+atide+ldec+stdstl+lunaph+etide+lmean,p.vec=seq(
1.3,1.9,length=10),do.plot=TRUE,do.smooth=TRUE,do.ci=TRUE,method="interpol
ation")
p=power$p.max

# Fitting a Tweedie model to this data, with "p" value found
# using the profile likelihood function and a log link function
glmmodel<-
glm(rain~sdec+atide+ldec+stdstl+lunaph+etide+lmean,family=tweedie(var.power=p,
link.power=0),x=TRUE)
# Initialize values - variables used in the first repetition
fits<-glmmodel$fitted.values
beta=glmmodel$coefficients
phi=power$phi.max
n=length(rain)
# Let "r" be the number of beta values (number of covariates +1)
r=glmmodel$rank
# Set the variables to be used in the convergence criteria
dev=sum(tweedie.dev(rain,fits,p))
devold=100*dev
epsilon=1e-8

################################################################
# Create the recursive (repeating steps 2 to 5), using a
# convergence criteria.

# Create a new set of fitted values for the new beta values found

# Convergence criteria

while(abs(dev-devold)/(0.1+abs(dev))>epsilon)  {

#______________________________________________________________________

# Step 2 - Compute the Pearson's residuals for the model

p.residuals = (rain-fits)/sqrt(fits^p)

#______________________________________________________________________

# Step 3a - Calculate alpha

# Calculate the new phi value and alpha

phi <- sum(p.residuals^2)/(n-r)

# Initialize alpha

alpha = NULL

# Obtain alpha value

alpha = sum(p.residuals[1:(n-1)]*p.residuals[2:n])

alpha = ((phi)*alpha)/(n*(n-r))

# Step 3b - Calculate R using the alpha values found in

# step 3a (using AR(1))

index  <- seq(0,n-1,by=1)

longindex <- c(seq(n-1,1,by=-1),index)

# Calculate R

i = 0

R = matrix(nrow=n,ncol=n)

while(i<n){
```

```
R[i+1,] = alpha^longindex[(n-i):(2*n-i-1)]
i = i+1
}
#______________________________________________________________________
# Step 4 - Calculate an estimate of the covariance matrix V
# using R found in step 3b.
# Calculate A:
A = diag(fits)^(p/2)
# Calculate V:
V = (A %*% R %*% A)
#______________________________________________________________________
# Step 5 - Find an updated version of beta
# Firstly find (partial mu / partial beta), let
# (partial mu/partial beta) be matrix "D"
xmat = glmmodel$x
D = matrix(nrow=n,ncol=r)
#Add the values to matrix "D"
for(i in (1:r)){
D[,i] = fits*xmat[,i]
}
# To find matrix beta(r+1) use the following notations
# beta = beta+inverse(C)*B,


#
```

```
# where C = transpose(D)*inverse(V)*D and

# B = transponse(D)*inverse(V)*(actual-fitted)

# Firstly find C

C = t(D) %*% solve(V) %*% D

# Find B

B = t(D) %*% solve(V) %*% (rain-fits)

beta=beta + (solve(C) %*% B)

# Fit the new values of dev, devold and fits for use in the

# covergence criteria

fits <- exp(t(beta) %*% t(xmat))

fits <- as.vector(fits)

devold <- dev

dev <- sum(tweedie.dev(rain,fits,p))

}

###################################################################

# DIAGNOSTICS

# Calculate QICu

# Calculate the quasi-likelihood first

quasi <- sum((rain*fits^(1-p)/(1-p))-((fits^(2-p))/(2-p)))

# Next calculate the QICu which is to find the best covariates to use

qicu <- (-2*quasi)+(2*r)

#__________________________________________________________________

# Calculate the Marginal R squared

marginal = (1/n)*sum(rain) # marginal component of R^2
```

```
top = sum((rain-fits)^2) # numerator of R^2

bottom = sum((rain-marginal)^2) # denominator of R^2

R2 = 1-(top/bottom)

#____________________________________________________________________

# Calculate the Wald-Wolfowitz run test to detect if the model

# is adequate and residuals are random.


# Calculate the raw residuals

residuals=rain-fits

# Initialise the values to use

run = NULL

nn = 0

np = 0

j = 1

# Start the test

while(j <= n){

if(residuals[j] <= 0){

run[j] = -1

nn = nn+1}

if(residuals[j] > 0){

run[j] = 1

np = np+1}

j = j+1

}
```

```
# Find the components E(T) and V(T) needed in the randomness test

ET = (2*np*nn)/(np+nn)+1

VT = (2*np*nn*(2*np*nn-np-nn))/((np+nn)^2*(np+nn-1))

# Find the total number of observed runs in the sequence


T=0

j=1

while(j<=(n-1)){

if(run[j]!=run[j+1]){

T=T+1}

j=j+1

}

T = T+1

# Find the test statistic W

W = (T-ET)/sqrt(VT)


#Print out all the relevant information

output1 <- data.frame(Diagnostic = c("alpha","QICu","R2","W"),

Data = c(alpha,qicu,R2,W))

output2 <- data.frame(BetaValues = c(beta),

Names = c(names(glmmodel$coefficient)))

print(output1)

print(output2)

####################################################################
```

```
##################################################################
# RESIDUALS PLOT FOR MODEL (1 + month + soiph)
# Plot of the Raw residuals
win.graph(width=11,height=7) # graphic size
plot(residuals,xlab="Observation Number",ylab="Raw Residuals",main="Plot of the
raw residuals using the month factor")
abline(0,0) # add a horizontal line at 0
dev.print(pdf, "D:/Documents and Settings/Administrator/My
Documents/February2012/rawresidualsmonth.pdf")
# Plot of Pearson Residuals versus linear predictor (eta=log(mu))
win.graph(width=11,height=7)
plot(log(fits),p.residuals,xlab="Linear Predictor", ylab="Pearson
Residuals",main="Plot of pearson residuals versus linear predictor (month)")
dev.print(pdf, "D:/Documents and Settings/Administrator/My
Documents/February2012/linearresidualsmonth.pdf")
##################################################################
# PREDICTED VALUES for (1 + month + soiph)
# Plot of predicted values for Nairobi
win.graph(width=12,height=6) # graphic size
# A time series plot of the amount of rain recorded during each
# dry and wet month and a plot of the predicted values for the
# amount of rain per month
# Observed rainfall:
plot(ts(rain,start=c(1959,1),frequency=12),
plot.type="single",col="blue", xlab="Year",ylab="Amount of rain
(mm)",main="Nairobi Monthly Rainfall")
abline(h=c(0,100,200,300,400,500),v=c(1960,1965,1970,1975,1980,1985,1990,1995
,2000),lty=2,lwd=.1,col="gray",las=2)
#Predicted Rainfall:
Nairobi.fitted<-ts(fits,start=c(1959,1),frequency=12)
points(Nairobi.fitted,type="l",col="red") #Add to the plot
dev.print(pdf, "D:/Documents and Settings/Administrator/My
Documents/February2012/Nairobiobsandpredict.pdf")
####################################################################################
# Finding a suitable link function
glmmodel<-glm(rain~sdec+atide+ldec+stdstl+lunaph+etide+lmean,
family=tweedie(var.power=p, link.power=0),x=TRUE) # Logarithm
glmmodel.other<-glm(rain~sdec+atide+ldec+stdstl+lunaph+etide+lmean,
family=tweedie(var.power=p),x=TRUE) # Canonical
#Deviances
glmmodel$deviance
glmmodel.other$deviance
#Df Residuals
glmmodel$df.residual
glmmodel.other$df.residual
####################################################################################
# Normal probability plot for the model (1 + month + soiph)
# First print the profile log-likelihood plot
```

```
win.graph(width=6,height=6)
power=tweedie.profile(rain~sdec+atide+ldec+stdstl+lunaph+etide+lmean,
p.vec=seq(1.3,1.9,length=10),do.plot=TRUE,do.smooth=TRUE,do.ci=TRUE,metho
d="interpolation")
dev.print(pdf,"D:/Documents and Settings/Administrator/My
Documents/February2012/logplotone.pdf")
p=power$p.max
glmmodel<-
glm(rain~sdec+atide+ldec+stdstl+lunaph+etide+lmean,family=tweedie(var.power=p,
link.power=0),x=TRUE)
win.graph(width=6,height=6) # graphic size
quantile=qres.tweedie(glmmodel) # Quantile residuals
qqnorm(quantile, main = "Normal probability plot \n for Nairobi
model",xlab="Standard Normal Quantiles",ylab="Quantile Residuals")
qqline(quantile) # Normality line
dev.print(pdf,"D:/Documents and Settings/Administrator/My
Documents/February2012/quantileonemonth.pdf")


p
R2
qicu
```

APPENDIX II

Beta values for all factors model.

rain~sdec(factor)+month(factor)+atide(factor)+atideph(factor)+soi(factor)+ldec(factor)+ soiph(factor)+qboph(factor)+ qbo(covariate)+stdstl(covariate)+ssph(factor). Diagnostic Data

```
1      alpha  2.048879e-03
2       QICu  2.839177e+04
3         R2  6.367770e-01
4          W -2.464647e+00
5         p = 1.544898
```

```
     BetaValues       Names
1    5.597500634 (Intercept)
2   -0.909023389 sdec[T.-21]
3   -0.325725032 sdec[T.-20]
4    0.454581261 sdec[T.-18]
5   -0.494059654 sdec[T.-13]
6   -1.562061648 sdec[T.-12]
7   -0.248279504  sdec[T.-9]
8   -0.854812813  sdec[T.-8]
9   -0.084215894  sdec[T.-2]
10   0.118942592  sdec[T.-1]
11  -1.479303845   sdec[T.2]
12  -1.788639032   sdec[T.3]
13   0.615739903   sdec[T.9]
14  -1.115206535  sdec[T.13]
15   0.697181716  sdec[T.18]
16   0.566785424  sdec[T.19]
17  -1.320627465  sdec[T.21]
18  -0.629060191  sdec[T.23]
19   0.406463919  atide[T.1]
20   0.565539541  atide[T.2]
21   0.811257037  atide[T.3]
22   0.394712574  atide[T.4]
23   0.342601238  atide[T.5]
24   0.501149141  atide[T.6]
25   0.784041140  atide[T.7]
26   0.552890810  atide[T.8]
27  -1.124155199  atide[T.9]
28  -4.250259274 soi[T.-5.8]
29  -3.926768565 soi[T.-4.7]
30  -0.655557648 soi[T.-4.4]
31  -1.694546484 soi[T.-4.3]
32  -1.608298648   soi[T.-4]
33  -0.871730179 soi[T.-3.7]
34  -1.041975077 soi[T.-3.4]
35  -3.755765403 soi[T.-3.3]
36  -1.688455181   soi[T.-3]
37  -2.423551141 soi[T.-2.9]
38   0.070211519 soi[T.-2.8]
39  -2.366020744 soi[T.-2.6]
40  -2.223418705 soi[T.-2.5]
41  -1.716012020 soi[T.-2.4]
42  -1.785445370 soi[T.-2.3]
43  -2.363316529 soi[T.-2.2]
44  -2.193756813 soi[T.-2.1]
45  -1.598173594   soi[T.-2]
46  -1.519375671 soi[T.-1.9]
47  -0.820647046 soi[T.-1.8]
48  -2.045809277 soi[T.-1.7]
49  -2.065455608 soi[T.-1.6]
```

50	-1.670541519	soi[T.-1.5]	72	-2.140230666	soi[T.0.7]	94	-3.551094352	soi[T.2.9]
51	-1.960444994	soi[T.-1.4]	73	-2.093700477	soi[T.0.8]	95	-2.030527813	soi[T.3]
52	-2.179657369	soi[T.-1.3]	74	-1.786085893	soi[T.0.9]	96	-0.929437519	soi[T.3.1]
53	-1.568082594	soi[T.-1.2]	75	-2.548319721	soi[T.1]	97	-2.430539173	soi[T.3.2]
54	-1.795224933	soi[T.-1.1]	76	-1.748629405	soi[T.1.1]	98	-1.834479780	soi[T.3.3]
55	-1.932400419	soi[T.-1]	77	-1.999821783	soi[T.1.2]	99	-0.614250002	soi[T.3.4]
56	-1.304399729	soi[T.-0.9]	78	-2.363690372	soi[T.1.3]	100	-1.875693531	soi[T.3.5]
57	-2.250413166	soi[T.-0.8]	79	-1.862530599	soi[T.1.4]	101	-3.525291903	soi[T.3.8]
58	-1.986763946	soi[T.-0.7]	80	-1.910133191	soi[T.1.5]	102	-1.916656299	soi[T.4]
59	-2.171466492	soi[T.-0.6]	81	-1.527998308	soi[T.1.6]	103	-2.138158996	soi[T.4.3]
60	-1.519189904	soi[T.-0.5]	82	-1.528363233	soi[T.1.7]	104	1.054267340	ldec[T.-27]
61	-2.176064142	soi[T.-0.4]	83	-1.686378683	soi[T.1.8]	105	0.425310068	ldec[T.-26]
62	-1.725591315	soi[T.-0.3]	84	-1.384599788	soi[T.1.9]	106	1.181525176	ldec[T.-25]
63	-2.005892298	soi[T.-0.2]	85	-1.306027526	soi[T.2]	107	0.432656623	ldec[T.-24]
64	-1.968763016	soi[T.-0.1]	86	-1.497822809	soi[T.2.1]	108	0.664673116	ldec[T.-23]
65	-1.349041925	soi[T.0]	87	-2.736990991	soi[T.2.2]	109	0.181169008	ldec[T.-22]
66	-1.974671868	soi[T.0.1]	88	-1.362158240	soi[T.2.3]	110	0.749358526	ldec[T.-21]
67	-1.647112589	soi[T.0.2]	89	-1.686523073	soi[T.2.4]	111	0.603964906	ldec[T.-20]
68	-2.627824476	soi[T.0.3]	90	-0.684832151	soi[T.2.5]	112	0.757348017	ldec[T.-19]
69	-1.926823973	soi[T.0.4]	91	-3.265009275	soi[T.2.6]	113	0.547698161	ldec[T.-18]
70	-2.165347131	soi[T.0.5]	92	-1.524190849	soi[T.2.7]	114	1.018329128	ldec[T.-17]
71	-2.183746717	soi[T.0.6]	93	-2.319040518	soi[T.2.8]	115	0.644080428	ldec[T.-16]

116	0.689702202	ldec[T.-15]	136	1.243941367	ldec[T.5]	156	0.443410070	ldec[T.25]
117	0.468838138	ldec[T.-14]	137	-0.137655587	ldec[T.6]	157	0.724683411	ldec[T.26]
118	0.461770630	ldec[T.-13]	138	0.229466553	ldec[T.7]	158	0.779387571	ldec[T.27]
119	0.248208100	ldec[T.-12]	139	0.609805981	ldec[T.8]	159	1.149875234	ldec[T.28]
120	0.525656823	ldec[T.-11]	140	1.408521505	ldec[T.9]	160	-0.306960043	soiph[T.2]
121	0.141652193	ldec[T.-10]	141	0.006909431	ldec[T.10]	161	-0.085750868	soiph[T.3]
122	0.792569648	ldec[T.-9]	142	0.448191123	ldec[T.11]	162	-0.094975974	soiph[T.4]
123	0.026086575	ldec[T.-8]	143	0.527695881	ldec[T.12]	163	-0.092357651	soiph[T.5]
124	0.922610712	ldec[T.-7]	144	0.793299939	ldec[T.13]	164	-0.035602476	soiph[T.6]
125	0.629114442	ldec[T.-6]	145	0.488906314	ldec[T.14]	165	-0.274412600	qboph[T.2]
126	0.418371910	ldec[T.-5]	146	0.745274216	ldec[T.15]	166	-0.38180146	qboph[T.4]
127	0.983783195	ldec[T.-4]	147	0.589435910	ldec[T.16]	167	-0.1330240	qboph[T.5]
128	0.787219433	ldec[T.-3]	148	0.347380356	ldec[T.17]	168	-0.1585512	qboph[T.6]
129	0.955113776	ldec[T.-2]	149	0.648492995	ldec[T.18]	169	-0.008878874	qbo
130	0.049197483	ldec[T.-1]	150	-0.015970717	ldec[T.19]	170	-0.150930294	stdstl
131	-0.079215051	ldec[T.0]	151	0.784670265	ldec[T.20]	171	-0.046291112	ssph[T.2]
132	0.247899411	ldec[T.1]	152	0.790335953	ldec[T.21]	172	0.046916289	ssph[T.3]
133	0.682139894	ldec[T.2]	153	0.394726083	ldec[T.22]	173	0.052201940	ssph[T.4]
134	1.250880705	ldec[T.3]	154	0.309990996	ldec[T.23]			
135	0.640117251	ldec[T.4]	155	0.844097545	ldec[T.24]			

APPENDIX III

Beta values for the model 12.3

rain~ sdec(F)+atide(F)+ldec(F)+stdstl(F)+lunaph(F)+etide(F). Section 4.4.1

```
1      alpha  8.828505e-04        11  -1.32908324    sdec[T.2]    28   1.03633662  ldec[T.-27]
2       QICu  2.207748e+04        12  -1.54433140    sdec[T.3]    29   0.62064499  ldec[T.-26]
3         R2  5.225058e-01        13   0.5475040    sdec[T.9]     30   1.07895584  ldec[T.-25]
4          W -1.967940e+00        14  -1.48406711   sdec[T.13]    31   0.84251407  ldec[T.-24]
                                  15   0.80771756   sdec[T.18]    32   0.37954092  ldec[T.-23]
> print(output2)                  16   0.34383814   sdec[T.19]    33   0.345413023 ldec[T.-22]
      BetaValues        Names     17  -1.83752574   sdec[T.21]    34   0.67328711  ldec[T.-21]
1    3.121645714  (Intercept)     18  -0.91437115   sdec[T.23]    35   0.72148902  ldec[T.-20]
2   -0.948907811 sdec[T.-21]      19   0.236665190 atide[T.1]     36   0.599130297 ldec[T.-19]
3   -0.25971855  sdec[T.-20]      20   0.2161033     atide[T.2]   37   0.43904532  ldec[T.-18]
4    0.38196194  sdec[T.-18]      21   0.249654381   atide[T.3]   38   0.89782494  ldec[T.-17]
5   -0.58105408  sdec[T.-13]      22   0.330429063   atide[T.4]   39   0.33356006  ldec[T.-16]
6   -2.216406993 sdec[T.-12]      23   0.063540494   atide[T.5]   40   0.615511467 ldec[T.-15]
7   -0.42585197   sdec[T.-9]      24   0.335320762   atide[T.6]   41   0.091629584 ldec[T.-14]
8   -0.598690837   sdec[T.-8]     25   0.401749492   atide[T.7]   42   0.563766592 ldec[T.-13]
9   -0.120214816   sdec[T.-2]     26   0.960526664   atide[T.8]   43   0.67526790  ldec[T.-12]
10   0.15206064   sdec[T.-1]      27  -0.625583256   atide[T.9]   44   0.49918324  ldec[T.-11]
```

45	0.21685133	ldec[T.-10]	64	1.637626784	ldec[T.9]	83	0.68064165	ldec[T.28]
46	1.05393771	ldec[T.-9]	65	0.15549187	ldec[T.10]	84	0.786065590	stdstl[T.18]
47	-0.767512849	ldec[T.-8]	66	0.63098437	ldec[T.11]	85	1.0971404856	stdstl[T.19]
48	0.345310892	ldec[T.-7]	67	0.54612030	ldec[T.12]	86	0.50414133	stdstl[T.20]
49	0.504856939	ldec[T.-6]	68	0.68099059	ldec[T.13]	87	0.41795094	stdstl[T.21]
50	0.214054307	ldec[T.-5]	69	0.62854862	ldec[T.14]	88	0.87264546	stdstl[T.22]
51	0.84317045	ldec[T.-4]	70	0.65741639	ldec[T.15]	89	1.05626034	stdstl[T.23]
52	0.800854583	ldec[T.-3]	71	0.73388159	ldec[T.16]	90	0.41510860	stdstl[T.24]
53	0.838829262	ldec[T.-2]	72	0.38993848	ldec[T.17]	91	0.71421715	stdstl[T.25]
54	0.187141670	ldec[T.-1]	73	0.565141536	ldec[T.18]	92	0.59004232	stdstl[T.26]
55	0.422252065	ldec[T.0]	74	0.03039593	ldec[T.19]	93	0.49172636	stdstl[T.27]
56	0.094392568	ldec[T.1]	75	0.81896849	ldec[T.20]	94	0.82420259	stdstl[T.28]
57	0.840679234	ldec[T.2]	76	0.53822974	ldec[T.21]	95	0.73926507	stdstl[T.29]
58	0.953659577	ldec[T.3]	77	0.62169479	ldec[T.22]	96	-0.2330166	lunaph[T.2]
59	0.788665884	ldec[T.4]	78	-0.00254907	ldec[T.23]	97	-0.1050344	lunaph[T.3]
60	1.015993287	ldec[T.5]	79	0.57337699	ldec[T.24]	98	-0.02519105	lunaph[T.4]
61	-0.000526230	ldec[T.6]	80	0.27446374	ldec[T.25]	99	0.00760904	etide[T.1]
62	-0.143617144	ldec[T.7]	81	0.75400187	ldec[T.26]	100	0.03293729	etide[T.2]
63	0.568590918	ldec[T.8]	82	0.52430738	ldec[T.27]			