PREDICTING RAINFALL AND DROUGHT USING THE SOUTHERN OSCILLATION INDEX IN DROUGHT PRONE ZIMBABWE

BY

R. Chifurira and

D. Chikobvu

December 2010

CHAPTER 1

INTRODUCTION

1.1 Background to the study

Zimbabwe is situated in Southern Africa between latitudes $15^0 30''$and $22^0 30''$ South of the equator and between longitudes 25^0 and $33^0 10''$ East of the Greenwich Meridian. It is a land locked country that shares its border with Mozambique to the east, South Africa to the south, Botswana to the west and Zambia to the north. It has a land area of approximately 390 757 square kilometres.

Zimbabwe has in the past years been severely affected by erratic rainfall patterns and sometimes droughts. During the 1991 to 1992 rainy season, Zimbabwe and some SADC countries in Southern Africa experienced the worst drought in living memory (Zimbabwe Central Statistical Office Report, 1994). In the year 2000, Zimbabwe was ravaged by cyclone $Eliñe$. Between 2001 to 2003, Zimbabwe had rainfall in the first half of the rainfall season and a dry spell in the second half resulting in severe drought in some parts of the country. Between 2004 to 2008, Zimbabwe received average rainfall in the northern parts of the country and other parts received very little rainfall or no rainfall. The 2009 to 2010 rainfall seasons, Zimbabwe received below average rainfall in the first half of the rainfall season and above average rainfall in the second half of the rainfall season (Zimbabwe Central Statistical Office Report, 2010). Following these constant changes in the rainfall patterns in Zimbabwe, research on the prediction of the amount of rainfall to be received by the country becomes very paramount, to help farmers plan well for each rainfall season.

Zimbabwe‘s economy is agro based thus vulnerable to the effects of climatic change despite the country’s insignificant contributions to the global climatic change through industrialization. The severe impact of climate change is due to the fact that the country does not have adequate resources or technology for adaptation to the conditions that come with climate change. The challenges range from droughts, floods and cyclones to the more recent high seasonal rainfall variability (Washington and Preston, 2006). Droughts occur frequently and are severe, impacting negatively on the country’s economic performance with more than 50% of the gross domestic product (GDP) being derived from rain-fed agriculture (Jury, 2002). Thus, for a country such as

Zimbabwe, where agriculture is the main driver of the economy and upon which over 80 percent of the population is directly dependent on, it is imperative that a simple tool be advanced to predict rainfall patterns as early as a year in advance and as accurately as possible. The objective of this study is to advance a model of the relationship between Southern Oscillation Index (SOI), a climatic determinant, and Zimbabwe annual rainfall and droughts.

Zimbabwe rainfall

Zimbabwe's rainfall is very seasonal, with one wet season running from mid – November through to mid – March with the peak of the season stretching from December to January and February (Torrance, 1981). The dates of the onset and end of the rainy season vary from one season to the other. At times the season starts as early as October and extending well into April. The main feature of the season is the Inter Tropical Convergence Zones (ITCZ) which moves southwards with the sun bringing with it copious rain. Over the south , the dryish south easterly air flow persists and rainfall tends to last for a few days and alternated with dry spells. The main rains are associated with the behaviour of the ITCZ, whose oscillatory behaviour is influenced by changing pressure patterns to the north and south of the country. Records shows that the heaviest 10% of rain days account for almost 45% of the entire annual precipitation (Buckle , 1996).

Zimbabwe lies in the South West Indian Ocean zone that is often affected by tropical cyclones. Tropical cyclones are low pressure systems which in the Southern hemisphere have a well defined clock wise wind circulations spiralling towards the centre with great intensity. The strongest winds and heaviest rains occur in the region close to the centre. Cyclones that develop over the western side of the Indian ocean occasionally have an impact on the rain season. The amount and intensity of rainfall during a given wet spell is enhanced by the passage of upper westerly waves of mid – latitude origin (Smith, 1985, Buckle, 1996).

Drought

A meteorological drought is an insidious natural hazard characterised by lower than expected or lower than normal precipitation (in the Zimbabwean case lower than 75 percent of 630mm) which is insufficient to meet the demands of human activities and the environment (World

Meteorological Organisation, 2006). A meteorological drought is a normal part of climate, although its spatial extent and severity will vary on seasonal and annual time scales.

At this point, it is important to mention the various definitions of drought. Zimbabwe Meteorological Services classifies droughts into meteorological, agricultural and hydrological droughts.

> **Meteorological drought** – is usually defined by a precipitation deficiency threshold over a predetermined period of time. The threshold chosen, such as 75 percent of the normal precipitation, and duration period, for example, six months, will vary by location according to user needs or applications. A meteorological drought is a natural event and results from multiple causes, which differ from region to region.
>
> **Agricultural and hydrological drought** – agricultural and hydrological drought place greater emphasis on the human or social aspects of drought highlighting the interaction or interplay between the natural characteristic of meteorological drought and human activities that depend on precipitation to provide adequate water supplies to meet societal and environmental demands. Agricultural drought is defined more commonly by the availability of soil water to support crop and forage growth than the decline of normal precipitation over some specified period of time. Hydrological drought is even further removed from the precipitation deficiency since it is normally defined by the decline of surface and subsurface water supplies like lakes, reservoirs, aquifers and streams from some average condition at various points in time (World Meteorological Organisation, 2006).

Prior to the advent of 1982 to 1984 drought, which caused widespread environmental impacts in the country, prediction of drought was not taken seriously. With severe droughts recurring in 1992 and 1997, it has become mandatory to predict drought. This research, tries to predict meteorological drought using the Southern Oscillation Index.

Southern Oscillation Index (SOI)

The Southern Oscillation gives a simple measure of the strength and phase of the difference in sea level air pressure between Tahiti and Darwin **(see section 3.2.2)**. A strong and consistent positive SOI pattern is related to $La\,Niña$. Conversely, a deep and consistent negative SOI

pattern is related to $El\ Niño$. $El\ Niño$ (associated with negative SOI phases) is usually (but not always) associated with below normal rainfall and $La\ Niña$ (associated with positive SOI phases) is associated with above normal rainfall. $El\ Niño$ is the abnormal warming of surface ocean waters in the eastern tropical Pacific Ocean. While, $La\ Niña$ is the cooling of surface ocean waters in the eastern tropical Pacific Ocean. If $El\ Niño$ takes place in the eastern tropical Pacific Ocean (Northern Hemisphere), $La\ Niña$ will simultaneously take place in the western tropical Pacific Ocean. The changes in temperature of the ocean waters affect surface air pressure in the Pacific Ocean, a phenomena known as Southern Oscillation. Southern Oscillation is the see-saw pattern of reversing surface air pressure between the eastern and western tropical Pacific Ocean: when the surface pressure is high in the eastern tropical Pacific Ocean, it is low in the western tropical Pacific Ocean and vice versa.

Because the ocean warming and pressure reversal are, for most parts, simultaneous, the phenomenon is called $El\ Niño$/ Southern Oscillation (ENSO)

$ENSO$ stands for $El\ Niño\ Southern\ Oscillation$

ENSO refers to both $El\ Niño$ and $La\ Niña$.

Source: International Research Institute for Climate Prediction (2010)

ENSO is more about positive phases and negative phases of Southern Oscillation and it is difficult to separate it from the SOI. This research uses monthly SOI values to predict annual rainfall and drought in Zimbabwe .

1.2 STATEMENT OF THE RESEARCH PROBLEM

The research seeks to predict the annual rainfall and droughts in Zimbabwe from the SOI level. It aims to identify the month and maximum lag whose SOI level explains annual rainfall and drought in the country.

1.3 Aims and Objectives of the study

The research's main aim and objectives is to identity the reliability of the SOI index in the predicting of annual rainfall and drought in Zimbabwe .

This shall be achieved by:

1. Investigating the relationship between SOI and Zimbabwe rainfall.
2. Determining the month whose SOI level and also maximum lag which best explains the annual rainfall in Zimbabwe.
3. Identifying a regression model for the prediction of annual rainfall in Zimbabwe.
4. Identifying a binary model for the prediction of drought in Zimbabwe .
5. Carrying out model diagnostic checking of the identified models – checking model assumptions, goodness of fit and model validation using historical rainfall and drought data obtained from the Department of Meteorological Services in Zimbabwe.
6. Suggesting recommendations and possible areas of further research.

1.4 Significance of study

The research is enrichment to knowledge; it aims to explore an area that has not been researched before, since many researchers have focused on correlations between SOI phases (ENSO) and monthly or annual rainfall of some stations in the country. A more accurate rainfall predicting model will help many Zimbabwean farmers who rely on rainfall rather than irrigation for their farming activities. Ultimately, the Zimbabwean economy, which is agro- based will be the greatest beneficiary from this research.

1.5 Delimitations

The research will be focusing on identifying the month whose SOI level and the maximum lag which best explains the annual rainfall and Drought in Zimbabwe. The research will use 1974 to 2009 mean annual rainfall data obtained from Zimbabwe Department of Meteorological Services.

1.6 Limitations

The research faced challenges in securing adequate rainfall data since the researcher had envisaged to gather data as far back as 1950. Securing the data was a major challenge.

1.7 Project layout

Chapter two discusses an overview of documented researches on predictability of rainfall and drought. Chapter three presents the methodology of analysing the data. Chapter four consists of a discussion of the data and detailed analysis of results from the models of predicting annual rainfall and drought in Zimbabwe. Chapter five summaries the research findings and gives the recommendations from the study.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

This chapter discusses some documented researches on prediction of rainfall and drought using Southern Oscillation Index and other factors which influence rainfall. It also discusses other related studies on prediction of rainfall in other countries.

2.2 Documented Researches

The relationships between the Zimbabwean rainfall and the SOI have been well documented. Significant differences in rainfall amount, temporal and spatial distribution have been found to occur in the country between opposite extremes in the phases of the ENSO (Matarira, 1990). Matarira found that during the warm phase of ENSO, the rainfall tends to be depressed across much of the country, whereas the converse is true for the cold phase. The study used an average of the SOI during the preceding 12 months (January to December) and found that there is positive correlation of +0.42 with November to April rainfall in Zimbabwe. Matarira and Unganai (1994), observed a peak correlation of +0.56 with November, December, January rainfall in the Southeast part of the country using monthly SOI at 1 to 2 month's lag. They concluded that Southern Oscillation can therefore explain up to about 30 percent of the inter – annual variation in summer rainfall in some parts of the country. Torrance (1990) compared SOI anomalies with the Zimbabwean seasonal rainfall, and found that positive values of SOI coincide with rainfall levels of 101to125 percent of the normal. Negative value of SOI are characterised generally with rainfall below normal rainfall. Torrance's study focused on correlations between positive and negative values of SOI with rainfall. His research grouped positive SOI levels into a positive phase and negative SOI levels into negative phase. In contrast, this research does not group SOI into phases (positive and negative SOI phases) but aims to determine a particular month and lag whose SOI explains total annual rainfall in Zimbabwe, that is, it seeks to determine the explanatory variable with a lag period.

Makarau and Jury (1997) found that high (low) ENSO phases and extreme above (below) normal rainfall occurrences are associated. According to the study's analysis, when the SOI level is

within one standard deviation from the long term mean, there will be a high probability that rainfall in Zimbabwe will be within 10 percent of the mean. The study used a 41 year record and found a high correlation of +0.44 between SOI and the Zimbabwean summer rainfall using August to October average SOI value. Rocha (1992) found that South East Zimbabwean rainfall correlate significantly with SOI (+0.4) at lead time of 4 to 5 months.

Researches done so far do not only focus on finding the impact on SOI on rainfall, but also used SOI to model maize production (Martin et al, 1999). Martin et al concluded that water stress for primary maize growing regions is related to ENSO indices. They determined water stress as the soil moisture content which cannot sustain the survival of crops i.e. soil water which is limited relative to crop requirements. SOI is used as an indicator of ENSO (Jones,1991). Cane et al (1994) achieved remarkable success in their assessment of forecasting both maize yield and rainfall in Zimbabwe using $Niño-3$ index of the $El\ Niño$ Southern Oscillation (ENSO). The $Niño-3$ index was calculated by spatially averaging Sea Surface Temperatures anomalies over the Pacific Ocean. Martin et al (1999) averaged ENSO indices over the following 3 – month periods of July to September; August to October; September to November; October to December; November to January and found that the water stress time series and ENSO indices were highest at a 4 – month lead with respect to a May harvest. The study found that SOI has +0.67 correlations with South African water stress time series (STS), with a regression model:

$$STS_{predicted} = \hat{\beta}_0 + \hat{\beta}_1 SOI$$

Where the estimates of β_0 and β_1 are 64.9 and 8.8 respectively. However, some observed natural phenomenon often have their variance increasing with the SOI levels making simple linear regression inadequate. For the Zimbabwean water stress time series (ZTS), the authors found that it is better predicted by the $Niño-3$ index($r = 0.38, p < 0.01$), with a regression model;

$$ZTS_{predicted} = \hat{\beta}_0 + \hat{\beta}_1 (Niño - 3)$$

Where the estimates of β_0 and β_1 are 86.6 and -5.0 respectively. However, for Zimbabwe to the best of our knowledge no one has come up with a SOI index based model for predicting mean annual rainfall. Most other researches refer to the simple correlations coefficient between annual rainfall and SOI levels.

Martin et al also made a comparison of rainfall forecasts for South African seasonal rainfall using SOI and found a correlation of +0.53 ($p < 0.02$) with the training data and a forecast correlation of +0.60 ($p < 0.005$) with the validation data. In Zimbabwe seasonal rainfall yielded a correlation of +0.40 with$Niño - 3\ index$. This research intends to use SOI values rather than Sea Surface Temperature (SST) values used by Martin et al (1999).

Makarau and Jury (1997) divided the Zimbabwean area rainfall into early and late summer seasons which are; November to mid – January and mid – January to March respectively. They developed a multivariate linear regression statistical model using a wide ranging predictor data set in a forward step wise approach. They used 120 predictor variables, including SST area indices, air pressure, surface and upper wind indices, convective indices, SOI etc. The multivariate algorithms for Zimbabwe were:

Early summer rain

$$+0.37(Wi\ Sip) + 0.5(NEolr) + 0.25(ATpc2) - 0.40(AtlW) + 0.34(WCi\ ABp)=71$$

Late summer rain

$$+0.36(OSTang) + 0.65(aCIst) - 1.09(oCIst) - 0.37(oSocns) + 0.30(aSwv) = 80$$

Where $WiSip$ is a measure of the west to south east pressure gradient across the Tropical Indian Ocean; $NEolr$, the convective anomaly in the north east Tropical Indian Ocean; $ATpc2$, Atlantic SST principal component 2 which is positively loaded to the south east of Brazil; $AtlW$, the 200hPa zonal wind anomaly over the equatorial central Atlantic Ocean and $WCi\ ABp$, a measure of the South to North pressure gradient across the western tropical Indian Ocean. In summer rainfall model: $oSTang$, is the September to November value of Atlantic SST anomalies north west of Angola; $aCIst$, is the central Indian Ocean SST key area; $aSWv$, is the meridional surface wind component in the south west Indian Ocean subtropics south of Madagascar in January, April and September and $oSocns$, is the Southern Ocean SST. 71 and 80 are percentage correlations, the respective models produces in the jack-knife skill test for the period 1971 to 1992.

Using an average 200hpa equatorial Atlantic zonal wind index, Makarau and Jury (1997) showed that up to 49 percent of the observed variance in the Zimbabwean summer rainfall is predictable.

However, the study did not investigate how much variance is already accounted for by the effects of ENSO.

Waylen and Henworth (1995) investigated the association between monthly precipitation totals and the SOI throughout Zimbabwe. They used monthly precipitation data from 68 meteorological sites possessing at least 20 years of complete records. Simple lag cross correlations between the SOI and precipitation totals at both the annual and monthly time scales were used to determine significant (at 0.05 level) association. Simple lag cross correlation between the SOI and annual precipitation revealed significant positive correlations with almost 30 percent of the stations at lag zero. They found a negative correlation at lag 1. The research found that significant correlation was geographically restricted, mainly to the east i.e in Manicaland and to the north i.e. in Mashonaland North. Significant correlations were also found in Mashonaland South and Midlands. Areas around Harare, the west and south of the country, significant correlations were absent. The study found that periods of greatest positive association are months of the rainy season October to April which are correlated to synchronous values of the SOI and those in the preceding June to September period. March's precipitation was found to be correlated strongly. Over 70 percent of the stations in their study reported significant correlations between March's precipitation and SOI in the preceding July and at least 25 percent of all stations reported similar associations to monthly SOI from the preceding May (lag 1) through to February (lag 1). The correlation was found to extend through the end of the rainy season into April and May. November and December were found to display weaker correlation to SOI values in the same period. January and February were found to be uncorrelated with the SOI.

Mason and Goddard (2000) investigated the influence of extreme ENSO states on global precipitation anomalies using contingency tables. Precipitation anomalies that are only weakly positive or negative are considered near normal and are not counted in the climate impacts. For a total of n years, of which b are "dry", and from which r years are selected at random (the r strongest $El\ Niño$ years, for e.g), letting the number of dry years that are selected to be denoted by X be equal to x [where $0 \leq x \leq \min(r, b)$]. They assumed there is a total of $r\ El\ Niño$ years and that x of them are dry, then the significance is defined as the probability of selecting x

or more dry years in a random sample of r years. This probability is equivalent to the right tail area of the hypergeometric distribution (Agresti, 1996) and is given by:

$$P_x(X \geq x) = H(x; r, b, n) = \sum_{k=x}^{\min(r,b)} \frac{\binom{b}{k}\binom{n-b}{r-k}}{\binom{n}{r}}$$

Manatsa et al (2007) used correlation analysis to identify the period lags for which the SOI and Darwin pressure anomalies are significantly correlated with the Zimbabwean summer precipitation index. They found that progressive lagged four months averaged Darwin pressure anomalies and the SOI are correlated with the Zimbabwean summer precipitation:

	JFMA	FMAM	MAMJ	AMJJ	MJJA	JJAS	JASO	ASON	SOND
SOI	0.184	0.247	0.237	0.269	0.333	0.366	0.394	0.420	0.398
Darwin	-0.198	-0.293	-0.326	-0.303	-0.300	-0.297	-0.312	-0.341	-0.320

Ropelewski and Halpert (1998) showed that the high SOI – precipitation relationships show the opposite sign of those documented from the low index. They found that precipitation relationships were consistent holding for over 70 percent of the high SOI years. The study revealed that high SOI is associated with enhanced precipitation for the monsoons of India and Northern Australia.

Although a lot has been documented on the predictability of Zimbabwean rainfall, the studies only determined correlations between rainfall of some stations with SOI. Water inflow into the Gariep dam was modelled by incorporating the effect of the SOI. The model used the previous year's SOI of different months and found that the SOI level of October of the previous year has a high correlation with total annual inflow into the dam (Bekker et al, 2010). The aim of this research is to model the annual rainfall in Zimbabwe incorporating the effect of the SOI. The study therefore departs from previous researches by attempting to model the annual rainfall in Zimbabwe by incorporating the effect of the SOI.

CHAPTER 3

METHODOLOGY

3.1 Introduction

This chapter outlines the data sources and the methodology applied in gathering, analyzing and discussing the data. The chapter also discusses in detail statistical tests used in the research.

3.2 Data sources

3.2.1 Rainfall

The historical annual mean rainfall data for Zimbabwe dating from as far back as 1974 to 2009 was collected from the Central Statistical Office (CSO) and Zimbabwe Department of Meteorological Services. The mean annual rainfall values were calculated from averaging the monthly rainfall totals of the summer period that stretches from October to March. The drought years from 1974 to 2009 were determined using a formula supplied by the Department of Meteorological Services:

Drought year if: Mean annual rainfall $< 0.75 \times average\ expected\ annual\ rainfall$

The $average\ expected\ annual\ rainfall$ is taken as the average rainfall from 1980 to 2010 and is 630mm. Any mean annual rainfall less than 75% of 630mm is regarded as a drought year. The Department also uses a 30- year time series obtained from aerially averaging ten rainfall stations with long enough rainfall data sets. The Department classifies the degree of wetness and dryness relative to the 30 years. The 630mm is the average of the ten rainfall stations from 1980 to 2010.

3.2.2 Southern Oscillation Index

The SOI data was obtained from the internet *http://www.longpaddock.qld.gov.au*. The SOI is calculated from the monthly or seasonal fluctuations in the air pressure difference between Tahiti and Darwin. The SOI gives a simple measure of the strength and phase of the difference in sea-level pressure between Tahiti (in the mid-Pacific) and Darwin (in Australia). The difference is given in terms of an index. A strong negative value usually indicates that the oscillation has entered an $El\ Niño$ phase. A strong positive value usually indicates a $La\ Niña$ phase.

SOI is calculated using the formula:

$$SOI = 10\left(\frac{Pdif - Pdiffav}{SD(Pdif)}\right)$$

where:

Pdiff = (average Tahiti MSLP for the month) – (average Darwin MSLP for the month)

Pdiffav = Long term average of Pdiff for the month in question, and

SD(Pdiff) = Long term standard deviation of Pdiff for the month in question.

Source: The Australian Bureau of Meteorology (2010)

The multiplication by 10 is a convention. Using this convention the SOI ranges from about -35 to about +35, and the value of the SOI can be quoted as a whole number.

3.3 Regression Analysis

Regression analysis includes any statistical technique of modelling and analyzing several variables, when the focus is on the relationship between a dependent (response) variable and one or more independent (explanatory) variables. Regression analysis helps us understand how the typical value of the dependent variable changes when any one of the independent variables is varied, while the other independent variables are held fixed. Regression analysis is now the most widely used statistical technique for example linear regression to handle data with a linear relationship:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_p x_p + \varepsilon$$

where y is the dependent variable, x_i is the independent variable, β_i's are the model parameters, ε is the random error term and p is the order of the multiple regression model.

This research used linear regression to determine the relationship between rainfall (dependent variable) and the Southern Oscillation index (independent variable).

3.3.1 The Assumptions of linear regression

The basic assumptions for regression analysis which need to be checked are:

1. $linearity$: the dependent and the independent variables should have a linear relationship.
2. $normality$: the errors ε_t's at each time period t are normally distributed. Where t is the length of the series.
3. $zero\ mean$: the error is assumed to be a random variable with a mean zero conditional on the explanatory variable.

$$E(\varepsilon_t) = 0$$

4. $homoscedasticity$: the variance of the errors is constant across observations.

$$Var(\varepsilon_t) = \sigma^2$$

5. $no - autocorrelation$: the errors are uncorrelated.

$$Cov(\varepsilon_i; \varepsilon_j) = 0, \qquad \text{for times} \quad i \neq j$$

Summarily, the random error term ε_t, are independent and identically normally distributed with mean zero and constant variance σ^2.

$$\varepsilon_t \sim N(0; \sigma^2)$$

These assumptions imply that the parameter estimates will be unbiased, consistent and efficient in the class of linear unbiased estimators (Dielman, 1991).

3.4 Estimation of the Regression Coefficients

In this research, the ordinary least square technique is used to estimate the regression coefficients for the simple linear regression model:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

The estimates are:

$$\hat{\beta}_1 = \frac{\sum_{t=1}^{T}(x_t - \overline{x})(y_t - \overline{y})}{\sum_{t=1}^{T}(x_t - \overline{x})^2}$$

and $$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$$

where $\overline{x}$ is the mean of the $x\ values$ and $\overline{y}$ is the mean of the $y\ values$. Under the assumption that the population error term has a constant variance and the estimate of the variance is given by:

$$s^2 = \frac{SSE}{N-2}$$

SSE is the sum of square for the errors, s^2 is called the Mean Square Error (MSE) of the regression. The standard errors of the parameter estimates are given by:

$$s_{\beta_0} = \hat{\sigma}_\varepsilon \sqrt{\frac{1}{n} + \frac{\overline{x}^2}{\sum_{t=1}^{T}(x_t - \overline{x})^2}}$$

$$s_{\beta_1} = \hat{\sigma}_\varepsilon \sqrt{\frac{1}{\sum_{t=1}^{T}(x_t - \overline{x})^2}}$$

The main aim of performing the regression analysis is to find out if:

1. the independent variable truly influence the dependent variable.
2. there is adequate fit of the data to the model.
3. the model adequately predict responses.

3.5 Model selection

There are several techniques of selecting a good model and some of the methods used include:

- $p\ \boldsymbol{values}$

The probability of drawing a $t\ statistic\ (or\ a\ z\ statistic)$ as extreme as the one actually observed, under the assumption that the errors are normally distributed, or that the estimated coefficient are asymptotically normally distributed will be used. This probability is also known as the as the $p\ value$. A $p\ value$ of lower than the significance level α is taken as evidence to reject the null hypothesis of a zero coefficient.

- *Information Criterion*

  The notion of an information criterion is to provide a measure of information that strikes a balance between the measure of goodness of fit and parsimonious specification of the model.

  a) Akaike Information Criterion

  The Akaike Information Criterion (AIC) assumes that the model errors are normally and independently distributed. AIC is computed as:

$$AIC = -\frac{2l}{T} + \frac{2k}{T}$$

  where l is the log-likelihood; k is the number of parameters to be estimated using T observations. The model with the lower AIC value is preferred and hence selected. AIC is often used in model selection for non- nested alternatives.

  b) Schwaiz Criterion

  The Schwaiz Criterion also known as the Bayesian Information Criterion (BIC) is an alternative to the AIC that imposes a large penalty for additional coefficients. Like the AIC the BIC assumes that the model errors are normally distributed and the model with the least BIC is also selected. BIC is computed as:

$$BIC = -\frac{2l}{T} + \frac{(klog\,T)}{T}$$

  where T is the sample size, l is the maximized value of the likelihood function for the estimated model (Tsay, 2002).

- $F - \boldsymbol{statistic}$

  The F- statistic test the hypothesis that the slope coefficient in the regression is zero.

$$F = \frac{R^2/(k-1)}{(1-R^2)/(T-k)}$$

  under the null hypothesis with normally distributed errors this statistic has an F distribution with $k-1$ numerator degrees of freedom and $T-k$ denominator degrees of

freedom. Where k is the number of parameters to be estimated. If the $p\ value$ is less than the significance level α, the null hypothesis that is, the slope coefficient is equal to zero is rejected. $F\ tests$ are criticised for the fact that the test is a joint test, so that even if all the $t\ statistics$ are insignificant, the $F\ statistic$ will be highly significant. However we have a variable (SOI) as the explanatory and so in this research, models will be selected using the probability of drawing a $t\ statistic(or\ a\ z\ statisic)$, AIC and BIC values. The model with least AIC and BIC values with a $p\ value$ less than significance level will be selected (Tsay, 2002).

3.6 Model diagnostic techniques

3.6.1 Residual Analysis

Inferences concerning relationships of any system must be based on a satisfactory model, that is, a model which seems to fit the data well. A model is plausible or satisfactory if none of its assumptions are grossly violated. Thus, before a model is used to make inferences it must be subjected to diagnostic checking for adequacy.

3.6.1.1 Test of normality

Violation of normality of residuals compromise the estimation of regression coefficients. Sometimes the error distribution is skewed by the presence of a few large outliers since parameter estimation is based on the minimization of squared error. A few extreme observations can exert a disproportionate influence on parameter estimation.

a) The Jarque-bera Test

The Jarque-bera test is a two-sided goodness of fit test suitable when a fully-specified null distribution is unknown and its parameters must be estimated. The test statistic is:

$$JB = \frac{n}{6}\left(s^2 + \frac{(K-3)^2}{4}\right)$$

The JB statistic has an asymptotic chi-square distribution with two degrees of freedom and can be used to test the null hypothesis that the data are from a normal distribution.

The null hypothesis is a joint hypothesis of both the skewness and excess kurtosis being zero, since samples from a normal distribution have an expected skewness of zero and an expected excess kurtosis of zero. Any deviation from skewness of zero and kurtosis of zero increases is the JB statistic. The weakness of the JB is that it is not a powerful test for small samples of $n \leq 2000$ (Jarque and Bera, 1987). In this research the sample size is far less than 2000, hence, JB test will not be used to test normality of the residuals. The best advice recommended by statisticians is to generate a Quantile – Quantile (Q – Q) plot using the normal distribution. A straight line in a Q – Q plot indicates normality (Chambers et al, 1983). Thus, this research will use the quantile – quantile plots to test for normality.

Quantile – Quantile plots

A Q – Q plot is a probability plot which is a graphical method for comparing two distributions by plotting their quantiles against each other. If the two distributions being compared are similar, the points in the Q – Q plot will approximately lie on the line $y = x$. If the distributions are linearly related, the points in the Q – Q plot will approximately lie on a line, but not necessarily on the line $y = x$. If the points follow the line $y = x$ they suggests that the data are normally distributed (Chambers et al, 1983).

3.6.1.2 Homoscedasticity

Violations of homoscedasticity make it difficult to gauge the true standard deviation of the forecast errors, usually resulting in confidence intervals that are too wide or too narrow.

a) Plot of residuals against predicted values

If the assumptions of linearity, independence, homoscedasticity and normality of General Linear Model are held, then it implies that a plot of residuals against predicted values should show a good fit characterized by small residuals with no apparent structure or pattern.

b) Plot of residuals against independent variables

A plot of residuals against predicted values results in a horizontal band to indicate a good model.

3.6.1.3 Independence of residuals

Residuals must not be auto - correlated. Serial correlation in the residuals means that there is room for improvement in the model, and extreme serial correlation is often a sign of a mis-specified model. Serial correlation is also sometimes a by product of a violation of the linearity assumption.

a) Durbin – Watson Statistic

The Durbin-Watson Test is used to test for presence of auto-correlation. The hypothesis to be tested is:

$H_0 : \rho = 0$ Against $H_1 : \rho \neq 1$

The above hypothesis is tested indirectly by testing the hypothesis

$H_0 : \mu_d = 2$ Against $H_1 : \mu_d \neq 2$

Where μ_d = E (d) and the test statistic d is given by:

$$d = \frac{\sum_{t=2}^{n}(\hat{u}_t - \hat{u}_{t-1})^2}{\sum_{t=1}^{n}\hat{u}_t^2}$$

$$\hat{\rho} = \frac{\sum_{t=2}^{n}\hat{u}_t\hat{u}_{t-1}}{\sum_{t=1}^{n}\hat{u}_t^{\,2}} \text{ for large } n.$$

the decision rule is,

- If $d < d_L$ reject H_0 in favour of H_1, i.e in favour of positive correlation.
- If $d > 4 - d_U$ reject H_o if in favour of H_1, i.e in favour of negative correlation.
- If $d_U < d < 4 - d_U$ accept if there i.e there is no auto-correlation.
- If $d_L < d < d_U$ the test is considered inconclusive.
- If 4- $d_L < d < 4 - d_L$ the test is again inconclusive.

(Gujarati, 1995)

b) Autocorrelation plots of the residuals

A plot of the Autocorrelation Function, of the residuals $\{u_t\}$ against the lag is often performed. If there is no auto-correlation then the Autocorrelation Function coefficients should lie within the 95% confidence band $\pm\frac{1.96}{\sqrt{n}}$

Where n, is the sample size. If outside the 95% confidence band there is auto-correlation of some sort. The advantage of this graphical test is that it applies not only to first order auto-correlation, but also to all forms of auto-correlation (Dielman, 1991).

3.7 Presence of Heteroscedasticity

If analysis of residuals against the fitted values shows that the assumption of constant variance, a property called homoscedastic is not true. Unequal variances for different setting of the independent variable(s) is said to be heteroscedastic. Weighted regression, autoregressive moving average (ARMA) error models and other models can be used to correct for the effects heteroscedasticity.

3.7.1 Stabilizing the variance

In this work, the variance of the error terms is stabilized in order to satisfy the standard regression assumption of homoscedasticity using:

a) Weighted least square regression.

b) Simple least square with ARMA error terms.

3.7.1.1 Weighted least square regression

The least square criterion weighs each observation equally in determining the estimates of the parameters. The procedure treats all of the data equally giving less precise measured points more influence than they should have and gives highly precise points too little influence. The weighted least squares weighs some observations more heavily than others giving each data point its proper amount of influence over the parameter estimates, and this maximizes the efficiency of

parameter estimation. Weighted least square reflects the behaviour of the random errors in the model. To find the parameters of the weighted least square method we minimize:

$$WSSE = \sum_{t=1}^{T} w_t (y_t - \hat{y}_t)^2$$

$$= \sum_{t=1}^{T} w_t \, (y_t - \hat{\beta}_0 - \hat{\beta}_1 x_{t-1})^2$$

where w_t is the weight assigned to the t^{th} observation. The weight w_t in this case is taken as the reciprocal of the variance of that observation's error term, σ_t^2 , i.e

$$w_t = \frac{1}{\sigma_t^2}$$

Observations with larger error variances will receive less weight (and hence have less influence on the analysis) than observations with smaller error variances.

Parameter estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ for the model $y_t = \beta_0 + \beta_1 x_{t-1} + \varepsilon_t$ are derived as:

$$WSSE = \sum_{t=1}^{n} w_t (y_t - \hat{\beta}_0 - \hat{\beta}_1 x_{t-1})^2$$

$$\frac{\partial WSSE}{\partial \hat{\beta}_0} = -2 \sum_{t=1}^{n} w_t \left(y_t - \hat{\beta}_0 - \hat{\beta}_1 x_{t-1}\right) = 0$$

$$-\sum_{t=1}^{n} y_t w_t + \hat{\beta}_0 \sum_{t=1}^{n} w_t + \hat{\beta}_1 \sum_{t=1}^{n} x_{t-1} w_t = 0$$

$$\hat{\beta}_0 = \frac{\sum_{t=1}^{n} y_t w_t}{\sum_{t=1}^{n} w_t} - \hat{\beta}_1 \frac{\sum_{t=1}^{n} x_{t-1} w_t}{\sum_{t=1}^{n} w_t}$$

and

$$\frac{\partial WSSE}{\partial \hat{\beta}_1} = -2 \sum_{t=1}^{n} x_{t-1} w_t \left(y_t - \hat{\beta}_0 - \hat{\beta}_1 x_{t-1}\right) = 0$$

$$-\sum_{t=1}^{n} y_t x_{t-1} w_t + \hat{\beta}_0 \sum_{t=1}^{n} x_{t-1} w_t + \hat{\beta}_1 \sum_{t=1}^{n} x_{t-1}^2 w_t = 0$$

substituting $\hat{\beta}_0$

$$\hat{\beta}_1 = \frac{\sum_{t=1}^{n} y_t x_{t-1} w_t - \frac{(\sum_{t=1}^{n} y_t w_t)(\sum_{t=1}^{n} x_{t-1} w_t)}{\sum_{t=1}^{n} w_t}}{\sum_{t=1}^{n} x_{t-1}^2 w_t - \frac{(\sum_{t=1}^{n} x_{t-1} w_t)^2}{\sum_{t=1}^{n} w_t}}$$

Derived from: $WSSE = \sum_{t=1}^{n} w_t (y_t - \hat{y}_t)^2$, advanced by Mendenhall and Sincich (1989)

The biggest disadvantage of weighted least squares is the fact that the theory behind this method is based on the assumption that the weights are known exactly. This is almost never the case in real applications, instead estimated weights are used (Carrol and Ruppert, 1988).

3.7.1.2 Simple least square regression with ARMA error terms

For the simple least square regression model: $y_t = \beta_0 + \beta_1 x_{t-1} + a_t$, an alternative approach to stabilize the variance of a_t, the error term, can be done by adding a moving average term. The series a_t of the error term can also be expressed in terms of random errors of its past values, which is then a moving average MA(q) model,

where, $a_t = c - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \ldots\ldots\ldots - \theta_q \varepsilon_{t-q+} + \varepsilon_t$ where $\theta_j (j = 1,2 \ldots\ldots q)$ are the weights for the moving average terms. a_t is assumed to be Gaussian white noise, which are independent, identically distributed random variables with mean of zero and constant variance $\varepsilon_t \sim N(0; \sigma^2)$ for all t and $-1 < \theta_j < 1$ (Macdonald and Mackinnon, 2001).

3.8 BINARY DATA

To predict the probability of drought in a given year, a year can be declared a meteorological drought if annual rainfall is less than 75 percent of the normal rainfall. Presence or absence of drought can be viewed as a Bernoulli trial, where success trial is the presence of drought and failure trial is the absence of drought. Thus a year is either classified as a drought or not a drought year. The Department of Meteorological Services in Zimbabwe defines drought in terms of annual rainfall less than 75 percent of 630mm. Rainfall data will be categorized into drought year and non-drought year. Mean rainfall amount less than 473mm will be categorized as a drought, indexed 1, and mean rainfall more than 473mm as no drought, indexed 0. Thus drought index can be treated as binary data. A binary response variable arises by classification as success (drought) when a quantitative test observation falls outside specifications limits (<473mm). The specifications limits were eventually adjusted to < 540mm in line with other reported drought years from other researches on drought patterns in Zimbabwe. (CSO Environmental Report, October 2004).

The binary response random variable is defined as:

$$D_t = \begin{cases} 1 & if\ the\ mean\ annual\ rainfall\ is\ less\ than\ 540mm \\ 0 & Otherwise \end{cases}$$

We view d as a realization of a random variable D_t that can take the value one and zero with probabilities π_t and $1-\pi_t$ respectively. The distribution of D_t is called a Bernoulli distribution with parameter π_t and can be written as:

$$Pr\{D_t = \mathrm{d_t}\} = \pi_t^{d_t}(1-\pi_t)^{1-d_t}$$

for $d_t = 0,1$.

and if there are T such realisation of the random variables $d_1 \ldots\ldots\ldots d_T$ which are independent, with $Prob(d_t = 1) = \pi_t$, then their joint probability is

$$\prod_{t=1}^{T} \pi_t^{d_t}(1-\pi_t)^{1-d_t} = exp\left[\sum_{t=1}^{T} d_t\ log\left(\frac{\pi_t}{1-\pi_t}\right) + \sum_{t=1}^{T} log(1-\pi_t)\right]$$

The expected value and variance of D_t is

$$E(D_t) = \mu_t = \pi_t \text{ and}$$

$$var(D_t) = \sigma_t^2 = \pi_t(1-\pi_t)$$

The mean and variance depend on the underlying probability π_t. Any factor that affects the probability will alter the mean and variance of the observations. This suggests that a linear model that allows the predictors to affect the mean but assumes that the variance is constant will not be adequate for the analysis of binary data.

When all the π_t's are equal we can define:

$$D = \sum_{t=0}^{T} d_t$$

so that D represents the number of successes in T "$trials$", then the distribution of D is binomial with parameters π and T.

$$D \sim Bin(T; \pi)$$

The probability distribution function of D is given by:

$$\Pr(D = d) = \binom{T}{d} \pi^d (1-\pi)^{T-d} \quad , d = 0,1 \dots \dots \dots . T.$$

The expected value and variance of D are

$$E(D) = \mu = T\pi \quad \text{and}$$

$$var(D) = \sigma^2 = T\pi(1-\pi)$$

3.8.1 Models for binary responses

To investigate the relationship between the response probability π_t and the covariate $x_t^{/}$, it is convenient to construct a formal model capable of describing the effect on π_t of changes in $x_t^{/}$.

The model embodies assumptions such as:

a) zero correlation.
b) lack of interaction and
c) linearity of residuals.

Suppose therefore that the dependence of x_t on $x_t^{/}$ occurs through the linear combination

$$\eta = \sum_{j=0}^{p} x_j \beta_j$$

for unknown coefficients $\beta_0 \dots \dots \dots \dots . \beta_p$ and the probabilities of observing a value of one is modeled as:

$$\Pr(d_t = 1 | x_t, \beta) = 1 - F(-x_t^{/}\beta)$$

Where $\beta = [\beta_0, \dots \dots . , \beta_p]^{/}$ and $x_t^{/} = [1\ x_1 \dots \dots \dots \dots . . x_p]$

Where F is a continuous, strictly increasing function that takes a real value and returns a value ranging from zero to one and F determines the type of binary model. It follows that;

$$\Pr\left(d_t = 0 \middle| x_t^{/}\beta\right) = F(-x_t^{/}\beta)$$

The specifications of the F function yields a Logit model (logistic function), Probit model (inverse Normal function) or Extreme value model(complementary log-log function), all with a systematic part:

$$g(\pi_t) = \eta_t = \sum_{j=1}^{p} x_{tj}\beta_j \quad \text{t} = 1 \dots \dots \dots \dots T$$

This systematic part is referred to as the link function. All the link functions are continuous and increasing on (0;1).

The logistic and the probit function are almost linearly related over the interval $0.1 \leq \pi_t \leq 0.9$. For this reason, it is usually difficult to distinguish between these two functions on the grounds of goodness of fit for small values of π_t, the complementary log-log function is close to the logistic, both being close to $\log \pi_t$. As π_t approaches 1, the complementary log-log function approaches infinity much slower than either the logistic or the probit function.

3.8.1.2 Logistic Regression Model with one independent variable and a lag of 1

The logit of the underlying probability π_t is a linear function of the predictors

$$logit(\pi_t) = x_t^{/}\beta$$

Where x_t is a vector of covariates i.e. $x_t^{/} = [\,1\; x_{t-1}\,]$ and β is the vector of regression coefficients and the pdf of logistic distribution for one independent variable is:

$$E(D_t) = \frac{\exp\,(\beta_0+\beta_1 x_{t-1})}{1+\exp\,(\beta_0+\beta_1 x_{t-1})}$$

equivalently the model may be written in terms of the odds of a positive response giving

$$\frac{\pi_t}{1-\pi_t} = \exp\,(\beta_0 + \beta_1 x_{t-1})$$

thus the logit link function is

$$\log\left(\frac{\pi_t}{1-\pi_t}\right) = \beta_0 + \beta_1 x_{t-1}$$

If we set

$$D_t^{\,*} = log\left(\frac{\pi_t}{1-\pi_t}\right)$$

The transformed logistic model

$$D_t^{\,*} = \hat{\beta}_0 + \hat{\beta}_1 x_{t-1}$$

is now linear in the $\beta' s$ and the model of least squares can be applied. Applying the method of least squares directly to the binary observations has limitations:

a) The linear transformation fails to capture the true probability π_t which remains unknown. It assumes that the probability increases linearly with the regressor, and also the model does not guarantee that the conditional probability will occur between zero and one (Gujarati, 1995), since the linear predictor $\beta_0 + \beta_1 x_{t-1}$ can take any real value so there is no guarantee that the predicted values will be in the correct range unless complex restrictions are imposed on the coefficients.

b) Since D_t takes only the values 0 and 1 $D_t^2 = D_t$ and $var(D_t) = \pi_t(1-\pi_t)$ which is non constant violating the assumption for least square regression except when $\pi_t = \pi$.

Thus the model parameters are estimated using standard maximum likelihood procedure.

3.8.1.2.3 Maximum likelihood estimation

In the case of linear logistic models, we have

$$g(\pi_t) = \eta_t = log\{\pi_t/(1-\pi_t)\} = \sum_j^p x_{t-1,j}\beta_j \qquad 3.8.1$$

and the log likelihood of a binomial distribution may be written in the form

$$l(\pi; d) = \sum_{t=1}^{T}\left[d_t log\left(\frac{\pi_t}{1-\pi_t}\right) + m_t \log(1-\pi_t)\right] \qquad 3.8.2$$

where $m_t = T - d_t$

substituting 3.8.1 into 3.8.2 gives

$$l(\beta; d) = \sum_{t}^{T}\sum_{j}^{p} d_t x_{t-1,j}\beta_j - \sum_{t}^{T} m_t log\left(1 + exp\sum_{j}^{p} x_{t-1,j}\beta_j\right)$$

we now derive the likelihood equations of the parameters β that appear in (3.8.1). First the derivative of log-likelihood function of the general binomial function with respect to π_t is

$$\frac{\partial l}{\partial \pi_t} = \frac{d_t - m_t\pi_t}{\pi_t(1-\pi_t)}$$

using the chains rule, the derivative with respect to β_r is

$$\frac{\partial l}{\partial \beta_r} = \sum_{t=1}^{T} \frac{d_t - m_t\pi_t}{\pi_t(1-\pi_t)} \frac{\partial \pi_t}{\partial \beta_r}$$

and setting the derivative to zero estimates of β_r are obtained (Dobson, 1990).

3.8.1.3 Probit Model

The probit model is a specification model for a binary response model which employs a probit link function:

$$g(\pi_t) = \eta_t = \Phi^{-1}(\pi_t) = \beta_0 + \beta_1 x_{t-1}$$

where $\beta_1 = \frac{-\mu}{\sigma}$ and $\beta_2 = \frac{1}{\sigma}$, and the link function is ϕ^{-1}, the inverse of the cdf of the standard Normal distribution. Parameters β are estimated by maximum likelihood estimation procedure. The Probit model assumes that the random errors in the model are independent and identically distributed with a mean of zero (Dueker, 1997), while for many time series applications this is not a plausible assumption. According to Estrella and Mishkin (1998), the probit model has an overlapping data problem such that the forecast errors are likely to be serially correlated. This raises the possibility that tests of significance of the variables using conventional test statistics may provide meaningless results. The problem can be corrected by a method proposed by Dueker (1997). Dueker observes that adding a lag of the dependent variable increases the validity of the assumption that error terms has a mean of zero, conditional on availability of

information over time $t-k$. The new model proposed by Dueker in the case of modelling drought using the probit link function would then be:

$$\widehat{D}_t = \hat{\beta}_0 + \hat{\beta}_1 x_{t-\mathrm{k}} + \hat{\beta}_2 \widehat{D}_{t-\mathrm{k}}$$

where estimates of β_0, β_1 and β_2 are the maximum likelihood estimates.

3.8.2 Binary model validation

Deviance and log – likelihood statistic can be used to test goodness of fit of the models. The thumb rule is $Deviance > T - p$ where T is the number of observations and p is the number of parameters to be estimated. Deviance value is used to test the goodness of fit of the Binary models.

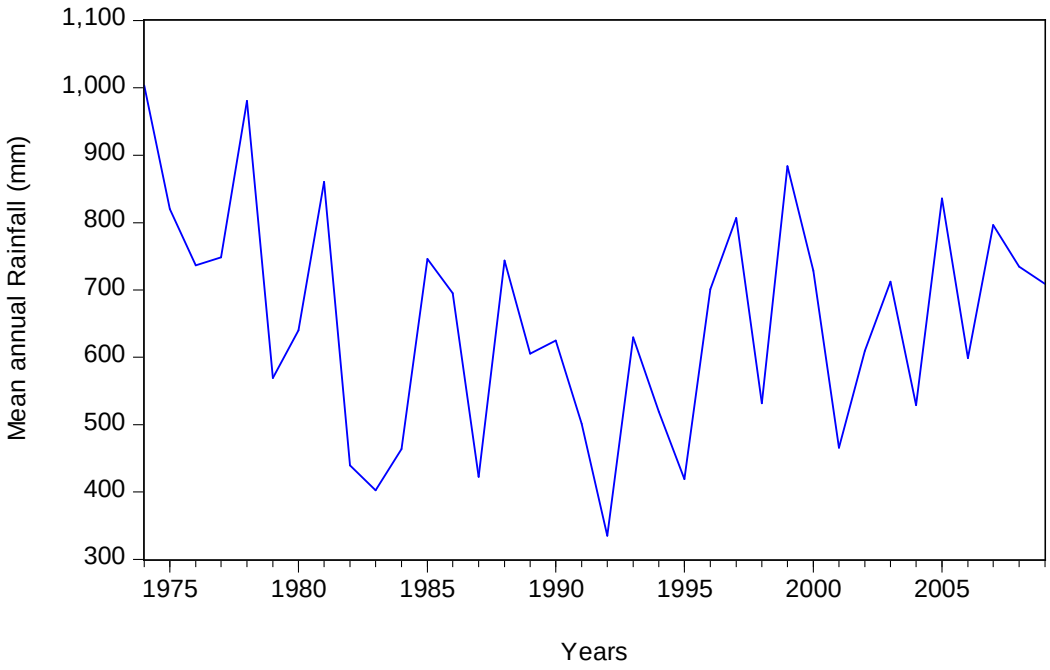
CHAPTER 4

ANALYSIS

4.1 Introduction

The chapter presents the results from the ordinary regression analysis, weighted regression and binary data analysis using the logistic and probit regression models.

4.2 The Zimbabwean Mean Annual Rainfall Patterns and Southern Oscillation



SOURCE: Zimbabwe Department of Meteorological Services

Fig 4.1 The Zimbabwean Mean Annual Rainfall for the period 1974 to 2009

Fig 4.1 shows the time series of the mean annual rainfall for Zimbabwe from 1974 to 2009. The highest rainfall was received in 1974, while the lowest rainfall was received in 1992 (the worst drought in the given history of the country). Rainfall less than 473mm is categorized by the Department of Meteorological Services as a meteorological drought.

The highest correlation between SOI and the Zimbabwean mean annual rainfall using the 1974 to 2009 rainfall data is +0.45 using the current April and January SOI value. This is in agreement with previous researches for the country. Makarau and Jury (1997), for example, used a 41 year period rainfall record and obtained the highest correlation between SOI and the Zimbabwean summer rainfall of +0.44 using the August to October mean SOI value. The lowest correlation of +0.23 was obtained between mean annual rainfall with the February SOI value. September SOI value has a +0.33 correlation with the rainfall, which is again in agreement with Matarira (1990). Matarira studied the relationship between SOI and area average rainfall over the south eastern parts of Zimbabwe and found correlations of +0.3 to +0.42 at lead times of 1 to 4 months. However, the focus of this study is to determine a particular month's SOI which has a high correlation with mean annual rainfall at **a lead time of a year or more**. This is a clear departure from other researches. Using a lag of one year, the highest correlation (+0.36) between the SOI and the Zimbabwean mean annual rainfall is with SOI for September. At a lag of more than a year the correlations between the SOI and annual rainfall are insignificant.

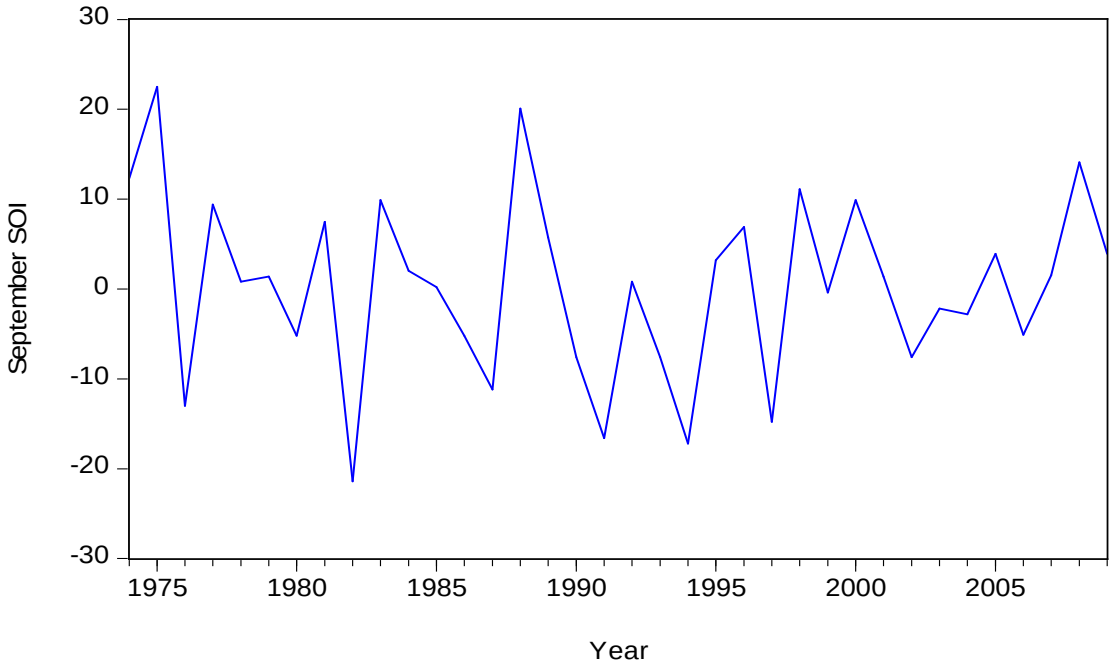


Fig 4.2 The Southern Oscillation Index for September 1974 to 2009

4.3 Relationship between SOI and Mean Annual Rainfall

Table: 4.1 Simple Regression Models

$Variable$	$coefficient$	$p\ value$	AIC	BIC
$SOI_{january}$	$\hat{\beta}_0 = 651.1446$ $\hat{\beta}_1 = 6.740371$	0.0000 0.0054	12.91146	12.99943
SOI_{april}	$\hat{\beta}_0 = 673.8462$ $\hat{\beta}_1 = 5.775406$	0.0000 0.0057	12.91420	13.00217
$SOI_{sept(-1)}$	$\hat{\beta}_0 = 643.1394$ $\hat{\beta}_1 = 5.525738$	0.0000 0.0311	12.89074	12.97942

Table 4.1 shows three possible models for predicting the Zimbabwean mean annual rainfall patterns. The model with the least AIC and BIC value was selected as the best model. The best linear regression model is the $\hat{y}_t = \hat{\beta}_0 + \hat{\beta}_1 x_{t-1}$ ($p\ value = 0.0311\ for\ SOI\ variable$), where y_t is the predicted annual rainfall and x_{t-1} is the September SOI for the previous year. The estimates of β_0 and β_1 are 643.1394 and 5.525738 respectively. The ACF and PACF correlogram (see appendix 2) show that the residuals are not correlated. The Durbin – Watson statistic of 2.1748 indicates that the residuals are independent thus uncorrelated. The model suggest that if September SOI of the previous year is zero, the mean annual rainfall is 643.1394 mm. The predicted rainfall increases by 5.525738 mm for a unit increase in the September SOI of the previous year.

4.3.1 Checking Model Assumptions

Testing for Normality of residuals

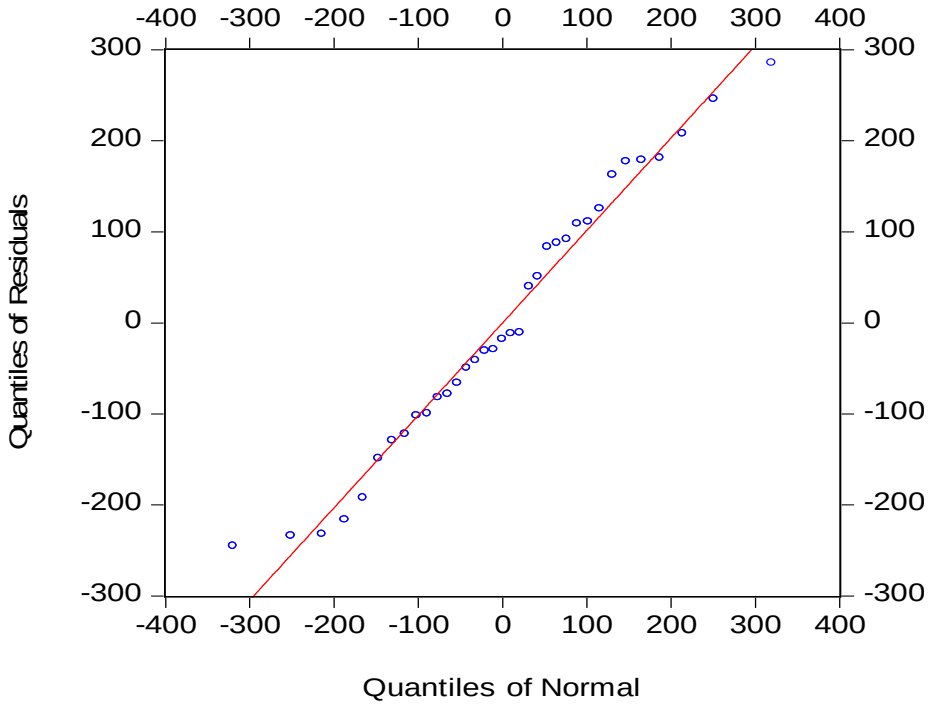


Figure 4.3 A Normal Q – Q plot of residuals for Simple Regression Model

The normal probability plot of the residuals showed an almost straight line suggesting that the residuals are normally distributed. The model thus does not violate the normality assumptions.

Testing for Constant variance

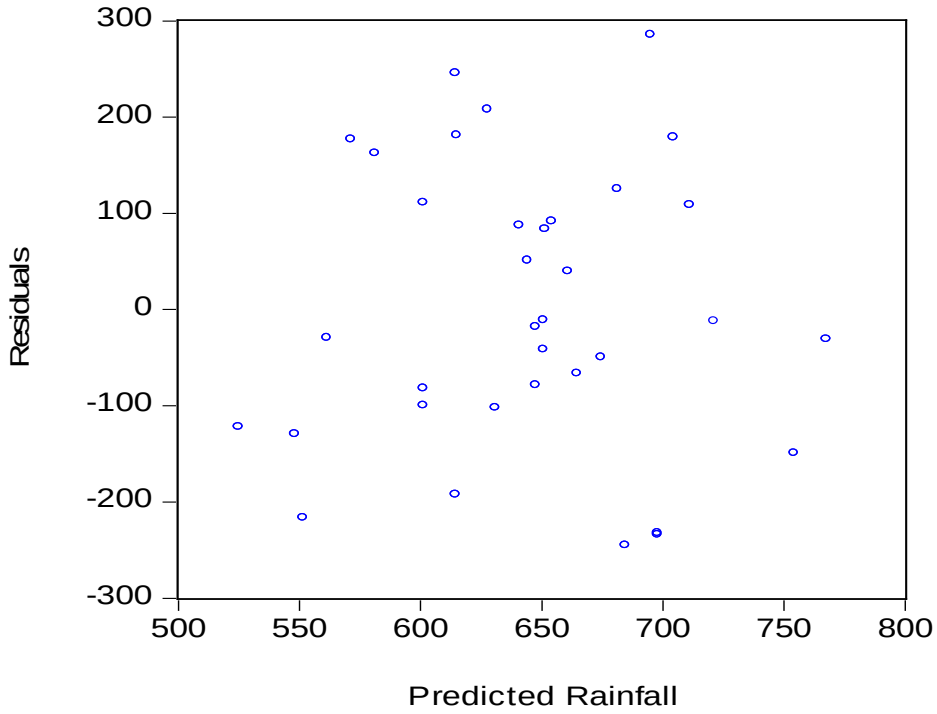


Figure 4.4 Residuals versus Predicted values for Simple Regression Model

Figure 4.4, a plot of residuals against predicted values indicates a cluster suggesting that the model violates the assumption of constant variance. This means the model can be improved by stabilizing the variance. The variance was stabilized using the weighted least square method. Simple regression with moving average error terms was also used to capture variability in the simple regression model.

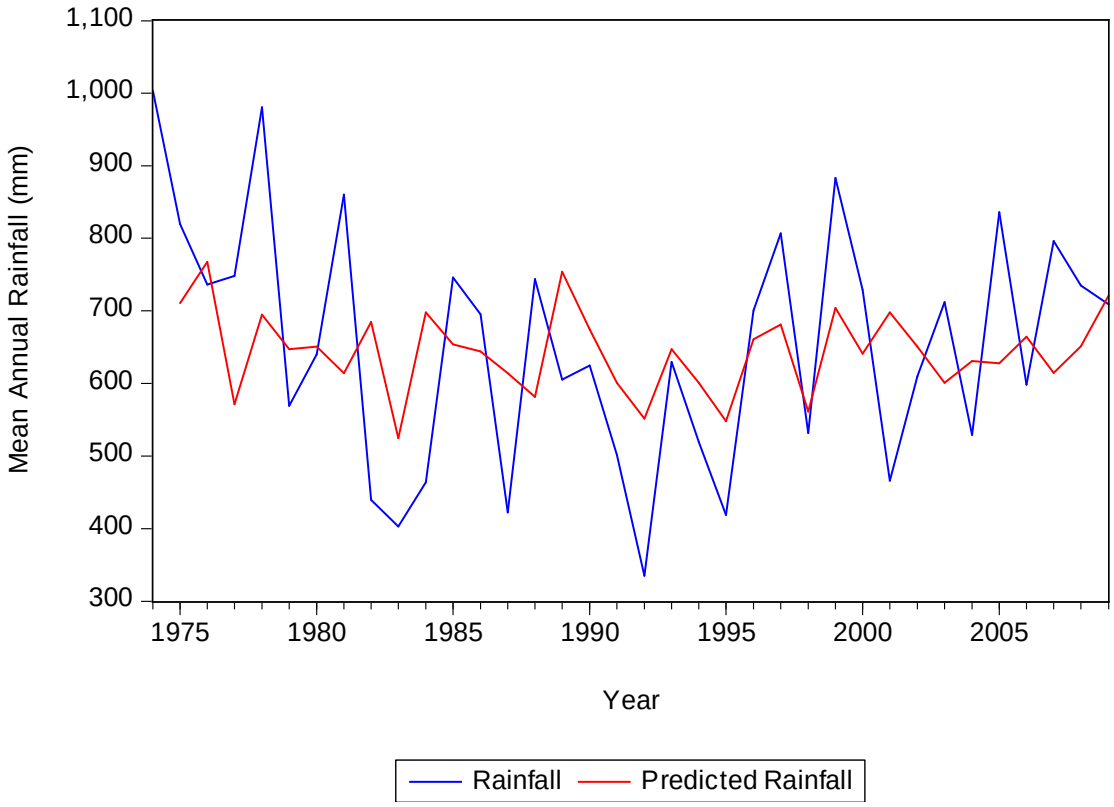


Figure 4.5 Rainfall versus Rainfall forecast (Simple Regression Model)

Figure 4.5 clearly shows that the model fails to capture the variance in the observed values. Thus the model needs to be improved to capture the variability.

4.4 Weighted Regression Model

Weighted regression was done to capture variability the observed values.

Table: 4.2 Weighted Regression models

$Weight$	$Coefficient$	$p\ value$	AIC	BIC
SOI_{sept}	$\hat{\beta}_0$ =672.3622 $\hat{\beta}_1$ =10.87731	0.0000 0.0182	13.14399	13.24356
$\frac{1}{SOI_{sept}}$	$\hat{\beta}_0$ =708.7970 $\hat{\beta}_1$ =5.483862	0.0000 0.0367	12.62942	12.72900
SOI_{sept}^2	$\hat{\hat{\beta}}_0$ =701.0582 $\hat{\beta}_1$ =16.76346	0.0000 0.0000	11.59687	11.68584

Table 4.2 shows the weighted linear regression models for predicting rainfall y_t, with SOI for September of the previous year as an independent variable. The model with SOI_{sept}^2 as the variance stabilizing weight was selected, it is the model with the least AIC and BIC. The model is significant ($p\ value = 0.0000$) for both parameters. The model is:

$$\hat{y}_t^* = \hat{\beta}_0 + \hat{\beta}_1 x_{t-1}$$

where x_{t-1} is SOI for September of the previous year and the estimates of β_0 and β_1 are 701.0582 and 16.76346 respectively. The residuals are independent and uncorrelated (see ACF and PACF correlogram in Appendix 3).

Checking Model Assumptions

Testing for Normality of residuals

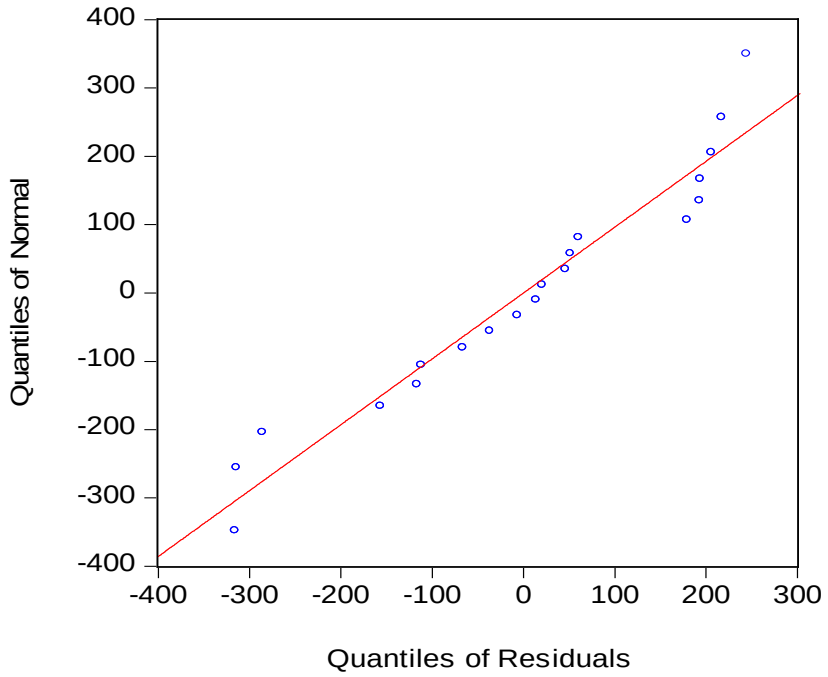


Figure 4.6 A Normal Q – Q plot of residuals for Weighted Regression Model

The normal probability Q – Q plot of residuals is linear suggesting that the residual are almost normally distributed. The model thus does not violate the assumption of normality.

Testing for constant variance

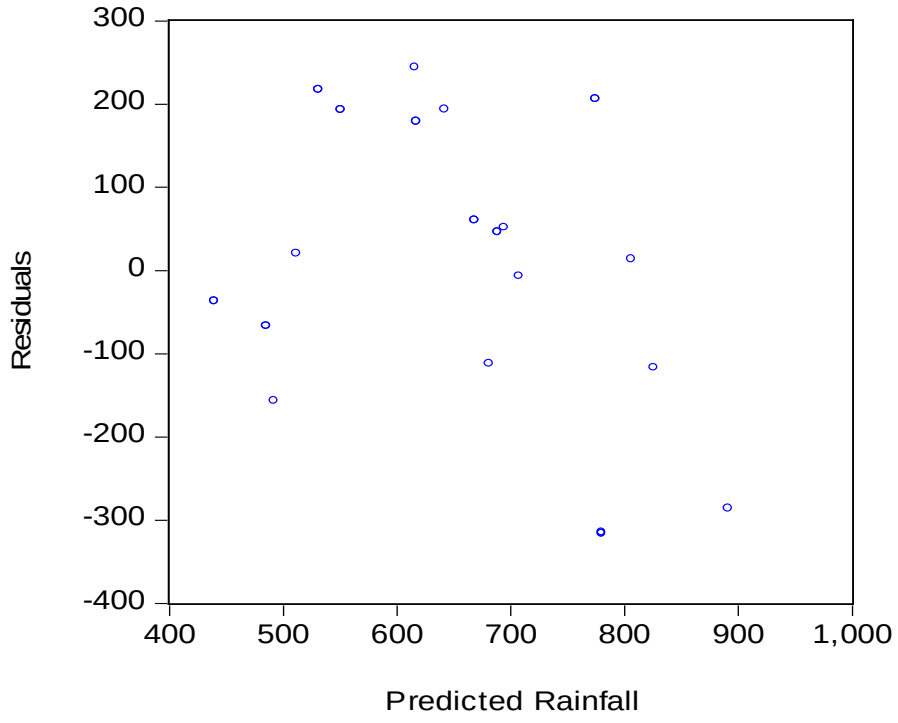


Figure 4.7 Residuals verses Predicted values

Figure 4.7 shows that the model almost has a constant variance thus it does not violates the model assumptions grossly. The model can be selected to forecast rainfall.

Rainfall versus Predicted Rainfall

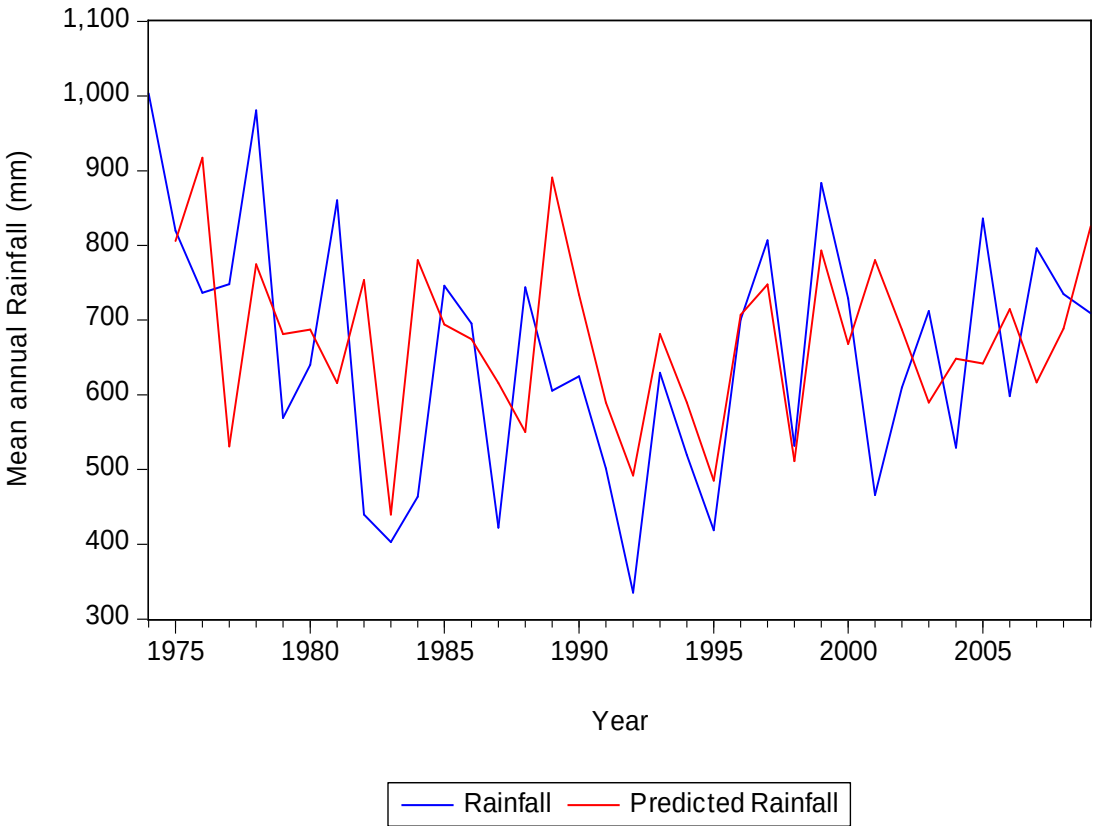


Figure 4.8 Rainfall against Forecasted Rainfall

Figure 4.8 shows that the weighted least square model has captured a lot of the variation. There is significant improvement in the forecasting power of this model.

4.5 Regression Model with MA error term

The simple regression model failed to capture variance of the observed values. Apart from stabilizing variance using the weighted least square method, this research stabilize, the variance by adding an MA error term to the simple regression model.

Table: 4.3

Variance stabilizer	***Coefficient***	***p value***	***AIC***	***BIC***
$MA(8)$	$\hat{\beta}_0 = 644.3668$ $\hat{\beta}_1 = 6.3329$	0.0000 0.0020	12.16043	12.29375
$MA(16)$	$\hat{\beta}_0 = 642.8118$ $\hat{\beta}_1 = 6.3235$	0.0000 0.0091	11.78474	11.91806
$MA(18)$	$\hat{\beta}_0 = 661.8678$ $\hat{\beta}_1 = 5.8676$	0.0000 0.0248	11.61895	11.75226

Table 4.3 Shows the linear regression models of predicting rainfall y_t with $x_{t-1,}$ SOI for September of the previous year as an explanatory variable stabilized with moving average error term. The model with MA(18) as a variance stabilizing agent was selected, since it has the least AIC and BIC. The model is:

$$y_t = \hat{\beta}_0 + \hat{\beta}_1 x_{t-1} + a_t$$

where a_t is a series of the residuals. $a_t = \varepsilon_t - \theta\varepsilon_{t-18}$, the estimates of β_0 and β_1 are 661.8678 and 5.867551 respectively and $\theta = -0.916365$

The expected value of y_t is then given by:

$$E(y_t) = \hat{\beta}_0 + \hat{\beta}_1 x_{t-1}$$

since $E(a_t) = 0$

The model is significant $(p\ value = 0.0248 < 0.05)$ and the moving average term is significant $(p\ value = 0.000 < 0.05)$. The residuals of the model are uncorrelated with a Durbin – Watson statistic of 2.46. Appendix 4 shows the ACF and PACF correlogram of the model.

Checking Model Assumptions

Testing for Normality Assumption

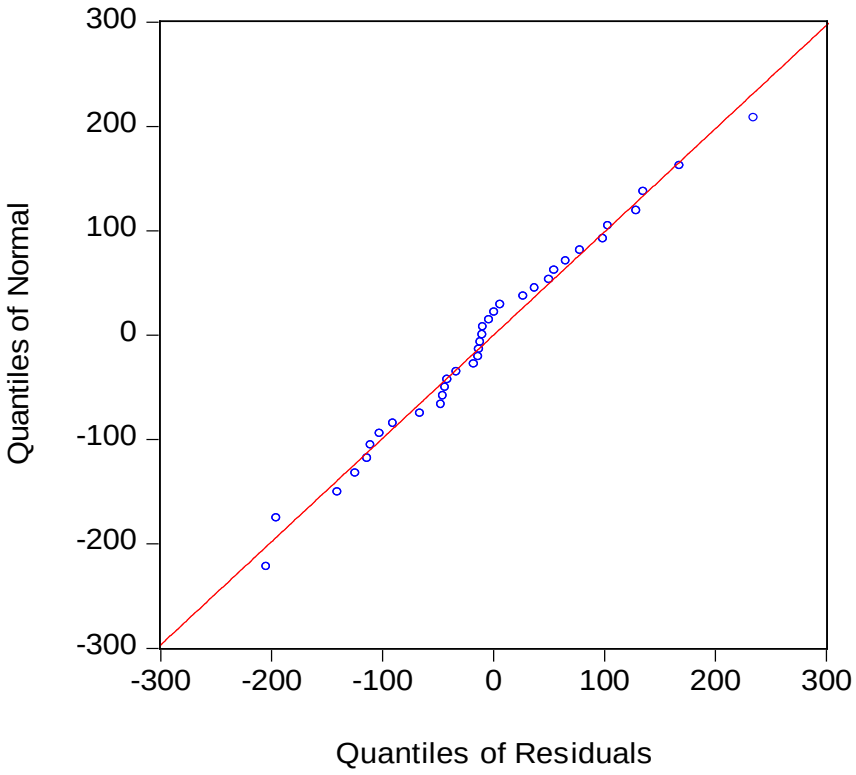


Figure 4.9 A normal Q – Q plot for Simple Regression model with MA errors

The normal probability plot of residuals is straight, the residuals are normally distributed. The model does not violate the assumption of normality.

Testing constant variance assumption

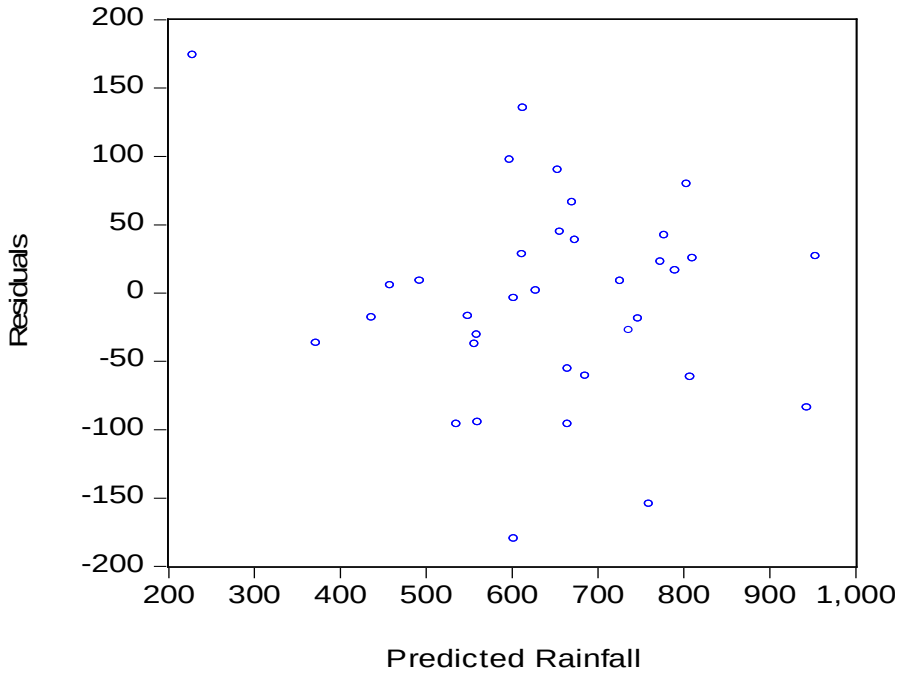


Figure 4.10 Residuals versus Predicted values

Figure 4.10 shows that the residuals can be assumed to have constant variance, the scatter plot does not show any pattern although there are some few residuals which are still very big. The model however does not grossly violate the assumption of constant variance, thus the model can be selected for forecasting.

Checking model forecasting power

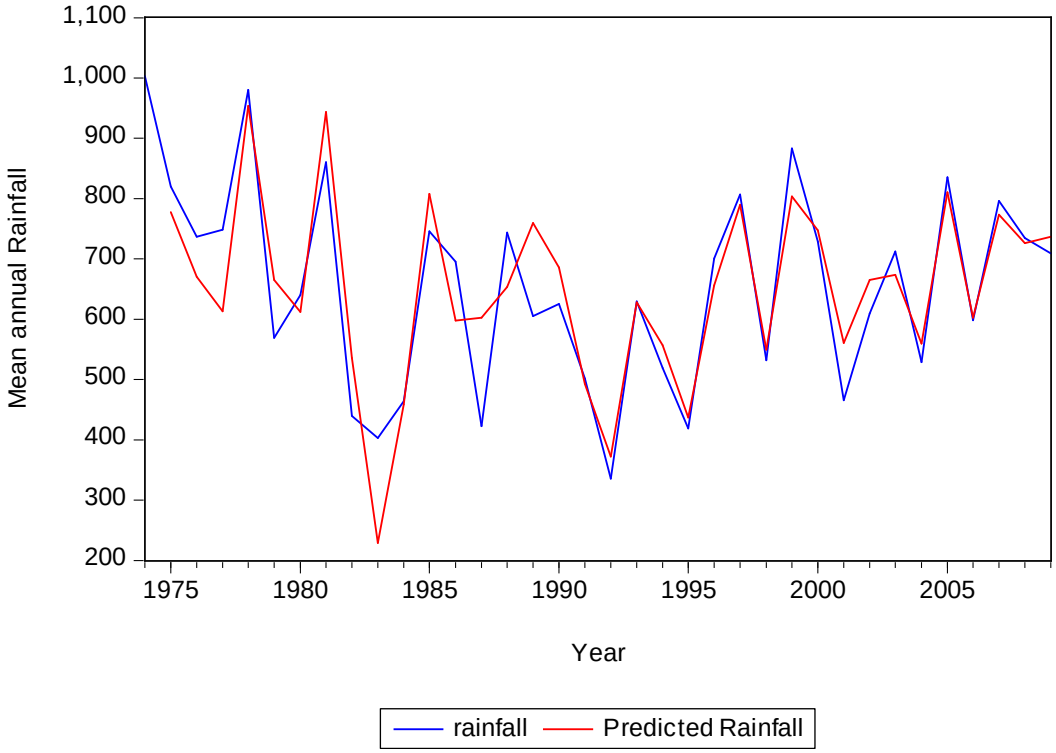


Figure 4.11 Rainfall verses Forecasted Rainfall

Figure 4.11 shows that the simple linear regression model with the MA(18) error term has a high forecasting power and can be selected for forecasting rainfall in Zimbabwe using SOI for September of the previous year.

4.6 Drought Models

The binary drought data was modelled using the Logistic and Probit regression.

Table: 4.4 Binary Models (473 mm threshold)

$\boldsymbol{Variable}$	$\boldsymbol{Link\ function}$	$\boldsymbol{Coefficient}$	$\boldsymbol{p\ value}$	$\boldsymbol{Deviance}$
$SOI_{sept(-1)}$	$probit$	$\hat{\beta}_0 = -0.876217$ $\hat{\beta}_1 = -0.032764$	0.0005 0.1801	33.11136
$SOI_{sept(-1)}$	$Logit$	$\hat{\beta}_0 = -1.490361$ $\hat{\beta}_1 = -0.032764$	0.0012 0.1663	32.94040

Table 4.4 shows that both models for predicting drought (with threshold of 473mm) do not fit the data so well since deviance is almost equal to degrees of freedom ($Deviances \leq degrees\ of\ freedom = 33$) and the models are not significant ($P\ values > 0.05$). Deviance of a good Binary model must be greater than $T - p$ degrees of freedom, where T is different covariates and p is the number of parameters to be estimated. The drought years were adjusted in line with other recorded research drought data (CSO Environmental Report, October 2004).

Table: 4.5 Binary models for adjusted data (540mm threshold)

$Variable$	$Link\ function$	$Coefficient$	$p\ value$	$Deviance$
$SOI_{sept(-1)}$	$Probit$	$\hat{\beta}_0 = -0.446515$ $\hat{\beta}_1 = -0.056530$	0.0554 0.0251	39.21368
$SOI_{sept(-1)}$	$Logit$	$\hat{\beta}_0 = -0.759067$ $\hat{\beta}_1 = -0.094721$	0.0577 0.0305	39.15936
$SOI_{sept(-1)}$	$Probit$	$\hat{\beta}_1 = -0.053708$	0.0282	43.02260
$SOI_{sept(-1)}$	$Logit$	$\hat{\beta}_1 = -0.085903$	0.0363	43.10996
$SOI_{sept(-1)}$ $Drought_{(-1)}$	$Logit$	$\hat{\beta}_1 = -0.102287$ $\hat{\beta}_2 = -1.177757$	0.0180 0.1086	40.13624

Table 4.5 shows the binary models for drought after adjusting the rainfall threshold figures in line with drought years in other researches (CSO Environmental report, October 2004). Annual rainfall below 540mm was categorized as a drought, indexed 1, and rainfall above 540mm indexed 0. For all the models:

$(Deviance > T - p = 33\ degrees\ of\ freedom)$

The deviances are greater than degrees of freedom, thus, all the models fits the data well. The best model to be selected is the model with the largest deviance. The model is:

$$D_t = \hat{\beta}_1 x_{t-1}$$

with a logit link function, where $D_t = \log\left(\frac{\pi_t}{1-\pi_t}\right)$ and x_{t-1} is the SOI for September of the previous year. The estimate of the parameter β_1 is -0.085903. If the SOI for September of the previous year is negative, the model predicts a drought. If the SOI for September of the previous year is positive, the model predicts no drought. The simple logistic model, as pointed out by

Dueker (1997), has one deficiency of assuming that the error terms are independent and identically distributed with a mean of zero.

The modified logistic model proposed by Dueker (1997) which corrects serial correlation of residuals in the simple logistic model is:

$$D_t = \hat{\beta}_1 x_{t-1} + \hat{\beta}_2 D_{t-1}$$

However, it is not significant, because slope parameter for d_{t-1} is marginally not significant ($p\ value = 0.1086 > 0.05$).

Checking model forecasting power

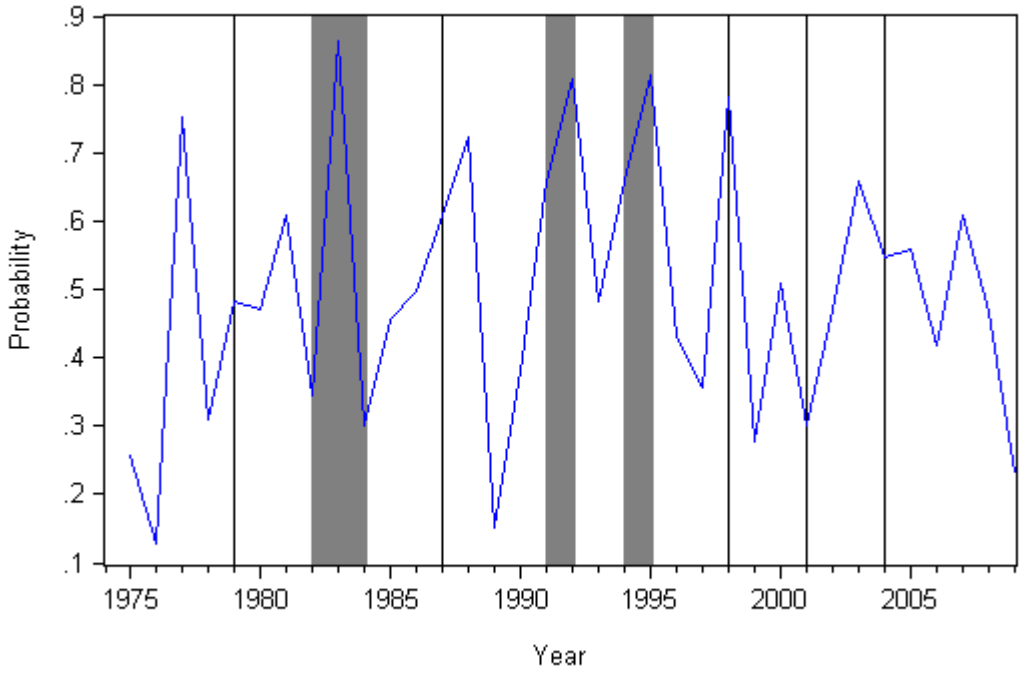


Figure: 4.12 Drought verses Predicted Drought probability

Figure 4.12 shows the graph of drought years in grey lines or grey shading verses predicted drought probability, indicating that the model can be used to predict drought in Zimbabwe. Although the model proposed by Dueker is marginally insignificant, figure 4.13 shows the graph

of predicted drought using the simple logistic model verses predicted drought probability using the model:

$$D_t = \hat{\beta}_1 x_{t-1} + \hat{\beta}_2 D_{t-1}$$

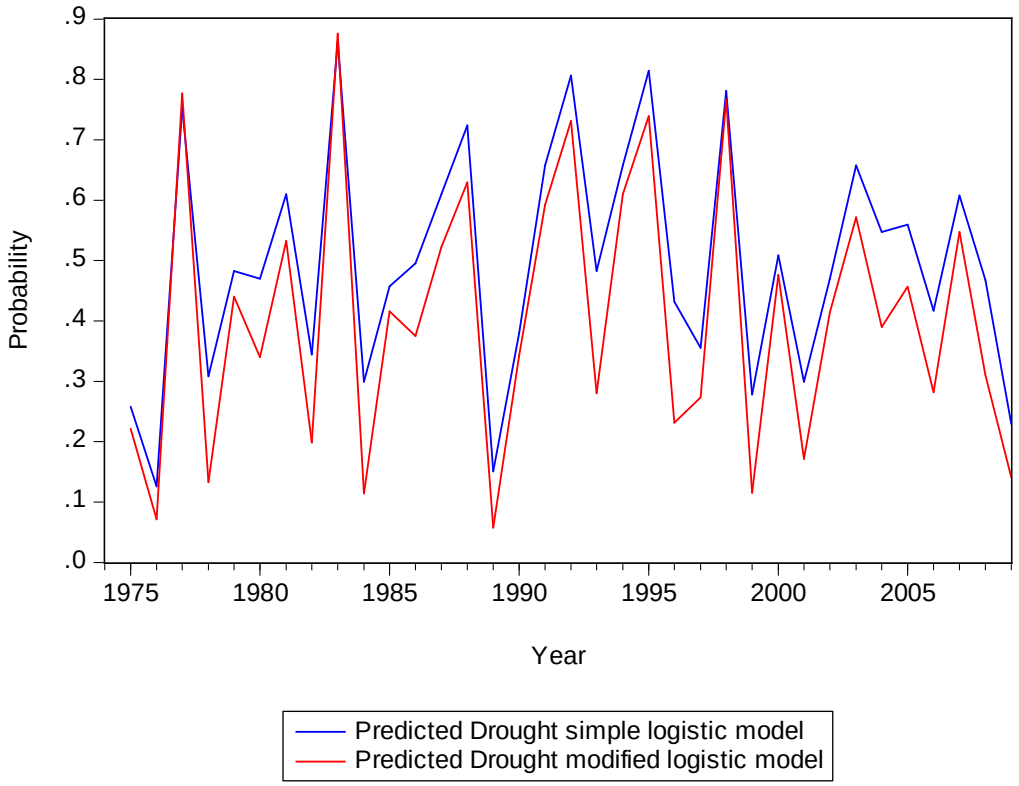


Figure 4.13 Predicted Drought Probability: Simple model verses Modified model

Figure 4.13 shows that the prediction of the simple logistic model is almost the same with predictions using the modified logistic model, since the parameter of the model modified is insignificant.

4.7 Forecasting Rainfall and Drought for the year 2011

Using the September SOI of 2010 which is +25 , the predicted annual rainfall for 2011 using the weighted regression model is: $y_{2011} = 701{,}0582 + 16.76386(25) = 1120mm$, thus 2011 should receive a good rainfall. The model with MA(18) error terms predicts 2011 rainfall of: $y_{2011} = 661.8678 + 5.8676(25) = 808.5578mm$, again which is high. The variations in the

predictions of the two models can be explained using the difference in the predicting powers of the models. Thus 2011 promises to be a good agricultural year. Farmers are advised to prepare for a good rainy season in 2011. The logistic model predicts:

$$E(D) = \frac{\exp(-0.085903(25))}{1+\exp(-0.08593(25))}$$

$$E(D) = 0.1045580477$$

a 10 percent chance of a drought in 2011. The logistic model is not predicting a drought in 2011 as alluded by the weighted regression model and the simple linear model with MA error term.

Table 4.6 below show the probability of drought as predicted by the logistic model.

Table: 4.6 Probability of a drought as predicted by the best Logistic model

$\boldsymbol{Prob}$ %	10	20	30	40	50	60	70	80	90
$\boldsymbol{Sept\ SOI}(-\mathbf{1})$	25.6	16.1	9.9	4.7	0	-4.7	-9.9	-16.1	-25.6

In the Table 4.6 the SOI for September of the previous year of 0, predicts a 50 percent chance of drought , while negative September SOI (< -4.7) predicts at least 60 percent chance of having a drought in the year ahead. See appendix A1 for the formula to determine the SOI for September in table 4.6.

Table: 4.7 Probability of drought as predicted by the best Probit model

$\boldsymbol{Prob}$ %	10	20	30	40	50	60	70	80	90
$\boldsymbol{Sept\ SOI}(-\mathbf{1})$	23.9	15.7	9.8	4.7	0	-4.7	-9.8	-15.7	-23.9

In the Table 4.7 the SOI for September of the previous year of 0, predicts also 50 percent probability of drought. The SOI for September of the previous year of -23.9 predicts a 90 percent chance of drought, while September SOI of the previous year of more than 23.9 predicts less than 10 percent chance of drought.

CHAPTER 5

Conclusions and Recommendations

5.1 Introduction

This chapter summaries the research findings and conclusions of the study. It concludes by suggesting recommendations.

5.2 Conclusions

Despite its short observational period, this research has shown that Zimbabwe annual rainfall and drought patterns can be predicted using Southern Oscillation Index. The study has shown that the Zimbabwean mean annual rainfall is highly correlated (+0.45) with January and April SOI. The aim of the research was to establish the relationship between the Zimbabwean mean annual rainfall patterns and the SOI of a particular month at a maximum lag for effective planning purposes. At a lag of one year, the highest correlation of (+0.36) was found between the mean annual rainfall and the SOI for September. Correlations at a lag of more than a year are insignificant.

The weighted regression model using squared SOI for September as a variance stabilizing weight was found to be significant. The model: $y_t = 701.0582 + 16.76346x_{t-1}$, where x_{t-1} is the SOI for September of the previous year was found to capture the variability between observed and predicted values. Using the current SOI for September 2010 of +25, the model predicts an annual rainfall of 1120mm, a figure well above the drought threshold of 540mm. If the SOI value for September of the previous year is less than – 13, then the model predicts an annual rainfall below the drought threshold.

MA(18) was also used to stabilize variance of the simple regression model. The model is $y_t = 661.8678 + 5.867551x_{t-1}$. The model was found to have stabilized the variance between observed and predicted values, thus has good forecasting power. The model predicts annual rainfall of 809mm in the year 2011 using the current SOI for September 2010.

Annual rainfall was categorized into drought year i.e. annual rainfall less than 75 percent of 630mm, indexed 1, and above 473mm as a non drought year, indexed 0. The probit and logit

models were found to be insignificant. The rainfall threshold was adjusted in line with other studies on droughts in Zimbabwe. Annual rainfall below 540mm was categorized as a drought year indexed 1 and index 0, for rainfall above 540mm. The best model was found to be the logistic model: $D_t = -0.085903x_{t-1}$, where d_t is in the interval [0;1]. Using the current SOI for September 2010, the model forecasts for the year 2011 is a low 10 percent probability of a drought. It was found that a 50 percent chance of a drought is predicted, if the SOI for September of the previous year is zero and 90 percent chance of a drought if the SOI for September of the previous year is – 25.6. The best probit model was found to be $D_t = -0.053708x_{t-1}$. Using the probit model, a 50 percent chance of a drought is predicted, if the SOI for September of the previous year is zero. A 10 percent chance of drought is predicted if the SOI value for September of the previous year is 23.9 and a 90 percent chance of drought is predicted, if SOI for September of the previous year is -23.9.

5.3 Recommendations

This study recommends further study in:

a) Developing a model of forecasting the September SOI such the current research models can be used to forecast annual rainfall and drought patterns for Zimbabwe at a lead time of more than a year. Forecasted SOI values could then be used to predict rainfall with a lead of more than 1 year.
b) Apply the Bayesian statistics approach in modelling the relationship between the September SOI and the Zimbabwean annual rainfall and drought patterns. Arguably, the Bayesian approach gives more information than the classical approach.
c) Advancing this research's finding to establish models using annual rainfall patterns per climatic region in the Zimbabwe, since different regions in the country receive different total annual rainfalls.
d) The department of meteorological services in Zimbabwe may need to revise or update its formula for calculating the threshold for declaring a meteorological drought year.

REFERENCES

Agresti, A. (1996). *An Introduction to Categorical Data Analysis.* John Wiley: London.

Bekker, S. J, Pretorius, J. H and de Waal, D. J. (2010). Modelling Inflows into Gariep Dam. *Technical report No. 405, Department of Mathematical Statistics and Actuarial Science; University of the Free State, South Africa.*

Buckle, C. *Weather and Climate in Africa.* Longman: Harlow.

Cane, M. A , Enhel, G and Backland, R. W. (1994). Forecasting Zimbabwean maize yield using eastern equatorial Pacific Sea Surface Temperatures. *Nature:37: 204 – 205*.

Carroll, R. J and Ruppert, D. (1988). *Transformation and Weighting in Regression.* Chapman and Hall: New York.

Chambers, J. William, C. Beat, K and Paul, T (1983). *Graphical Methods for Data Analysis.* Wadsworth: New York.

Cox, R. D. (1970). *The Analysis of Binary Data.* Spoltiswoode, Ballantyne & CO LTD: London.

Dielman, T. E. (1991). *Applied Regression Analysis for Business and Economics.* PWS-KENT: Boston.

Dobson, A. J. (1990*). An Introduction to Generalized Linear Models.* Chapman and Hall: London.

Dueker, M. J. (1997). Strengthening the case for Yield Curve as a Predictor of US Recessions. *Federal Reserve Bank of St Louis Economic Review 79: 41 – 51*

Estrella, A and Hardouvelis, G. (1991). The term structures as a predictor of real economic activity. *The journal of Finance 46: 555 – 570.*

Gujarati, D. (1995). *Basic Econometrics* 4^{th} *Edition.* McGraw-Hill: Boston.

Jarque, C. M and Bera, A. K. (1987). A test of normality of observations and regression residuals. *International Statistical Review, Volume 55: 163 – 172.*

Jury, M. R. (1996). Regional tele-connection pattern associated with summer rainfall over South Africa, Namibia and Zimbabwe. *International Journal of Climatology*.

Khomo, M. M and Aziakpono, M. J (2007). Forecasting recession in South Africa; A comparison of the yield curve and other economic indicators. *South African Journal of Economics volume 75 2 June 2007: 194 – 212.*

MacDonald, G. M and MacKinnon, J. G. (1985).Convenient methods for estimation of linear regression models with MA(1) errors. *Canadian Journal of Economics 17 No. 1: 106 -116.*

Makarau, A and Jury, M. R. (1997). Predictability of Zimbabwe Summer Rainfall. *International Journal of Climatology 17: 1421 – 1432.*

Manatsa, D , Chingombe, W , Matsikwa, H and Matarira, C. H. (2007). The superior influence of the Darwin Sea Level Pressure anomalies over ENSO as a single drought predictor in Southern Africa. *Theoretical and Applied Climatology D01 10.1007/s00 704 – 007 – 0315 – 3.* www.springerlink.com/index/8703666K421773hpdf

Manatsa, D , Chingombe, W , Matsikwa, H and Matarira, C. H. (2008). The impact of the positive Indian Ocean dipole on Zimbabwe droughts. *International Journal of Climatology 28: 2011 – 2029.*

Matarira, C. H. (1990). Drought over Zimbabwe in a regional and global context. *International Journal of Climatology 10: 609 – 625.*

Martin, R. V, Washington, R and Downing, T. E. (1999). Seasonal Maize Forecasting for South Africa and Zimbabwe from an Agroclimatological model. *Journal of Applied Meteorology 39: 1473 – 1479.*

Mendenhall, W and Sincich, T. (1989). *A second course in business statistics: Regression Analysis.* Collier Macmillan Publishers: London.

McCullagh, P and Nelder, J. A (1989). *Generalized Linear Models.* Chapman and Hall: London.

Rocha, A. M. C. (1992). The Influence of Global Sea Surface Temperatures on Southern African Summer Climate. *Phd thesis, University of Melbourne: 249pp*.

Ropelewski, C. F and Halpert, M . S. (1989). Precipitation patterns associated with the high phase of the Southern Oscillation. *Journal of Climate 2: 268 – 284.*

Simonoff, J. S. (2003). *Analyzing Categorical Data*: Springer: New York.

Smith, S. V. (1985). Studies of the effects of cold fronts during the rainy season in Zimbabwe. *Weather 40:* 198 – 203.

Torrance, J. D. (1990). The Southern Oscillation and the rainy season in Zimbabwe. *Zimbabwe Science News, 24: 4 – 6.*

Tsay, R. S. (2002). *Analysis of Financial Time Series.* Wiley: New York.

Unganai, S. L and Mason, S. L. (2002). Long – range predictability of Zimbabwe summer rainfall. *International Journal of Climatology, 22: 1091 – 1103*.

Waylen, P and Henworth, P. (1996). A note on the timing of precipitation variability in Zimbabwe as related to the Southern Oscillation. *International Journal of Climatology, 16: 1137 – 1148.*

World Meteorological Organization (1986). Guidelines to the Quality Control of Surface Climatological Data. *WCP – 85, WMO, AD – No. 111: 56*.

Zimbabwe Central Statistical Office handbook (October, 2004 & 2010). *Environmental Statistics Report* .

http://www.longpaddock.qld.gov.au