Modelling Mean Annual Rainfall for Zimbabwe

Retius Chifurira

January 2018

Modelling Mean Annual Rainfall for Zimbabwe

by

RETIUS CHIFURIRA
(2009057023)

THESIS

Submitted in fulfilment of the requirements for the degree of

PHILOSOPHIAE DOCTOR

in

STATISTICS

(APPLIED)

in the

FACULTY OF NATURAL AND AGRICULTURAL SCIENCES

DEPARTMENT OF MATHEMATICAL SCIENCES AND ACTUARIAL SCIENCE

at the

UNIVERSITY OF THE FREE STATE

BLOEMFONTEIN: JANUARY 2018

Thesis promoter: Dr Delson Chikobvu

UNIVERSITY OF THE FREE STATE
FACULTY OF NATURAL AND AGRICULTURAL SCIENCES
DEPARTMENT OF MATHEMATICAL STATISTICS AND ACTUARIAL SCIENCE
BLOEMFONTEIN, SOUTH AFRICA

Declaration

I, Retius Chifurira, declare that

1. The thesis is hereby submitted for the qualification of Doctor of Philosophy in Statistics at the University of the Free State.

2. The research reported in this thesis, except where otherwise indicated, is my original research.

3. This thesis has not been submitted for any degree or examination at any other University/Faculty.

4. This thesis does not contain other persons' writing, unless specifically acknowledged as being sourced from other researchers. Where other written sources have been quoted, then

   (a) their words have been re-written but the general information attributed to them has been referenced, or

   (b) where their exact words have been used, then their writing has been placed in italics and referenced.

5. I cede copyright of the thesis to the University of the Free State.

_____ _____

Retius Chifurira Date

Disclaimer

This document describes work undertaken as a PhD programme of study at the University of the Free State. All views and opinions expressed therein remain the sole responsibility of the author, and do not necessarily represent those of the institution.

A man's mind, stretched by new ideas, may never return to its original dimensions.

Oliver Wendell Holmes Jr.

Abstract

Rainfall has a substantial influence on agriculture, food security, infrastructure development, water quality and the economy. Zimbabwe, like most other Southern African countries, has distinctive meteorological features which are characterized by a high variability of temporal and spatial rainfall distributions, flash floods and prolonged drought periods. Because people struggle to adapt to these diverse rainfall patterns, a better understanding of rainfall characteristics, its distribution and potential predictors will help mitigate the effects of these adverse weather conditions. The aim of this thesis is to develop an early warning tool that can help predict a drought and/or flash flood in Zimbabwe, and to estimate the amount of rainfall during the year. In this thesis, mean annual rainfall figures from 1901 to 2015 obtained from 40 rainfall stations scattered throughout Zimbabwe were used.

The thesis consists of three sections. In the first section, appropriate statistical models are applied to a set of annual rainfall figures that have been divided by 12 to produce a mean annual rainfall figure for the year with a view towards finding potential predictors for rainfall in Zimbabwe. Monthly-based indicator variables associated with the Southern Oscillation Index (SOI) and the standardised Darwin sea level pressure readings (SDSLP) were considered as predictor variables with the SOI and SDSLP readings for August (two months before the onset of the rainfall season) producing the most important predictor variables for future rainfall in Zimbabwe.

In the second part of the thesis, several characteristics associated with the mean

annual rainfall for Zimbabwe are studied using an appropriately fitted theoretical probability distribution. More specifically, the annual rainfall figures from 1901 to 2009 were used to fit a gamma, lognormal and log-logistic distribution to the annual rainfall data. The relative performance of the fitted distributions were then assessed using the following goodness-of-fit tests, namely; the relative root mean square error (RRMSE), relative mean absolute error (RMAE) and the probability plot correlation coefficient (PPCC). All the fitted distributions, however, were not able to adequately predict periods of extreme rainfall. Extreme value distributions such as generalised extreme value and generalised Pareto distributions were then fitted to the mean annual rainfall data. The possibility that periods of extreme rainfall may be time-dependent and be influenced by weather/climate change drivers was then considered. This study shows that, although rainfall extremes for Zimbabwe are not time-dependent, they are highly influenced SDSLP anomalies for April.

In the third and last part of this thesis, we categorized rainfall data using a drought threshold value of 570 mm. We compared the relative performance of the logistic regression model in estimating drought probabilities for Zimbabwe with that of a generalised extreme value regression model for binary data. The department of meteorological services in Zimbabwe uses 75% of normal annual rainfall (usually a 30-year time series data) to declare a drought year. Results show that the GEVD regression model with SDSLP anomaly for April is the best performing model and can be used to predict drought probabilities for Zimbabwe.

Key words: Drought, early warning system, extreme value theory, floods, mean annual rainfall, southern oscillation index, standardized Darwin sea level pressure, Zimbabwe.

Acknowledgements

I gratefully acknowledge my supervisor Dr. D. Chikobvu for his inspiration, competent guidance, patience and encouragement through all the different stages of this research. I thank him for guiding me in conducting my PhD research in a challenging area of meteorological modelling application which can be used to benefit a drought-prone country such as Zimbabwe.

I would like to thank everyone in the School of Mathematics, Statistics and Computer Science at University of KwaZulu-Natal. In particular, a special thanks to Prof. Henry Mwambi for reading my draft thesis and to my best friend Knowledge Chinhamu for discussing statistics questions.

I would also like to thank Danielle Roberts for IT support and Jahvaid Hummurajudy for proof reading papers submitted for publication.

Finally, I would like to thank my wife Cornelia, my son Tinashe, my daughter Tirivashe and members of my family for their unwavering support and encouragement throughout my study.

Contents

List of Figures

List of Tables

Abbreviations

ACF	Autocorrelation Function
AD	Anderson-Darling
ADF	Augmented Dickey-Fuller
EVT	Extreme Value Theory
FAO	Food and Agriculture Organisation of the United Nations
GEVD	Generalised Extreme Value Distribution
GLM	Generalised Linear Model
GPD	Generalised Pareto Distribution
JB	Jacque Bera
KPSS	Kwiatkowski-Phillips-Schmidt Shin
LM	Lagrange Multiplier
MCMC	Markov Chain Monte Carlo
MLE	Maximum Likelihood Estimation
NLL	Negative Log-Likelihood
PACF	Partial Autocorrelation Function
PP	Phillips Perron
PPCC	Probability Plot Correlation Coefficient
Q-Q	Quantile to Quantile
RRMSE	Relative Root Mean Square Error
RMAE	Relative Mean Absolute Error
SDSLP	Standardised Darwin Sea Level Pressure
SOI	Southern Oscillation Index
SPEI	Standardised Precipitation Evapotranspiration Index

Research Output

A list of publication from this thesis is given below.

Peer-reviewed Journal Publications

1. **R. Chifurira**, D. Chikobvu, and D. Dubihlela (2016). Rainfall prediction for sustainable economic growth. *Environmental Economics* , 7(4), 120-129.

2. D. Chikobvu and **R. Chifurira** (2015). Modelling extreme minimum rainfall using the generalised extreme value distribution for Zimbabwe. *South African Journal of Science*, 111(9/10):1-8.

3. **R. Chifurira** and D. Chikobvu (2014). Modelling extreme maximum rainfall for Zimbabwe. South African Statistical Journal Proceedings: Proceedings of the 56th Annual Conference of the South African Statistical Association, 9-16.

4. D. Chikobvu and **R. Chifurira** (2012). Predicting Zimbabwe's annual rainfall using the Southern Oscillation Index: Weighted regression approach. *African Statistical Journal*, 15, 97-107.

5. **R. Chifurira** and D. Chikobvu (2017) Extreme Rainfall: Candidature probability distributions for mean annual rainfall data. Under Review. *Submitted to Journal of Disaster Risk Studies.*

Conference Presentations

1. **R. Chifurira** and D. Chikobvu. Modelling extreme minimum annual rainfall for Zimbabwe. 33rd Southern Africa Mathematical Sciences Association Annual Conference, 24-28 November 2014, Victoria Falls, Zimbabwe.

2. **R. Chifurira** and D. Chikobvu. Modelling extreme maximum annual rainfall for Zimbabwe**.** 56th Annual Conference of the South African Statistical Association, 28 - 30 October 2014, Rhodes University, Grahamstown, South Africa.

3. **R. Chifurira** and D. Chikobvu. Predicting Zimbabwe's Annual Rainfall using Darwin Sea Level Pressure Index . 54th Annual Conference of the South African Statistical Association,5-9 November 2012, Nelson Mandela Metropolitan University, Port Elizabeth, South Africa.

4. D. Chikobvu and **R. Chifurira**. Predicting Rainfall and drought using the Southern Oscillation Index in drought-prone Zimbabwe. 53th Annual Conference of the South African Statistical Association,31 October-4 November 2011, Council for Scientific and Industrial Research and Statistics South Africa, Pretoria, South Africa.

Chapter 1

Introduction

This chapter outlines the background of the study, statement of the problem, research aim, objectives, and the significance of the study. The chapter also summarizes related literature on modelling rainfall, contributions of the study and concludes with the thesis layout.

1.1 Background

There is an increasing concern in Southern Africa about the declining rainfall patterns as a result of global warming (Rurinda et al., 2014; Mushore, 2013; Mazvimavi, 2010). According to Muchunu at al. (2014) Southern Africa is a region of significant rainfall variability and is prone to serious drought and flood events. This raises serious concern, if proved correct, because rainfall has a substantial influence over agriculture, food security, infrastructure development and the economy. Zimbabwe's rain-fed agriculture production is the main driver of the economy (Mamombe at al., 2017) and any effort to revive the country's economy can be hampered by erratic rains. Although knowledge of rainfall patterns over an area may be used for strategic economic planning, it is one of the most difficult meteorological parameters to study because of a lack of reliable data and large variations of rainfall in space, and time. Therefore, developing methods that can suitably predict meteorological events is extremely valuable for both meteorologists and statisticians in light of global cli-

mate change. Zimbabwe's developmental goals and drought/floods mitigation efforts depend on long-lead time and accurate prediction of rainfall in a challenging global climate change environment. The upsurge of extreme weather events and global weather/climate change call for extensive research. Timely and accurate prediction of the amounts of rainfall in Zimbabwe is important not only in developing the economy, but also for assisting in decision-making on planning disaster risk reduction strategies by government agencies and citizens.

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 kilometers. Zimbabwe has, in previous years, been severely affected by erratic rainfall patterns and sometimes droughts. According to Rurinda et al. (2013), Zimbabwe is one of the 'hotspots' for climate change with predicted increases in rainfall variability and increased probability of extreme events such as droughts and flash floods. During the 1991 to 1992 rainy season, Zimbabwe and some Southern Africa Development Community (SADC) countries experienced the worst drought period (Zimbabwe Central Statistical Office Handbook, 1994). In the year 2000, Zimbabwe was ravaged by cyclone Eliṅe. Between 2001 to 2003, Zimbabwe only experienced rainfall in the first half of the rainfall season and a dry spell in the second half, resulting in severe droughts in some parts of the country. Between 2004 to 2008, Zimbabwe received an average amount of rainfall in the northern parts of the country while other parts received very little to no rainfall. During 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 Handbook, 2010). Due to these changes in rainfall patterns in Zimbabwe, research in predicting the amount of rainfall in the country is crucial.

1.2 Statement of the problem

Zimbabwe's economy is mainly agro-based and is thus vulnerable to the effects of weather and climatic change. The severe impact of weather and climatic change is due to the fact that the country does not have adequate resources or technology to deal with the conditions that accompany weather and climatic change. Challenges include droughts, floods, cyclones and more recently high variability in seasonal rainfall (Dodman and Mitlin, 2015; Washington and Preston, 2006). Therefore, there is a need to model annual rainfall patterns for Zimbabwe with the aim of predicting extreme annual rainfall trends. Such an analysis will be used to compute the return period, which is the mean period of time in years for a rare event such as droughts or floods of a given magnitude to be equaled or exceeded once.

1.3 Aim and objectives of the thesis

The main aim is to develop a modelling framework which can be used in the meteorological sector for carrying out accurate assessments of the frequency and level of occurrence of extreme mean annual rainfall.
The objectives are to:

(a) Propose predicting models for mean annual rainfall for Zimbabwe.

(b) Fit parent theoretical distributions to mean annual rainfall and select the best fitting distribution. Distributions considered include the gamma, normal, log-normal and log-logistic. Parent distributions concentrate on the location of the main body of the data.

(c) Fit extreme value distributions namely; two-parameter exponential, Generalised Extreme Value and Generalised Pareto Distributions to mean annual

rainfall and select the best fitting distribution. Extreme value distributions concentrate on what happens at the extremes, away from the main body of the data i.e. at the tails of the data.

(d) Identify the main weather/climatic change drivers of extreme mean rainfall for Zimbabwe.

(e) Propose prediction models for meteorological drought probabilities for Zimbabwe.

1.4 Literature review

This section reviews relevant studies in the modelling of rainfall that have been previously conducted.

1.4.1 Introduction to literature review

Natural disasters such as floods, droughts, earthquakes and other natural disasters are established in the literature as destructive to humans and their environment. The impacts of these natural disasters may be severe. Therefore, careful planning and mitigating efforts are required to reduce the risks associated with these natural disasters. This thesis focuses on reducing the risk of disasters that occur as a result of extreme rainfall such as floods and droughts. Modelling mean annual rainfall for a drought-prone country, such as Zimbabwe, is very important for decision-making in the agricultural sector and sectors involved in disaster risk reduction. Decision-making in the agricultural sector involves strategic planning with regard to water resource management and the selection of types of crops, and animals to raise. Whereas, in the disaster risk reduction sector, decision-making involves strategic planning on coping mechanisms for the country using information on the severity and occurrence (return period) of drought/flash flood. It is, therefore, important to produce very reliable predictions as the consequences of underestimation or over-

estimation can be extremely costly. Failure to accurately predict annual rainfall can result in loss of lives and low agriculture production. It can also result in economic challenges since the country's economy is agro-based.

1.4.2 Modelling rainfall for Zimbabwe

Previous studies on modelling rainfall in Zimbabwe mainly used correlation analysis. Unganai (1996) fitted a linear regression model to mean annual rainfall for Zimbabwe using the national data set for the period 1901-1994. Using temperature as a covariate, the study established that mean annual rainfall for Zimbabwe had declined by 10% during the period under investigation. Makarau (1995) also made a similar conclusion. Mason and Jury (1997) investigated quasi-periodicities in annual rainfall totals over Southern Africa. The study established that an El Niño Southern Oscillation event and sea surface temperature anomalies in the Indian and South Atlantic Oceans influence rainfall variability in Southern Africa. Chikodzi et al. (2013) used time series analysis to investigate trends in national rainfall data and Zaka rainfall station data. The study used data for the years 1930-2010 and showed a significant decline in rainfall amounts during this period.

Mazvimavi (2008) departed from the use of correlation analysis and employed non-parametric tests to investigate possible changes in extreme annual rainfall in Zimbabwe. The study used the Mann-Kendal test to investigate whether rainfall data from 40 rainfall stations in Zimbabwe vary with time. The main finding was that there was no evidence that annual rainfall data for the years 1892-2000 for each rainfall station had changed over time i.e. the Mann-Kendal test showed no significant trend with time for the data.

Mazvimavi (2010) investigated changes over time of annual rainfall for the same 40 rainfall stations using quantile regression technique. The total rainfall data set for

the period 1892-2000 was divided as follows (a) the early part of the rainfall season (October-November-December), (b) middle to end of the rainy season (January-February-March) and (c) the whole year. The study established that the annual rainfall at the 40 rainfall stations did not demonstrate evidence of changes with time in all the three time periods. The researcher noted that the absence of trends with time did not imply that global warming will not cause changes in rainfall in Zimbabwe, but the effects did not result in significant statistical changes in the historical rainfall record.

Shoko and Shoko (2014) analysed the relationship between mean annual rainfall amounts for Zimbabwe and the El Niño Southern Oscillation phases. The study used mean annual rainfall data for the period 1901-1997. The data set was divided into three categories, namely: drought (rainfall less than 488.4 mm), normal (rainfall between 489 mm and 839 mm) and wet (rainfall equal to 840 mm and more). Correlation analysis was used to find relationships between mean annual rainfall amounts in different categories and the El Niño Southern Oscillation phases. The results indicated a positive relationship between rainfall and the El Niño Southern Oscillation phases in all the categories.

Manatsa et al. (2008), using averaged seasonal rainfall for the period 1901-2000 showed the superiority of the Darwin sea level pressure anomalies over the Southern Oscillation Index (SOI) as a simple drought predictor for Southern Africa. Using correlation analysis, the study established that the averaged Darwin sea level pressure anomalies for the months March, April, May and June were the earliest and simplest predictor of drought in Zimbabwe. Hoell et al. (2017) using southern Africa's precipitation data of the months December to March for the period 1979 to 2014 showed that opposing El Niño Southern Oscillation and Subtropical Indian Ocean Dipole result in strong southern Africa climate impacts during December to March period. Manatsa et al. (2017) using correlation analysis investigated the relationship

between Nino 3.4 index an El Niño Southern Oscillation index and the standardised precipitation evapotranspiration index (SPEI) for Southern Africa. The study showed that Nino 3.4 values for the month of May are correlated to SPEI. Research on the influence of weather/climate drivers such as El Niño Southern Oscillation is not limited to Southern Africa's climate only, Roghani et al. (2016) using monthly rainfall data for the period 1951 to 2011 investigated the relationship between SOI and October to December rainfall in Iran. The study showed that average SOI and SOI phases during July to September were related with October to December rainfall.

However, in this thesis, we used a longer data set, i.e. data set for the period 1901-2009 as the in-sample data set and 2010-2015 as the out-of-sample data set. Secondly, we employed the use of extensive statistical models to (i) identify the simplest and earliest warning meteorological indicators that influence mean annual rainfall for Zimbabwe, (ii) identify the best suitable probability distribution of mean annual rainfall, (iii) describe extreme maxima and minima annual rainfall for Zimbabwe using extreme value theory and (iv) predict drought probabilities for Zimbabwe using extreme value theory. This was a clear departure from available literature on Zimbabwe and to our knowledge, there is no work which uses extreme value theory on Zimbabwean mean annual rainfall.

Much research has been conducted to study the physical and statistical properties of rainfall using observational data. Research has been done on finding the probability distributions of rainfall amount (Deka et al., 2009; Cho et al., 2004). It is generally assumed that a hydrological variable follows a certain probability distribution. Many probability distributions, in many different situations have been considered. Stagge et al. (2015), Husak et al. (2007),Cho et al. (2004), Aksoy (2000), Adiku et al. (1997),and McKee et al. (1993) modelled daily rainfall data using the gamma distribution. Suhaila et al. (2011), Deka et al. (2009), and Cho et al. (2004) inves-

tigated the performance of the lognormal distribution in fitting daily and monthly rainfall data. Fitzgerald (2005) and Ahmad et al. (1988) fitted the log-logistic distribution to daily rainfall data. Sakulski et al. (2014) fitted the log-logistic, Singh-Maddala, lognormal, generalised extreme value, Fréchet and Rayleigh distributions to spring, autumn and winter rainfall data from the Eastern Cape province, South Africa. The researchers found the Singh-Maddala distribution to be the best fitting distribution to all four seasons rainfall data. Stagge et al. (2015) fitted seven candidate distributions to standardised precipitation index (SPI) for Europe and recommended the two-parameter gamma distribution for modelling SPI. Suhaila and Jemain (2007) found that the mixed Weibull distribution was the better fitting distribution when compared to single distributions in modelling rainfall amounts in Peninsular Malaysia. Zin et al. (2009) also found that the generalised lambda distribution was the best fitting distribution for Peninsular Malaysia. However, the results obtained by Suhaila and Jemain (2007) differ from the results obtained by Zin et al. (2009). Each kind of probability distribution has its own applications and limitations. A regionalised study on the statistical modelling of annual rainfall is, therefore, very essential as statistical models may vary according to geographical location of the area and the length of the data series used.

1.5 Significance of the study

Understanding the processes governing rainfall is important for a wide range of human activities. This study provides the first application of statistical analysis of mean annual rainfall for Zimbabwe using parent and extreme value distributions. Thus, this study aims to enhance our understanding of rainfall patterns in order to develop tools that reduce the negative impact of extreme rainfall events in the agricultural and other rain-dependent sectors. The amount of rainfall is a key determinant of the success of the agriculture sector. Knowledge of rainfall behaviour at a lead-time before the onset of the rainfall season plays a pivotal role in guiding farmers and agriculture experts on the type of crops that are viable in the following rainfall season.

Knowledge of rainfall behaviour is also important in defining appropriate water harvesting and distribution plans for industrial and domestic use. Finally, knowledge of rainfall behaviour is essential as an early drought/flood monitoring tool for citizens, government agencies and non-governmental organizations involved in disaster management associated with drought and flood risk assessment. By estimating return periods of high rainfall amounts or extremely low rainfall amounts that can lead to floods or droughts, respectively, the management of the disaster risk that occurs as a result of extreme mean annual rainfall is supported.

1.6 Contributions

The main contribution of this thesis is to apply statistical techniques in modelling mean and extreme annual rainfall for Zimbabwe.
The specific contributions are to:

1. Propose a suitable model to explain the influence of Southern Oscillation Index (SOI) and standardised Darwin Sea level Pressure anomalies on mean annual rainfall for Zimbabwe.

2. Investigate the best theoretical parent distributions for mean annual rainfall for Zimbabwe.

3. Propose a suitable Generalised Extreme Value Distribution model for extreme maximum annual rainfall for Zimbabwe.

4. Propose a suitable Generalised Extreme Value Distribution model for extreme minimum annual rainfall for Zimbabwe

5. Propose a suitable Generalised Pareto Distribution model for extreme maximum annual rainfall for Zimbabwe.

6. Propose a Generalised Extreme Value Distribution binary regression model for predicting drought probabilities for Zimbabwe

1.7 Thesis layout

The thesis is organized as follows: **Chapter 2** reviews statistical tests and graphical methods used to analyse the data. **Chapter 3** describes the data used in this thesis. **Chapter 4**[1] deals with modelling mean annual rainfall using the Southern Oscillation index and standardised Darwin sea level pressure anomalies using the weighted regression approach. The best fitting candidate probability distribution for mean annual rainfall is selected in **Chapter 5**.[2] Goodness-of-fit tests, namely: RRMSE, RMAE and PPCC are used to select the best fitting distribution. The first part of the thesis shows that modelling mean annual rainfall using the 'averaging thinking' only is not enough. Therefore, in the second part of the thesis, extreme mean annual rainfall for Zimbabwe is modelled using different approaches.

In **Chapters 6 to 8**[3], modelling of the tail behaviour of the mean annual rainfall through the Extreme Value Theory (EVT) is discussed. The theory behind estimation of the parameters of the extreme value distributions are also discussed in the **Chapter 6**. In **Chapters 6** and **7**, it is revealed that extreme maxima and minima annual rainfall for Zimbabwe does not trend with time. It is also shown in the same chapters that incorporating meteorological climate change indicators improves the generalised extreme value distribution model for the data. **Chapter 8**[4] deals with modelling extreme maxima annual rainfall data using the peak-over-threshold approach. The chapter also confirms that extreme mean annual rainfall for Zimbabwe does not trend with time. In **Chapter 9**[5], we propose a binary model to predict

[1] Rainfall prediction for sustainable economic growth. Environmental Economics, Vol. 7(4), 2016, pp. 120-129; Predicting Zimbabwe's annual rainfall using the Southern Oscillation Index: Weighted regression approach. The African Statistical Journal, Vol. 15, 2012, pp. 87-107.

[2] Extreme Rainfall: Selecting the best probability distribution for mean annual rainfall: An application to Zimbabwean data

[3] Modelling extreme maximum annual rainfall for Zimbabwe. South African Statistical Journal: Peer-reviewed Proceedings of the 56th Annual Conference of the South African Statistical Association, 2014, pp. 9-16; Modelling of extreme minimum rainfall using generalised extreme value distribution for Zimbabwe, South African Journal of Science, 2015, pp. 1-8.

[4] Modelling mean annual rainfall extremes using a Generalised Pareto Distribution Model

[5] Generalised Extreme Value Regression with Binary Dependent variable: An application to predicting meteorological drought probabilities. Submitted to Journal of disaster risk studies, 2017

drought probabilities when the binary dependent variable is extreme. In order to adequately predict drought probabilities, we used the generalised linear model (GLM) with the quantile function of the generalised extreme value distribution (GEVD) as the link function. **Chapter 9** shows that the GEVD regression model performs better than the logistic model, thereby providing a good alternative candidate for predicting drought probabilities for Zimbabwe. The chapter also establishes that standardised Darwin sea level pressure anomalies for April of the same year is an earlier and skillful predictor of meteorological droughts in Zimbabwe.

The following statistical packages were used in this thesis: R, Eviews and SAS.

Chapter 2

Methodology

2.1 Introduction

This chapter mainly presents statistical tests and graphical methods used to analyse the mean annual rainfall data in Zimbabwe.

2.2 Tests for stationarity

A common assumption in many time series techniques is that the data is stationary. There are two types of stationarity, namely weakly stationary (usually referred to as stationary) and strongly stationary (usually referred to as strictly stationary). A time series $\{x_t, t \in \mathbb{Z}\}$ (where $\mathbb{Z}$ is the integer set) is said to be weakly stationary or just stationary if

(i) $E(x_t^2) < \infty \; \forall \; t \in \mathbb{Z}$.

(ii) $E(x_t) = \mu \; \forall \; t \in \mathbb{Z}$.

(iii) $\mathrm{Cov}(x_s, x_t) = \mathrm{Cov}(x_{s+h}, x_{t+h}) \; \forall \; s, t, h \in \mathbb{Z}$.

Thus, a stationary time series has the property that the mean, variance and autocorrelation structure is time invariant i.e. depends only on $(t - s)$ and not on s or t.

A time series $\{x_t, t \in \mathbb{Z}\}$ is said to be strongly or strictly stationary if the joint distributions

$$(x_{t_1}, ..., x_{t_k}) \stackrel{D}{=} (x_{t_{1+h}}, ..., x_{t_{k+h}})$$

for all sets of time points $t_1, ..., t_k \in \mathbb{Z}$ and integer h.

In this study, the augmented-Dickey Fuller (ADF), Phillips-Perron(PP) and Kwiatkowski-Phillips-Schmidt-Shin (KPSS) tests are used to formally test for stationarity in data. The null hypothesis for the ADF and PP tests is that the mean annual rainfall series is non-stationary while for the KPSS test is that the mean annual rainfall series is stationary.

2.2.1 Augmented-Dickey Fuller (ADF) test

The ADF test is one of the most commonly used tests for stationarity of time series. The test is also known as the unit root (non-stationary) test. There are three cases of the ADF test equation depending on the nature of the time series data being tested.

a) When the time series is flat (no trend) and potentially slow-turning to zero value. The test equation is:

$$\Delta x_t = \phi x_{t-1} + \theta_1 \Delta x_{t-1} + \theta_2 \Delta x_{t-2} + ... + \theta_k \Delta x_{t-k} + \epsilon_t$$

The equation has no intercept and no time trend.

b) When the time series is flat (no trend) and potentially slow-turning to non-zero value. The test equation is:

$$\Delta x_t = \alpha_0 + \phi x_{t-1} + \theta_1 \Delta x_{t-1} + \theta_2 \Delta x_{t-2} + ... + \theta_k \Delta x_{t-k} + \epsilon_t$$

The equation has an intercept (α_0) but no time trend.

c) When the time series has trend in it (either up or down) and is potentially slowly turning around a trend line you would draw through the data. The test

equation is:

$$\Delta x_t = \alpha_0 + \phi x_{t-1} + \gamma t + \theta_1 \Delta x_{t-1} + \theta_2 \Delta x_{t-2} + ... + \theta_k \Delta x_{t-k} + \epsilon_t$$

The equation has an intercept (α_0) and time trend γt. In all cases ϵ_t is a white noise error.

where $\epsilon_t \sim \text{i.i.d.}(0, \sigma^2)$ and the number of augmenting lags (k) is determined by minimising the Schwartz Bayesian information criterion or lags are dropped until the last lag is statistically significant. According to Niyimbanira (2013) the ADF test relies on rejecting the null hypothesis of the data need to be differenced to make it stationary (the series is non-stationary) in favour of the alternative hypothesis of the data is stationary and does not need to be differenced. Once a value for the test statistic:

$$ADF = \frac{\hat{\phi}}{\hat{\sigma}} \quad \text{where } \hat{\phi} = \rho - 1,$$

is computed it can be compared to the relevant critical value for the Dickey-Fuller test. $\hat{\sigma}$ is the OLS standard error of $\hat{\phi}$. If the test statistic is less than the critical value, then the null hypothesis of $\phi = 0$ is rejected and no unit root is present.

2.2.2 Phillips-Perron (PP) test

The PP stationarity test and the ADF test differ mainly on how they deal with serial correlation and heteroscedasticity in the errors. The ADF test uses a parametric auto-regression to approximate the auto-regressive moving average structure of the errors in the regression while the PP test ignore any serial correlation in the test regression. For the model

$$x_t = \alpha_0 + \phi x_{t-1} + \epsilon_t,$$

the test regression for the PP test is:

$$\Delta x_t = \alpha_0 + \delta x_{t-1} + \epsilon_t$$

where we may exclude the constant (α_0) or include a trend term γt and

$$\epsilon_t = \delta\epsilon_t + u_t, u_t \sim \text{i.i.d}(0, \sigma^2),$$

where i.i.d stand for independent and identically distributed. The null hypothesis to be tested is $H_0 : \delta = 0$. The PP test correct for any serial correlation and heteroscedasticity in the errors ϵ_t of the test regression by directly modifying the ADF test statistic. There are two statistics, Z_δ and Z_t, calculated as:

$$Z_\delta = n\hat{\delta} - \frac{1}{2}\frac{n^2(\text{s.e}(\hat{\delta}))}{\hat{\sigma}^2}(\hat{\lambda}_n^2 - \hat{\sigma}^2) \tag{2.1}$$

$$Z_t = \sqrt{\frac{\hat{\sigma}^2}{\hat{\lambda}_n^2}}\frac{\hat{\delta}_n - 1}{\text{s.e}(\hat{\delta})} - \frac{1}{2}(\hat{\lambda}_n^2 - \hat{\sigma}^2)\frac{1}{\hat{\lambda}_n^2}\frac{n(\text{s.e}(\hat{\delta}))}{\hat{\sigma}^2} \tag{2.2}$$

where $\hat{\sigma}^2$ and $\hat{\lambda}_n^2$ are the consistent estimates of the variance parameters

$$\lambda_n^2 = \lim_{n\to\infty} E\left[\frac{1}{n}\epsilon_t^2\right] \tag{2.3}$$

$$\sigma^2 = \frac{1}{n-p}\lim_{n\leftarrow\infty}\sum_{t=1}^{n} E(\epsilon_t^2) \tag{2.4}$$

where ϵ_t is the OLS residual, p is the number of covariates in the regression. Under the null hypothesis that $\delta = 0$, Z_t and Z_δ statistics have the same asymptotic distributions as the ADF statistic.

2.2.3 Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test

The KPSS test has been developed to complement tests for non-stationarity as the ADF and PP tests have low power with respect to near non-stationary and long-run trend processes. KPSS has stationary as the null hypothesis. Assuming there is no trend, then it follows that:

$$x_t = \boldsymbol{\beta}'\mathbf{D}_t + \mu_t + u_t \tag{2.5}$$

$$\mu_t = \mu_{t-1} + \epsilon_t, \epsilon_t \sim N(0, \sigma_\epsilon^2),$$

where $\mathbf{D}_t$ contains deterministic components (constant or constant plus time trend). The null hypothesis that x_t is stationary, $\mathbf{I}(0)$, is formulated as $H_0 : \sigma_\epsilon^2 = 0$, which implies that μ_t is a constant. The KPSS test statistic is the Langrange multiplier (LM) or score statistic for testing $\sigma_\epsilon^2 = 0$ against the alternative that $\sigma_\epsilon^2 > 0$ and is given by

$$KPSS = \left(n^{-2} \sum_{t=1}^{n} \hat{S}_t^2 \right) / \hat{\lambda}^2 \tag{2.6}$$

where $\hat{S}_t = \sum_{j=1}^{t} \hat{u}_j$, $\hat{u}_t$ is the residual of a regression of y_t and $\mathbf{D}_t$. $\hat{\lambda}^2$ is a consistent estimate of the long-run variance of u_t using $\hat{u}_t$.

2.3 Tests for randomness and serial correlation

2.3.1 Bartels rank test

This is the rank version of von Neumann's Ratio test for randomness. The test statistic is:

$$RVN = \frac{\sum_{i=1}^{n-1}(R_i - R_{i+1})^2}{\sum_{i=1}^{n}(R_i - (n+1)/2)^2} \tag{2.7}$$

where $R_i = \text{rank}(y_i)$, $i = 1, ..., n$. It is known that $(RVN - 2)/\sigma$ is asymptotically standard normal, where $\sigma^2 = \frac{4(n-2)(5n^2-2n-9)}{5n(n+1)(n-1)}$. The possible alternatives are "two sided", "left sided" and "right sided". By using the alternative "two sided" the null hypothesis of randomness is tested against non-randomness. By using the alternative "left sided" the null hypothesis of randomness is tested against a trend. By using the alternative "right sided" the null hypothesis of randomness is tested against a systematic oscillation. In this thesis we use the "two sided" alternative.

2.3.2 Ljung-Box test

The Ljung-Box test is widely applied in econometrics and other applications of time series analysis. Ljung-Box (1978) is a modification of the Box and Pierce (1970)'s

Portmanteau Statistic for auto-correlation

$$Q^*(m) = n \sum_{l=1}^{m} \hat{\rho}_l^2 \tag{2.8}$$

a test statistic for the null hypothesis $H_0 : \rho_1 = ... = \rho_m = 0$, that is the data are independently distributed. The Ljung-Box statistic is

$$Q(m) = n(n+2) \sum_{l=1}^{m} \frac{\hat{\rho}_l^2}{n-l}. \tag{2.9}$$

The null hypothesis of independence and identically distributed is rejected if $Q_m > \chi_\alpha^2$, where χ_α^2 denotes the $100(1-\alpha)^{th}$ percentile of the chi-squared distribution with m degrees of freedom. The Ljung-Box test has more power in finite samples than the Box and Pierce (1970) test.

2.3.3 Brock-Dechert-Scheinkman (BDS) test

The BDS test is a portmanteau test time based dependence in a series and can be used for testing against a variety of possible deviations from independence including linear dependence, non-linear dependence, or chaos. This test can be applied to a data series to check for independence and identically distribution (i.i.d.). The null hypothesis is the data series is i.i.d. against an unspecified alternative (Dutta et al., 2015; Kuan, 2008). The BDS test takes its root from the concept of a correlation integral, a measure of spatial correlation in m-dimensional space originally developed by Grassberger and Procaccia (1983). The correlation integral at embedding dimension m can be estimated by

$$C_{m,\epsilon} = \frac{2}{T_m(T_m-1)} \sum_{m \le s \le t \le T} I(x_t^m, x_s^m; \epsilon) \tag{2.10}$$

where $T_m = n - m + 1$ and $I(x_t^m, x_s^m; \epsilon)$ is an indicator function which is equal to one if $\mid x_{t-i} - x_{s-i} \mid < \epsilon$ for $i = 0, 1, ..., m-1$, and zero otherwise.

The correlation integral estimates the probability that any two m-dimensional points

are within a distance ϵ of each other. That is, it estimates the joint probability:

$$\Pr(\mid x_t - x_s \mid < \epsilon, \mid x_{t-i} - x_{s-i} \mid < \epsilon, ..., \mid x_{t-m+i} - x_{s-m+i} \mid < \epsilon).$$

If x_t are i.i.d. this probability should be equal to the following in the limiting case:

$$C_{1,\epsilon}^m = \Pr(\mid x_t - x_s \mid < \epsilon)^m.$$

Brook et al. (1996) define the BDS statistic as follows:

$$V_{m,\epsilon} = \sqrt{n}\frac{C_{m,\epsilon} - C_{1,\epsilon}^m}{s_{m,\epsilon}}, \tag{2.11}$$

where $s_{m,\epsilon}$ is the standard of $\sqrt{n}(C_{m,\epsilon} - C_{1,\epsilon}^m)$ and can be estimated consistently. Under moderate regularity conditions, the BDS statistic converges in distribution to the standard normal distribution so that the null hypothesis of i.i.d. is rejected at the 5% significance level whenever $\mid V_{m,\epsilon} \mid > 1.96$ (Brook et al., 1996).

Alternatively, we establish if the data series $\{x_1, x_2, ..., x_n\}$ are a realization of an i.i.d. process:

$$x_t \sim \text{i.i.d.}(\mu, \sigma^2)$$

using a correlogram. If the data $\{x_1, x_2, ..., x_n\}$ were really generated by an i.i.d. process, then about 95% of the autocorrelations $\hat{\rho}_1, \hat{\rho}_2, ..., \hat{\rho}_n$ should fall between the bounds $\pm\frac{1.96}{\sqrt{n}}$, i.e. if the process is i.i.d. we would expect 95% of the sample autocorrelations to lie within the dashed lines of the Auto-Correlation Function (ACF) plot (Maddala and Kim, 1998).

In this thesis, the ACF plot and the theoretical tests for randomness and serial correlation will be used.

2.4 Test for heteroscedasticity

2.4.1 ARCH-LM test

An uncorrelated time series can still be serially dependent due to a dynamic conditional variance process. A time series exhibiting conditional heteroscedasticity or autocorrelation in the squared series is said to have autoregressive conditional heteroscedastic (ARCH) effects. The ARCH is a Lagrange Multiplier (LM) test of Engle (1982) to assess the significance of ARCH effects. The test is usually referred to as the ARCH-LM test.

Consider a time series

$$x_t = \mu_t + \epsilon_t$$

where μ_t is the conditional mean of the process, and ϵ_t is an innovation with mean zero.

Suppose the innovations are generated as $\epsilon_t = \sigma_t \varepsilon_t$ where ε_t is an independent and identically distributed process with mean 0 and variance 1. Thus, $E(\epsilon_t \epsilon_{t+h}) = 0$ for all lags $h \neq 0$ and the innovations are uncorrelated.

Let H_t denote the history of the process available at time t. The conditional variance of x_t is:

$$\mathrm{var}(x_t \mid H_{t-1}) = \mathrm{var}(\epsilon_t \mid H_{t-1}) = E(\epsilon_t^2 \mid H_{t-1}) = \sigma_t^2.$$

Thus, conditional heteroscedasticity in the variance process is equivalent to autocorrelation in the squared innovation process.

Now we define the innovation (residual) series as:

$$\hat{\epsilon}_t = x_t - \hat{\mu}_t$$

and the squared series ϵ_t^2, is then used to check for conditional heteroscedasticity, which is also known as the ARCH effects. The presence of the ARCH effect is based on the linear regression

$$\epsilon_t^2 = \alpha_0 + \alpha_1 \epsilon_{t-1}^2 + ... + \alpha_m \epsilon_{t-m}^2 + \varepsilon_t, \quad \text{for } t = m+1, ..., n,$$

where ε_t denote the error term, m is a pre-specified positive integer, and n is the sample size. The null hypothesis corresponding to homoscedasticity is:

$$H_0 : \alpha_1 = ... = \alpha_m = 0$$

The test statistic for Engle's ARCH-LM test is the usual F statistic for regression on the squared residuals. Under the null hypothesis, the F statistic follows a χ^2 distribution with m degrees of freedom. A large critical value indicates rejection of the null hypothesis in favour of the alternative.

2.5 Tests for normality

2.5.1 Jarque-Bera test

In statistical analysis the assumption of data coming from a normal distribution is often made, however, most informed opinion have accepted that populations might be non-normal. The Jarque-Bera (JB) test is a goodness-of-fit test of departure from normality based on the sample skewness and kurtosis. The JB statistic is:

$$JB = n \left[\frac{s^2}{6} + \frac{(k-3)^2}{24} \right] \tag{2.12}$$

where s is the skewness and k is the kurtosis statistics. The skewness estimate statistic is defined as:

$$s = \frac{1}{n}\frac{\sum_{t=1}^{n}(x_t - \overline{x})^3}{(\hat{\sigma}^2)^{\frac{3}{2}}}$$

where x_t is each observation and

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{t=1}^{n}(x_t - \overline{x})^2.$$

The skewness gives a measure of how symmetric the observations are about the mean.

The kurtosis estimate is defined as:

$$k = \frac{1}{n}\frac{\sum_{t=1}^{n}(x_t - \overline{x})^4}{(\hat{\sigma}^2)^2}$$

The kurtosis gives a measure of the thickness in the tails of a probability density function.

The JB statistic follows a χ^2 distribution with two degrees of freedom. For large sample sizes, the null hypothesis of normality is rejected if the JB statistic is greater than the critical value from the χ^2 distribution with 2 degrees of freedom.

2.5.2 Shapiro-Wilk test

The Shapiro-Wilk test calculates a W statistic that tests whether a random sample $x_1, x_2, ..., x_n$ comes from a normal distribution. The W statistic is:

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_i\right)^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \tag{2.13}$$

where x_i are the ordered sample values and a_i are constants generated from the means, variances and covariances of the order statistics of a sample of size n from a normal distribution.

The W statistics is obtained as an F-ratio. The null hypothesis of normality is rejected if the calculated p-value is less than α the level of significance.

2.6 Model adequacy and goodness-of-fit tests

2.6.1 Probability-Probability plot

The probability-probability (PP) plot is a graphical technique for determining if the percentiles of the data and the fitted theoretical distribution come from the same underlying distribution. If the percentiles are identically distributed variables, then the PP plot will be a straight line configuration oriented from (0,0) to (1,1). The PP plot is usually sensitive to discrepancies in the middle of a distribution rather in the tails. Therefore, the PP plot is not desirable for model adequacy using heavy-tailed data.

2.6.2 Quantile-Quantile plot

The quantile-quantile (QQ) plot is a graphical technique for determining if the quantiles of the data and the fitted theoretical distribution come from the same distribution. If the two sets (quantiles of data and fitted theoretical distribution) come from a population with the same distribution, the points should fall approximately along the 45-degree reference line. Any significant departure from the reference line indicates that the quantiles of the data and the fitted theoretical distribution have come from populations with different distributions. The QQ plot tends to emphasise the comparative structure in the tails and to blur the distinctions in the middle of the distribution. For this reason, the QQ plot is a better graphical technique when fitting heavy-tailed distributions.

2.6.3 Kolmogorov-Smirnov test

The Kolmogorov-Smirnov (KS) test is used to decide if a sample comes from a population with a specific distribution. The KS test belongs to the supremum class of empirical distribution function statistics and is based on the largest vertical difference between the hypothesized and empirical distribution (Razali and Wah, 2011). Given that $x_1, x_2, ..., x_n$ are n ordered data points, Conover (1999) defined the test statistic proposed by Kolmogorov (1933) as

$$T = \sup_x \mid F^*(x) - F_n(x) \mid \tag{2.14}$$

where "sup" stands for supremum which means the greatest. $F^*(x)$ is the hypothesized distribution function whereas $F_n(x)$ is the empirical distribution function. If T exceeds the $1 - \alpha$ quantile as given by the table of quantiles for the KS test statistic, then we reject the null hypothesis that the data follows a specified distribution at the level of significance, α. The main disadvantage of the KS test statistic is that it places more weight at the center of the distribution than on the tails and therefore cannot adequately check goodness-of-fit of heavy tailed distribution.

2.6.4 Anderson-Darling test

The Anderson-Darling (AD) test is an improvement of the Kolmogorov-Smirnov (KS) test. It gives more weight to the tails of the distribution than the KS test does (Farrel and Rogers-Stewart, 2006). The AD statistic is defined as:

$$W_n^2 = n \int_{-\infty}^{\infty} [F_n(x) - F^*(x)]^2 \psi(F^*(x)) dF^*(x) \tag{2.15}$$

where ψ is a non-negative weight function which can be computed from

$$\psi = [F^*(x)(1 - F^*(x))]^{-1}.$$

In order to make the computation of the statistic easier, Arshad et al. (2003) redefined the AD statistic as

$$W_n^2 = -n - \frac{1}{2}\sum_{t=1}^{n}(2t-1)[\ln F^*(x_t) + \ln(1 - F^*(x_{n-t+1}))] \tag{2.16}$$

where $F^*(x_t)$ is the cumulative distribution function of the specified distribution, y_t is ordered data and n is the sample size.The null hypothesis of the AD test is that the data follows a specified distribution at the level of significance, α. According to Arshad et al. (2003), the AD test is the most powerful empirical distribution function test. Since the AD statistic is a measure of the distance between the empirical and hypothesized distribution functions, the fitted distribution with the smallest AD statistic value will be selected as the best fitting model.

In this thesis the QQ plot and AD test are used to check for model adequacy of the fitted models.

2.7 Model selection

In this thesis the Akaike information criterion (AIC) and natural forecast statistic namely; the relative root mean square error (RRMSE) and relative mean absolute error (RMAE) and probability plot correlation coefficient (PPCC) are used to select the best fitting model against competing models.

2.7.1 Akaike information criterion

AIC is a measure of how well the fitted model fits with the data in respect to candidate models. AIC estimates the quality of each model relative to the other models. The AIC is given by:

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

where l is the log likelihood, k is the number of parameters in the model and n is the sample size. The model with the smallest AIC value is usually regarded as the best

model for the data (Tsay, 2013).

2.7.2 Assessing model performance

In this thesis the measures of average error namely, mean absolute percentage error (MAPE) and root mean square error (RMSE) are applied to evaluate model performance. These measures are based on statistical summaries of $\epsilon_t(t = 1, 2, .., n)$. The average model-estimation error can be written generically as:

$$\bar{\epsilon}_t = \left[\frac{\sum_{t=1}^{n} v_t \mid \epsilon_t \mid^{\omega}}{\sum_{t=1}^{n} v_t} \right]^{\frac{1}{\omega}} \tag{2.17}$$

where $\omega \geq 1$ and v_t is a scaling assigned to each $\mid \epsilon_t \mid^{\omega}$ according to its hypothesized influence on the total error (Willmott and Matsuura, 2005). For the calculation of RMSE, $\omega = 2$ and $v_t = 1$. RMSE is measured in the same unit as the forecast and is given by:

$$RMSE = \left[n^{-1} \sum_{t=1}^{n} \mid \epsilon_t \mid^2 \right]^{\frac{1}{2}} \tag{2.18}$$

The MAPE is also measured in the same unit as the forecast, but gives less weight to large forecast errors than the RMSE. To obtain the MAPE we set $\omega = 1$ and $v_t = 1$ and is given by:

$$MAPE = \left[n^{-1} \sum_{t=1}^{n} \mid 100\epsilon_i \mid \right] \tag{2.19}$$

According to Willmott and Matsuura (2005) and Trúck and Liang (2012), MAPE is the most natural measure of average error magnitude and it is an unambiguous measure of the average error magnitude. The MAPE and RSME values can range from 0 to infinity and smaller values indicate a better model.

2.7.3 Forecast statistics

The relative root mean square error (RRMSE) and relative mean absolute error (RMAE) assesses the difference between the observed values and the expected values of the assumed distributions. The probability plot correlation coefficient (PPCC) measures the correlation between ordered values and the corresponding expected values of the assumed distribution. The formulae for the tests are

$$RRMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_{i:n} - \hat{Q}(F_i)}{x_{i:n}}\right)^2} \tag{2.20}$$

$$RMAE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{x_{i:n} - \hat{Q}(F_i)}{x_{i:n}}\right| \tag{2.21}$$

$$PPCC = \frac{\sum_{i=1}^{n}(x_{i:n} - \overline{x})(\hat{Q}(F_i) - \overline{Q}(F_i))}{\sqrt{\sum_{i=1}^{n}(x_{i:n} - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(\hat{Q}(F_i) - \overline{Q}(F))^2}} \tag{2.22}$$

where $x_{i:n}$ is the observed values of the ith order statistics of a random sample of size n. $\hat{Q}(F_i)$ is the estimated quantile value of the assumed distribution associated with the ith Landwehr plotting position, $F_i = \frac{i-0.35}{n}$. $\overline{Q}(F_i)$ and $\overline{x}$ are the averages of $\hat{Q}(F_i)$ and $x_{i:n}$ respectively. Heo et al. (2008) investigated the performance of different probability plotting positions for extreme value distribution and recommended the Landwehr plotting position for medium sample sizes. In this thesis we use the Landwehr plotting position to estimate quantiles of the fitted distributions. The fitted distribution with the smallest values of the RRMSE and RMAE is selected as the best fitting distribution while the distribution with the computed PPCC closest to 1 indicates the best fitting distribution.

Chapter 3

Rainfall data

3.1 Introduction

In this chapter we described the data used in this thesis. Rainfall data series from Zimbabwe, the Southern Oscillation Index (SOI) and standardised Darwin sea level pressure (SDSLP) anomalies were used in this study. The SOI and SDSLP were used to investigate their influence on the amount of rainfall in Zimbabwe. The descriptive statistics and properties, such as stationarity and distribution of the rainfall data set, are discussed. Lastly, we used data correlation techniques to select the SOI and SDSLP for a particular month at a lag which can significantly influence rainfall in Zimbabwe.

3.2 Mean annual rainfall data

We used the mean annual rainfall data series for the period 1901-2015 obtained from the Department of Meteorological Services in Zimbabwe. In this study, mean annual rainfall data was chosen over using monthly or daily rainfall data because (i) at monthly and daily time-lines, rainfall amount records for rainfall stations contains many zeros (no rainfall amount recorded), (ii) the focus of the thesis was to develop a national drought/flash floods early warning system. The data set was divided into the in-sample data set (1901-2009) and the out-of-sample data set (2010-2015).

The rainfall amounts were from 40 rainfall stations located across the country. The rainfall stations had monthly rainfall recordings for more than 100 years. Figure 3.1 shows the map and locations of these rainfall stations.

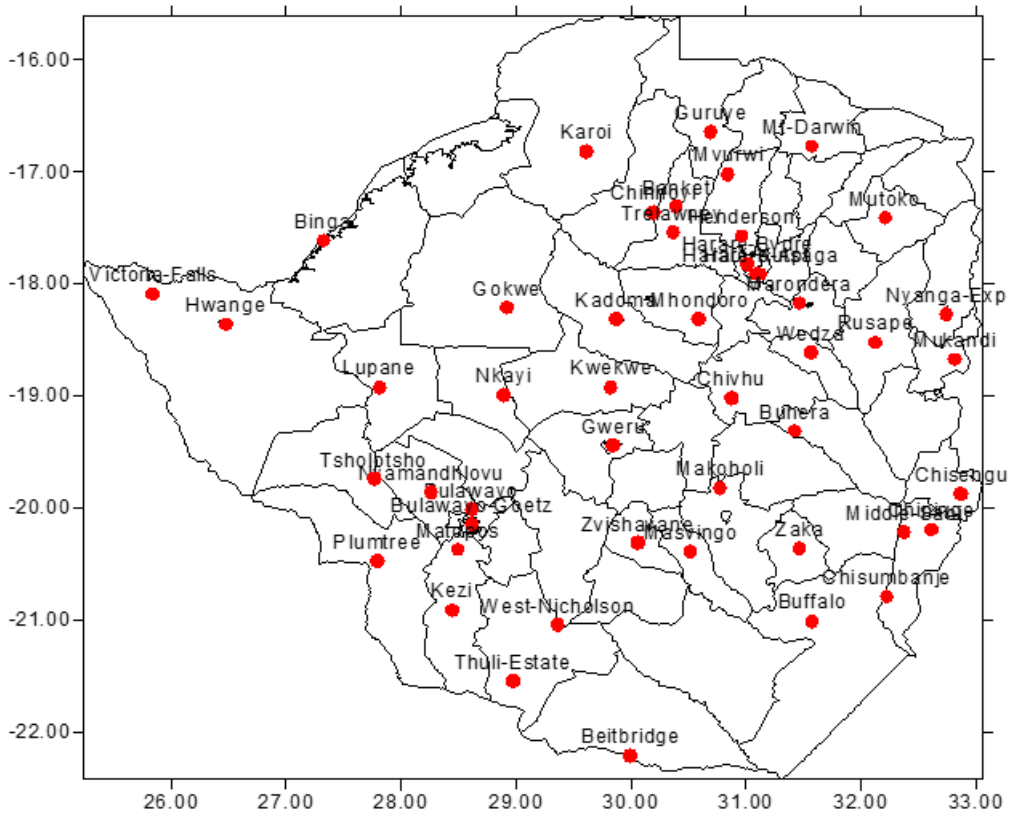

Figure 3.1: Location of the rainfall stations in Zimbabwe (selected for this study)

The mean annual rainfall for Zimbabwe was calculated. The mean annual rainfall were considered to improve the signal to noise ratio compared to the use of individual station data. The rainfall season in Zimbabwe stretches from mid November to mid-March of the following year (Mamombe, 2017). Thus, an average annual rainfall for example 2000 means the average rainfall recorded from October 2000 to April 2001. Figure 3.2 shows the time series plot of mean annual rainfall for the period 1901-2009.

The time series plot in Figure 3.2 shows that the mean annual rainfall appears to be

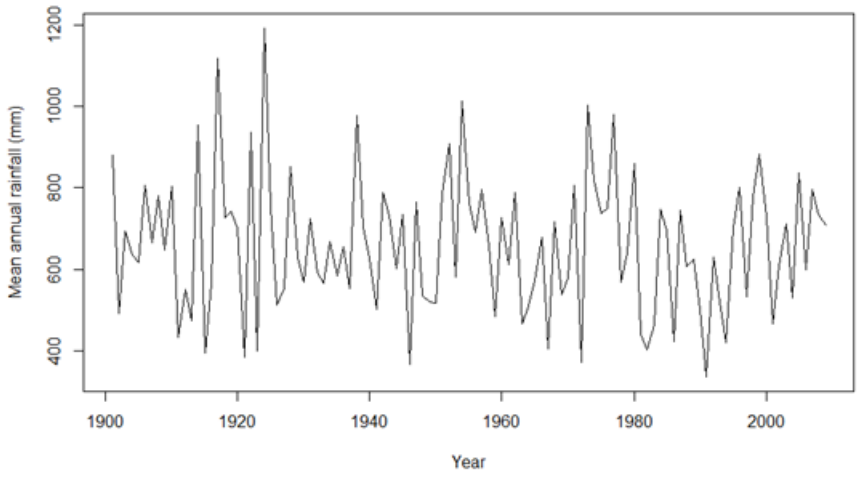

Figure 3.2: Time series plot of mean annual rainfall for Zimbabwe for the period 1901-2009

stationary. We formally tested for stationarity of mean annual rainfall data using the augmented Dickey-Fuller (ADF), Phillips-Perron (PP) and Kwiatkowski-Phillips-Schmidt-Shin (KPSS) tests. The results of testing for stationary of mean annual rainfall are reported in Table 3.1.

Table 3.1: Results of tests for stationarity of mean annual rainfall data for the period 1901-2009

Test	ADF	PP	KPSS
p-value	< 0.01	< 0.01	> 0.10

The null hypothesis for the ADF and PP statistic is that the data series is not stationary. Both the ADF and PP statistics with p-values $< 0.01 < 0.05$, reject that the null hypothesis at 5% significance level implying that the mean annual rainfall data are stationary. This is confirmed by the KPSS statistic with p-value > 0.05 which fails to reject the null hypothesis for stationarity at 5% level of significance. Table 3.2 presents the summary statistics for the mean annual rainfall data.

The small positive skewness and excess kurtosis clearly illustrates the normality of

Table 3.2: Summary statistics for mean annual rainfall data for Zimbabwe for the period 1901-2009

No. of obs.	Mean	Std. dev.	Min	Max	Skewness	Excess kurtosis
109	659.9312	169.2457	335.3000	1192.6200	0.4455	0.1222

the distribution of mean annual rainfall data. We formally tested for normality of the rainfall data using the Jarque-Bera and Shapiro-Wilk tests. Table 3.3 shows the p-values for test of normality.

Table 3.3: Results of tests for normality of mean annual rainfall data

Test	p-value
Jarque-Bera	0.4120
Shapiro-Wilk	0.1203

Both the Jarque-Bera and Shapiro-Wilk tests confirm that the mean annual rainfall data is normality distributed.

Modelling data with statistical distributions usually assumes that the data are independent and identically distributed (i.e., randomness), with no serial correlation and no heteroscedasticity. We tested for randomness and serial correlation using the ACF plot, Brock-Dechert-Scheinkman (BDS), Bartels (1982) and the Ljung-Box tests. The null hypothesis for the tests is that the annual rainfall is independent and identically distributed (i.i.d). The ACF of the mean annual rainfall is shown in Figure 3.3.

From Figure 3.3, all the sample correlations lie within the dashed lines suggesting that the mean annual rainfall data is independent and identically distributed and no serial correlation exist. We formally tested for independent and identically distributed property. The corresponding p-values, based on the mean annual rainfall are given in Table 3.4.

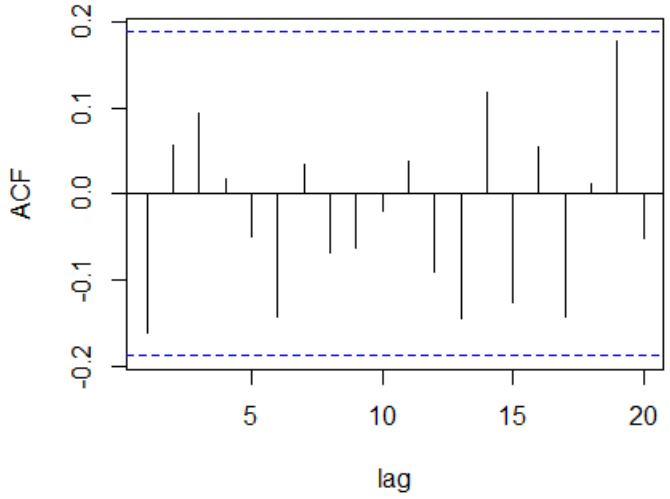

Figure 3.3: ACF plot of mean annual rainfall for Zimbabwe for the period 1901-2009

Table 3.4: Results of tests for i.i.d. of mean annual rainfall data for the period 1901-2009

Test	p-value
BDS	0.3970
Bartels (1982)	0.7570
Ljung-Box	0.1270

We tested for no heteroscedasticity using the ARCH LM test. The null hypothesis for the test is that the annual rainfall data has no presence of heteroscedasticity. The corresponding p-values based on the annual rainfall are given in Table 3.5.

Table 3.5: Results of test for heteroscedasticity of mean annual rainfall data for the period 1901-2009

Test	p-value
ARCH-LM	0.2790

The tests reported in Tables 3.4 and 3.5 are semi-parametric in nature, i.e., no strict distributional assumptions are made about the data. The tests confirm that the mean annual rainfall data are independent and identically distributed, with no serial cor-

relation and have no heteroscedasticity.

3.3 Weather/climate change determinants

In this section, we briefly described the two weather/climate change determinants we chose to investigate their influence on the annual rainfall for Zimbabwe. We also reported the significant correlations between annual rainfall and weather/climate change determinants values at some specified lag.

3.3.1 Southern oscillation index and standardised Darwin sea level pressure

The Southern oscillation index (SOI) and standardised Darwin sea level pressure (SDSLP) data were obtained from National Oceanic and Atmospheric Administration (NOAA). The SOI is calculated from monthly or seasonal fluctuations in the air pressure difference of the area between Tahiti (mid-Pacific) and Darwin (Australia). Sea level pressure (SLP) is the atmospheric pressure at mean sea level measured by stations at sea level (Mason and Jury, 1997). The SOI is a measure of the strength and phase of the difference in sea-level pressure between the two locations, expressed as an index. SOI is calculated (Satish et al., 2014) using the formula:

$$\text{SOI} = 10\frac{\text{Standardised Tahiti} - \text{Standardised Darwin}}{\text{MSD}}$$

where:

$$\text{Standardised Tahiti} = \frac{(\text{Actual Tahiti SLP} - \text{Mean Tahiti SLP})}{\text{Standard deviation Tahiti}},$$

where

$$\text{Standard deviation Tahiti} = \sqrt{\sum \frac{(\text{Actual Tahiti SLP} - \text{Mean Tahiti SLP})^2}{N}}$$

where $N =$ number of months

and

$$\text{Standardized Darwin} = \frac{(\text{Actual Darwin SLP} - \text{Mean Darwin SLP})}{\text{Standard deviation Darwin}},$$

where

$$\text{Standard deviation Darwin} = \sqrt{\sum \frac{(\text{Actual Darwin SLP} - \text{Mean Darwin SLP})^2}{N}}$$

where $N =$ number of months

and

$$MSD = \sqrt{\sum \frac{(\text{Standardised Tahiti} - \text{Standardised Darwin})^2}{N}}$$

where $N =$ number of months

Comparing lead times of SOI and SDSLP values

Monthly SOI and SDSLP values were matched to mean annual rainfall. Table 3.6 shows the correlations between mean annual rainfall and monthly weather/climatic determinants at current and at a lag of one year.

The focus of this thesis is developing an early warning system for disaster management for Zimbabwe. In order to inform government departments and other stakeholders involved in disaster risk reduction, a timeous prediction of extreme rainfall for the country is important. Predicting mean annual rainfall at a longer lead-time before the onset of rainfall season guards against the risk of disaster and can help the country prepare for extreme rainfall. Thus, we aim to determine a particular month's weather/climatic determinant which has a high correlation with mean annual rainfall at a longer lead-time. From Table 3.6, the highest correlation between the mean annual rainfall for Zimbabwe and monthly SOI values is 0.4289 (SOI value for August of the current year), which gave a lead-time of 2 months before the onset

Table 3.6: Correlations between mean annual rainfall data for the period 1901-2009 and climate change determinants at current and at a lag of one year

Month	SOI		SDSLP	
	current	lag of a year	current	lag of a year
January	0.0468	0.0792	0.0002	0.0780
February	0.1349	0.0120	-0.0784	0.0120
March	0.1748	0.0183	-0.2068	0.0181
April	0.1684	-0.0526	-0.2479	-0.0523
May	0.2570	-0.0429	-0.2000	-0.0424
June	0.2346	0.1118	-0.2022	0.1115
July	0.3027	-0.0484	-0.2166	-0.0482
August	0.4289	-0.0572	-0.3359	-0.0572
September	0.3866	-0.0666	-0.2586	-0.0664
October	0.3189	0.0679	-0.2669	-0.0465
November	0.3465	0.0754	-0.3357	-0.0368
December	0.3322	0.1055	-0.2754	-0.0622

of the rainfall season. The correlation of the SOI value for May of the current year and mean annual rainfall is 0.2570, which gave a lead-time of 5 months before the onset of the rainfall season. We also noticed that the highest correlation of mean annual rainfall and SDSLP values is -0.3359 (SDSLP value for August of the current year), which also gave a lead-time of 2 months before the onset of rainfall season. At a lead-time of at least six months before the onset of the rainfall season, the highest correlation is -0.2479 (SDSLP value for April of the current year), which gave a lead-time of 6 months before the onset of the rainfall season. At more than 6 months lead-time, there are no significant correlations between mean annual rainfall and the weather/climatic determinants.

Therefore, we selected the SOI values for August and May and SDSLP anomalies for August and April to be the weather/climate change determinants in this thesis. Since, we aimed to use both SOI and SDSLP values as input variables in our models, we checked for correlation between the input variables. Table 3.7 shows the correla-

tion analysis.

Table 3.7: Correlations between the weather/climate change variables

		SDSLP value for the month of:	
		April	August
SOI value for the month of:	May	-0.3934	-0.4585
	August	-0.3105	-0.7342

Table 3.7 shows that SOI and SDSLP values for August are highly correlated. In Chapter 4, the principal component analysis is used to extract orthogonal explanatory variables for the predictive regression model. We present the summary statistics of the selected weather/climatic variables in Table 3.8.

Table 3.8: Summary statistics for the selected weather/climate change variables

Climatic variable	Mean	Std. dev.	Min	Max	Skewness	Excess kurtosis
SOI_{MAY}	-0.9055	10.0495	-37.4000	21.8000	-0.6361	1.0659
SOI_{AUGUST}	-0.7495	10.2883	-23.6000	34.8000	0.1744	0.1986
$SDSLP_{\text{APRIL}}$	-0.2965	0.8857	-2.1640	1.8255	-0.0357	-0.3877
$SDSLP_{\text{AUGUST}}$	-0.4086	0.8471	-3.2460	1.5596	-0.1092	0.3364

From Table 3.8, the mean of all the selected weather/climate change variables are negative and the skewness values with t-statistics $<\mid 2 \mid$, are not significantly different from zero. This suggest that the weather/climate change variables are possibly symmetric.

3.4 Concluding remarks

In this chapter the data selected to be used in this thesis was discussed. The chapter established that the mean annual rainfall data was stationary, i.i.d. and had no het-

eroscedasticity. The chapter also established that mean annual rainfall for Zimbabwe was correlated with SOI values for May and August of the current year. At a lead time of more than 6 months before the onset of summer rainfall season, there was no significant correlations between mean annual rainfall and SOI values for a particular month. The chapter also established that mean annual rainfall was correlated to SDSLP values of April and August of the current year. Using these weather/climate change variables, we aim to develop simple but reliable models that can be used as an early warning tool for floods or droughts in Zimbabwe. In the following chapter, we develop a simple rainfall predictive model using the selected weather/climate change variables discussed in this chapter.

Chapter 4

Modelling mean annual rainfall using weather/climate change determinants: Weighted regression models

4.1 Introduction

Southern Africa's economic performance in the last decade was disappointing, with much of the region unable to break away from the paths of low or negative per capita income growth and balance of payments difficulties. In economic literature, one of the suggested cause of poor economic growth is persistent droughts that has been experienced in the region considering that southern Africa's economy is highly dependent on agriculture (Jayne et al., 2007). Agriculture is the backbone of the economy in Zimbabwe. It provides employment and income for 60 to 70 percent of the population, supplies 60 percent of the raw materials required by the industrial sector and contributes 40 percent of the total export earnings (Mazvimavi, 2010). Despite agriculture offering high level employment opportunities, it only contributes 20% to the annual Gross Domestic Product (GDP) of the country depending on the rainfall

patterns (Government of Zimbabwe, 2001; Jury, 1996). The contribution of this sector to the economy has not been fully realised. Concerns over economic growth, environmental issues and sustainable development is relatively recent and has captured the attention of researchers, aid agencies and development and environmental planners. This is because sustainable development may equate to sustainable economic growth (Lélé, 1991). Economic sustainability seeks to avoid extreme future imbalances in production (Harris, 2003). With long run economic sustainability, welfare is maximised over time. The interest rests in the need for development programmes to eradicate poverty and food insecurity in most economies (Perman et al., 2003).

Ample theories have been proposed to explain the relatively poor economic growth of Zimbabwe and other sub-Saharan countries (Manatsa et al., 2008; Jury, 1996). In essence the theories can be divided into those arising from politics and those due to exogenous factors. Political explanations usually refer to poor and inconsistent policies that are argued to have impacted negatively on economic growth in Zimbabwe and other sub-Saharan African countries (FAO, 2001; Mangoyana and Meda, 2001; UNECA, 2000). These include poor fiscal and trade policies, lack of good governance, corruption and ill functional financial and labour markets. Exogenous explanations include external aid allocation (Burnside and Dollar, 2004) and a lack of diversification of exports tropical climates including a lack of early drought warning tools (Sachs and Warner, 1997). Given the importance of agriculture to a developing country such as Zimbabwe and the dependence of this sector on rainfall as suggested by Manatsa et al., (2008) and Mangonyana and Meda, (2001), a decline in rainfall and lack of prediction tools can have severe consequences for sustainable growth. Additionally, this decline and lack of prediction tools may have a detrimental impact on the energy supply in Zimbabwe which heavily relies on hydro-power for electricity generation (Kaunda et al., 2012).

Rainfall patterns are affected by various natural phenomena. Firstly, large scale

weather/climatic variation that occur from one year to year (Panu and Sharma, 2002). This is the Southern Oscillation climatic condition, which manifests itself in the differential oceanic temperature phenomenon across the tropical Pacific Ocean. The Southern Oscillation Index (SOI) as defined is the difference between seasonally normalised sea level pressures of Darwin (in Australia) and Tahiti (in the Mid Pacific). Secondly, Darwin Sea Level Pressures (Darwin SLP) have been found to influence seasonal rainfall patterns in Zimbabwe (Manatsa et al, 2008).

Evidence of relationships among meteorological variables is well documented (Rocha, 1992; Ropelewski and Halpert, 1987; Webster, 1981). Most research on rainfall patterns for Zimbabwe have focused on correlations between phases of SOI and rainfall (Torrance, 1990; Matarira and Unganai, 1994; Waylen and Henworth, 1996; Richard et al., 2000). Rocha (1992) found that southeast Zimbabwean rainfall correlated significantly with the SOI (+0.4), with a lead time of four to five months. Using precipitation data from 68 meteorological sites with at least 20 years of complete records, Waylen and Henworth (1996) investigated the association between monthly precipitation totals and the SOI throughout Zimbabwe. Simple lag cross-correlations between the SOI and annual precipitation totals, both annual and monthly, were used to determine significant associations. The researchers revealed positive correlations, with almost 30% of the stations at lag zero. They found a negative correlation at lag -1 (month). The periods with the strongest positive association were the months of the rainy season from October to April, which were correlated to synchronous values of the SOI and those in the preceding June-to-September period. March precipitation was found to be positively correlated with the SOI value for the same month. More than 70% of the stations in the study reported significant correlations between March precipitation and the SOI in the preceding July and at least 25% of the stations reported similar associations with monthly SOI in the preceding May (lag -10) to February (lag -1). The correlation was found to extend through the end of the rainy season into April and May. November and December precipitation were less

strongly correlated with the SOI values in the same period. January and February precipitation were not correlated with monthly SOI values.

Matarira (1990) found that during the warm phase of ENSO, rainfall tends to be low across much of the country, whereas the converse is true for the cold phase. Using an average of the SOI during the preceding 12 months (January to December), Matarira (1990) found a positive correlation of +0.42 with total rainfall for the wet season (November to April). Matarira and Unganai (1994) reported a peak correlation of +0.56 with November to January rainfall in the south eastern part of the country, using the SOI one to two months earlier. They concluded that the SOI can explain up to 30% of the annual variance in summer rainfall in some parts of the country.

In a comparison of SOI anomalies with Zimbabwean seasonal rainfall, Torrance (1990) found that the positive values of the SOI coincided with amounts that were 101% to 125 % of normal. Negative SOI values were generally associated with below-normal rainfall. Torrance's study focused on correlations between SOI values and rainfall. He grouped positive SOI levels into a positive phase, and negative SOI levels into a negative phase. By contrast, this research aims to identify a particular month and lag whose SOI value explains total annual rainfall in Zimbabwe; that is, it seeks to identify the explanatory variable at a time lag of about one year in advance.

Makarau and Jury (1997) used a host of meteorological variables to predict summer rainfall in Zimbabwe. Makarau and Jury (1997) found an association between ENSO phases and extreme rainfall. Specifically, when the SOI is within one standard deviation of the long-term mean, the probability is high that rainfall in Zimbabwe will be within 10% of the mean. Based on a 41-year time series and using the average of August to October SOI values, the correlation between the SOI and Zimbabwean summer rainfall was +0.44. Ismail (1987) proposed an empirical rule from which

the mean seasonal rainfall over Zimbabwe can be predicted three months before the start of the rainy season and ten months before its end using SOI. Ismail (1987) concluded that SOI has an influence on the seasonal rainfall over Zimbabwe. Manatsa et al. (2008) used correlation analysis to identify the lag periods for which SOI and Darwin pressure anomalies are significantly correlated with the Zimbabwean Summer Precipitation Index. They concluded that progressive lagged four months averaged Darwin sea level pressure anomalies were correlated with the Zimbabwean precipitation index. Our work advances the work of Manatsa et al. (2008) by trying to propose a simple early warning rainfall predictive model using climatic determinants such as SOI value and SDSLP anomalies for Zimbabwe, at a longer lead time before the onset of the rainfall season. Annual rainfall patterns are crucial for agriculture, water management, hydro-power electricity generation and infrastructure design. Any meaningful planning requires information based on the rainfall patterns of the crucial months of the rainfall season. We are not aware of any literature relating to modelling mean annual rainfall for Zimbabwe using SDSLP and SOI values.

The rest of the chapter is organised as follows: Section 4.2 presents the research methodology. Section 4.3 presents the empirical results and discussion of the findings. Section 4.4 provides the concluding remarks. Finally, Section 4.5 and Section 4.6 (Appendix) presents the concluding remarks and the correlograms for residual analysis respectively.

4.2 Research methodology

In this chapter, dependent mean annual rainfall variable is expressed in terms of independent explanatory variables; SDSLP anomalies and SOI. Multiple linear regression can be used to model a relationship between the dependent variable and the explanatory variables. It allows for investigating the effect of changes in the various factors on the dependent variable. If the observations are measured over time, the model becomes a time series regression model. The resulting statistical relation-

ship can be used to predict values of rainfall. To ascertain the predictive power of the model, all assumptions of multiple linear regressions must be met.

4.2.1 General linear model

Probabilistic models that include more than one independent variables are called multiple regression. The model can be written as:

$$y_t = \beta_0 + \beta_1 x_{1,t} + \beta_{2,t} x_{2,t} + ... + \beta_p x_{p,t} + \epsilon_t, \quad t = 1, 2, ..., n \tag{4.1}$$

where y_t is the t^{th} dependent variable and $x_{j,t}$ is the corresponding observation of the explanatory variable whose predictive influence is of interest i.e. SOI and SD-SLP values. Parameters β_t are unknown and the probabilistic component of the model ϵ_t is the unknown error term. The value of the coefficient β_t determines the contribution of the independent variable $x_{j,t}$ given that the other independent variables are held constant. Using classical estimation techniques estimates for the unknown parameters are obtained. If the estimated values for $\beta_0, \beta_1, ..., \beta_p$ are given by $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_p$ then the dependent variable is estimated as:

$$\hat{y}_t = \hat{\beta}_0 + \hat{\beta}_1 x_{1,t} + \hat{\beta}_{2,t} x_{2,t} + ... + \hat{\beta}_p x_{p,t}, \quad t = 1, 2, ..., n \tag{4.2}$$

and the estimate $\hat{\epsilon}_t$ for the error term ϵ_t is determined as the difference between the observed and the predicted dependent variable; $\hat{\epsilon}_t = y_t - \hat{y}_t$.
The model (4.1) can be rewritten as:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{4.3}$$

where

$$\mathbf{y} = [y_1\ y_2 ...\ y_n]', \mathbf{X} = \begin{bmatrix} x_{11} & x_{21} & . & . & . & x_{p1} \\ x_{12} & x_{22} & . & . & . & x_{p2} \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ x_{1n} & x_{2n} & . & . & . & x_{pn} \end{bmatrix}, \boldsymbol{\beta} = [\beta_0\ \beta_1\ ...\ \beta_{p-1}]',$$

$\boldsymbol{\epsilon} = [\epsilon_1\ \epsilon_2\ ...\ \epsilon_n]'$,

and where the assumption that $y_1, y_2, ..., y_n$ is a random sample is equivalent to the assumption that $E(\boldsymbol{\epsilon}) = 0$ and $\text{var}(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}_n$. We shall also assume that the x_{tj}'s are chosen in such a way that the $\text{rank}(\mathbf{X}) \leq p$.

We can estimate the parameters $\boldsymbol{\beta}$ by the least squares method. The least squares estimator $\hat{\boldsymbol{\beta}}$ is such that $\sum_{t=1}^{n} \hat{\epsilon}_t^2 = \sum_{t=1}^{n}(y_t - \hat{y}_t)^2$ is a minimum. Thus $\hat{\boldsymbol{\epsilon}}'\hat{\boldsymbol{\epsilon}} = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$ must be minimised.

Theorem 4.1 *Let $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}+\boldsymbol{\epsilon}$ where $\mathbf{X} : n\times p$ is of full rank, $\boldsymbol{\beta} : p\times 1$ is a vector of unknown parameters, and $\boldsymbol{\epsilon} : n\times 1$ is a random vector with mean zero and $\text{cov}(\boldsymbol{\epsilon}, \boldsymbol{\epsilon}') = \text{var}(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}_n$. The least squares estimator for $\boldsymbol{\beta}$ is denoted by $\hat{\boldsymbol{\beta}}$ and is given by:*

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}. \tag{4.4}$$

In the theoretical model several assumptions are made about the explanatory variables and the error term. Firstly there must be insignificant correlation between the explanatory variables. When the explanatory variables are correlated, multicollinearity problems arise. The estimated parameters will be unstable and unreliable if highly correlated variables are used in the model as explanatory variables. In the study present, the predictive power of SOI and SDSLP values at a maximum lag is important. We aim to investigate the influence of SOI values (for the month of May and August) and SDSLP values (for the month of August and April) on mean annual rainfall. As shown in Section 3.3, SOI values for August were highly correlated to SDSLP anomalies for August, principal component analysis is used to produce

orthogonal explanatory variables that can be used in a general linear model.

4.2.2 Principal component analysis

Principal component analysis (PCA) also known as empirical orthogonal function has been used in many different disciplines including finance, agriculture, biology, chemistry, climatology, demography, ecology, psychology and meteorology. PCA is a technique used to combine highly correlated factors into principal components that are much less highly correlated with each other. This improves the efficiency of the model.

In this chapter, the predictive power of SOI values (I_1) and SDSLP values (I_2) is explored. Two new, uncorrelated factors, I(I_1^*)and (I_2^*), can be constructed as follows:
Let $I_1^* = I_1$

Then, we carry out a linear regression analysis to determine the parameters γ_1 and γ_2 in the equation:

$$I_2 = \gamma_1 + \gamma_2 I_1^* + \epsilon_i^* \tag{4.5}$$

γ_1 and γ_2 are the intercept and slope parameters of the regression model respectively and ϵ_i^* is the 'error' term, which by definition is independent of $I_1^* = I_1$.
We then set:

$$I_2^* = \epsilon_i^* = I_2 - (\gamma_1 + \gamma_2 I_1^*) \tag{4.6}$$

By construction I_2^* is uncorrelated with SOI values (I_1) since $I_2^* = \epsilon_i^*$, the residual term in the equation. Changes in I_2^* is interpreted as the change in the observed values of SDSLP (I_2) that cannot be explained by the observed change in SOI (I_1). I_2^* in the rainfall model (4.1) explains the component of rainfall that cannot be explained by the SOI.

The other assumptions of the rainfall model (4.1) are that there is no serial correlation and heteroscedasticity of error terms. These assumptions are likely to be violated in regression models with time series data. Autocorrelation (the error terms being correlated among themselves through time) leads to regression coefficients which are biased, inefficient and the standard errors are probably wrong making t and F tests unreliable. In a regression with auto-correlated errors, the errors will probably contain information that is not captured by the explanatory variables. The Durbin-Watson test is used to assess whether the residuals are significantly correlated. A Durbin-Watson statistic of 2 indicates the absence of autocorrelation. The autocorrelation function and partial autocorrelation function can also be used to detect autocorrelation among the residuals.

4.2.3 Weighted least squares model

The least squares method assumes that each data point provides equally precise information about the deterministic part of the total process variation i.e. the standard deviation of the error term is constant over all values of the explanatory variables. This assumption, however, clearly does not hold in all instances, even approximately, in every modelling application. In situations when it may not be reasonable to assume that every observation should be treated equally, weighed least squares weighs some observations more heavily than others, giving each data point its proper amount of influence over the parameter estimates. This maximises the efficiency of parameter estimation. Weighted least square reflects the behaviour of the random errors in the model.

To find the weighted least square parameters of the weighted model, we minimise the weighted sum of squared errors (WSSE):

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

where $w_t > 0$ is the weight assigned to the t^{th} observation. The weight w_t can be 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. The estimates are:

$$\hat{\boldsymbol{\beta}}^* = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{y}. \tag{4.7}$$

where $\mathbf{W} = [w_1, w_2, ..., w_n]$ is the weight vector.

The biggest disadvantage of weighted least squares is the fact that the theory behind this method is based on the assumption that the exact weights are known. This is almost never the case in real applications where estimated weights are used (Carrol and Ruppert, 1988). In this chapter, we aim to estimate reliable weights for modelling mean annual rainfall using SOI and SDSLP values.

4.3 The models

The data for this chapter is the mean annual rainfall data and selected SOI_{May}, $\text{SOI}_{\text{August}}$, $\text{SDSLP}_{\text{April}}$ and $\text{SDSLP}_{\text{August}}$ discussed in Chapter 3. It was shown in Chapter 3 that $\text{SOI}_{\text{August}}$ was highly correlated to $\text{SDSLP}_{\text{August}}$. We set $\text{SOI}_{\text{August}} = I_1 = I_1^*$ and $\text{SDSLP}_{\text{August}} = I_2$ and ran a regression $I_2 = \gamma_1 + \gamma_2 I_1^* + \epsilon_i^*$. We extracted the residuals ϵ_i^* and standardised them. The extracted standardised residuals denoted by I_2^* are orthogonal (not correlated to $\text{SOI}_{\text{August}} = I_1^*$. Therefore, we propose the following rainfall predictive models.

(i) $\hat{y}_t = \hat{\beta}_0 + \hat{\beta}_1 \text{SOI}_{\text{August}} + \hat{\beta}_2 I_2^*$ (to be referred to as Model 1)

(ii) $\hat{y}_t = \hat{\beta}_0 + \hat{\beta}_1 \text{SOI}_{\text{August}} + \hat{\beta}_2 \text{SDSLP}_{\text{April}}$ (to be referred to as Model 2)

(iii) $\hat{y}_t = \hat{\beta}_0 + \hat{\beta}_1 \text{SOI}_{\text{May}} + \hat{\beta}_2 \text{SDSLP}_{\text{April}}$ (to be referred to as Model 3)

where $\hat{y}_t$ is the estimated rainfall amount at time period t.

4.3.1 Assessing model performance

To evaluate the performance of the considered models, we applied the measures of average error namely, mean absolute percentage error (MAPE) and root mean square error (RMSE) discussed in Chapter 2.

4.3.2 Model selection criterion

The scope of model selection is to identify the model that is better suited to predict mean annual rainfall using SOI and SDSLP values. The AIC discussed in Chapter 2 is used for model selection.

4.4 Empirical results

In this section, results from general regression and weighted regression models are also presented.

We investigate the influence of the climatic variables using multiple regression approach. The $\text{SOI}_{\text{August}} = I_1^*$ and $\text{SDSLP}_{\text{August}} = I_2$ are highly correlated as shown in Chapter 3. We extract the residuals I_2^* of the simple regression of $\text{SOI}_{\text{August}}$ as the dependent variable and $\text{SDSLP}_{\text{August}}$ as the explanatory variable. The estimated simple regression model is:

$$\hat{\text{SOI}}_{\text{August}} = -4.3896 - 8.9086 \text{SDSLP}_{\text{August}},$$

with $R^2 = 0.5337$. The simple model is significant. The extracted residuals I_2^* are then used as one of the explanatory variables in multiple regression model. Table 4.1 shows the results of the multiple regression approach for Models 1, 2 and 3.

From Table 4.1, the models parameter estimates are significant at 10% level of sig-

Table 4.1: Parameter estimates for regression models (standard errors in brackets)

Model	$\hat{\beta}_0$ $\hat{\beta}_1$ $\hat{\beta}_2$	AIC	Adjusted R^2
1	$664.9689^{**}(14.8507)$ $6.7211^{**}(1.9672)$ $0.7234^{**}(2.8943)$	12.9426	0.1690
2	$657.5519^{**}(15.6949)$ $6.4106^{**}(1.5041)$ $-24.2283^{*}(17.4727)$	12.9252	0.1834
3	$653.0067^{**}(16.8821)$ $3.1855^{**}(1.6948)$ $-33.0797(19.2309)$	13.0506	0.0741

*Note: ** and * indicates significant at 5% and 10% level of significance*

nificance except $\hat{\beta}_2$ in Model 3. Based on the AIC values, the best model is Model 2. Adjusted $R^2 = 0.1834$ which indicates that the model only explains 18.34% of the variations in mean annual rainfall. The Durbin-Watson statistic for Model 2 is 2.3173. This indicate that the model does not violate the assumption of serial correlation of the residuals. However, the selected model has low forecasting power, thus we can improve the models by using different weights in the regression model. Various weights are considered in arriving at estimates using the weighted regression. A bigger class of models was however considered and the better models are listed in Table 4.2. The parameter estimates, AIC values and statistics for checking model adequacy for some of the weighted regression models are reported in Table 4.2.

We use the MAPE and RMSE measures to assess the forecasting performance of the models and the AIC to select the best fitting model. The weights are assumed to be proportional to the inverse standard deviation. Model 1 with $(I_2^*)^2$ as the weight is selected to be the best predictive model, since it has the least AIC value. The model has the least MAPE and RMSE values indicating that it performs better than the other models in forecasting mean annual rainfall. The model is also significant at

Table 4.2: Parameter estimates for weighted regression models (standard errors in brackets)

Model	Weight with a time lag	$\hat{\beta}_0$ $\hat{\beta}_1$ $\hat{\beta}_2$	AIC	Adjusted R^2	MAPE	RMSE
1	I_2^*	519.1459(67.3999) 6.8718(3.2014) 15.87054(7.7761)	16.9549	0.3786	28.6553	236.0212
1^*	$(I_2^*)^2$	716.5207(14.8638) 5.3913(1.0445) 5.4111(0.9335)	12.3958	0.6629	22.9341	162.9557
2	$(\text{SDSLP}_{\text{APRIL}})^2$	660.1174(18.0899) 6.0862(1.1068) $-7.5155(10.6878)$	13.6137	0.2665	29.6337	163.4481
3	$(\text{SDSLP}_{\text{APRIL}})^3$	838.5388(89.1952) $-0.6079(2.6977)$ $-239.6288(72.1891)$	12.4650	0.1734	50.2127	346.2853

Note: All the parameters are significant at 5% level of significance

5% significance level. The multiple adjusted $R^2 = 0.6629$, indicates that the model explains 66.3% of the variations in mean annual rainfall from just two variables. SOI explains only approximately 30% of the rainfall variability, which means that other factors should also be taken into account when predicting rainfall. Figure 4.1 shows the box plot and QQ-plot of selected model's residuals.

From Figure 4.1 it appears that the residuals do not violate the assumption of normality. The box plot seems to be symmetric and there is no serious departure from the reference line of the QQ-plot. The ACF and PACF correlogram (see Section 4.6) shows that the residuals are independent. This is confirmed by the Durbin-Watson statistic value equal to 2.1616, indicating that the residuals are approximately independent. The correlogram of squared residuals (see Appendix 4.1) shows that the model does not violate the assumption of homoscedasticity. This is confirmed by the White test, which produce an F-statistic of 19.7600 (p-value$= 0.0000 < 0.05$). It is

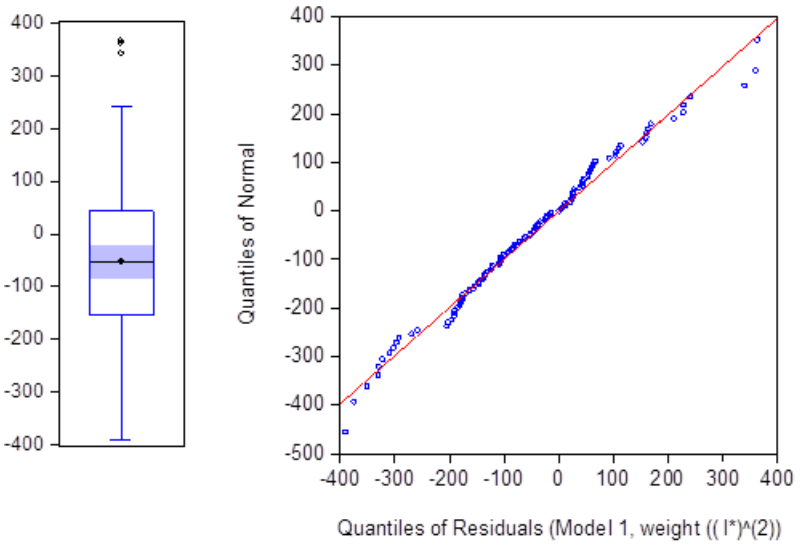

Figure 4.1: Box and QQ-plots of residuals of the selected model (Model 1)

important to check the in-sample forecasting power of the model. Figure 4.2 shows the plot of observed mean annual rainfall against predicted rainfall from the selected model.

From Figure 4.2 the model seems to be able to predict in-sample mean annual rainfall. The model seems to show little variability between predicted and actual rainfall. The model can be used to predict mean annual rainfall for Zimbabwe. The out-of-sample forecasts for the years 2010 to 2015 are shown in Table 4.3.

The out-of-sample forecasts seem to be reasonable. The MAPE value of 0.2018 and RMSE value of 141.65 are not so large considering that only two climatic explanatory variables have been used in the model. The model however seems to slightly over-forecast the mean rainfall. This suggests that the proposed model is reasonable and can be used as an early warning tool to inform the government, farmers and other stake-holders about the upcoming rainfall season. The model may still need improvement and can be complimented by other models.

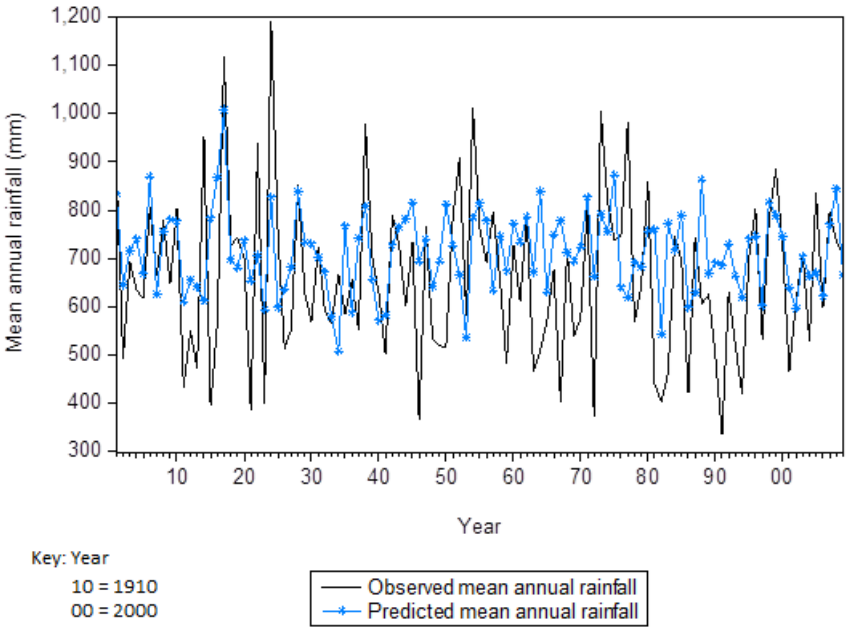

Figure 4.2: Mean annual rainfall versus predicted rainfall

4.5 Concluding remarks

We proposed a simple annual rainfall predicting tool for Zimbabwe using SOI and SDSLP values. The simple model can be used as part of a drought early warning system. This chapter's main finding is that annual rainfall for Zimbabwe correlates with SOI and SDSLP values of the month of August. We employed principal component analysis to construct orthogonal factors (non collinear variables) since SOI and SDSLP values are highly correlated. The combination of regression and time series analysis offers a powerful tool for predicting annual rainfall using SOI and SDSLP values of a particular month with a lag of at least 2 months. Developing a simple model for annual rainfall prediction provides important drought forecasts so that precautionary measures can be taken. These measures may involve an adjustment in the national budget expenditure in order to carter for the forthcoming drought, deciding on the type of seeds to plant and other possible measures of saving water as a scarce resource. The following measures may be taken based on the research results: (a) adaptive measures, thus create standby measures to deal with climate

Table 4.3: Out-of-sample forecasts

Year	Actual Mean annual rainfall (mm)	Predicted annual rainfall (mm)
2010	681.6	874.5316
2011	676.5	765.9985
2012	601.7	709.5376
2013	618.3	717.7716
2014	554.0	664.5861
2015	421.4	590.4050

related disasters, (b) shifting agricultural activities to be in line with the amount of anticipated rain. With regard to the data, it is clear that the explanatory variables incorporated in the model are limited. It would be interesting to include other climatic determinants such as Sea Surface Temperatures at Darwin and wind speed. However, the use of weighted regression gives an acceptable fit in the absence of these other factors. Extending the model with more factors may provide a better understanding of the rainfall patterns in Zimbabwe and is an area for further research.

4.6 Appendix

Diagnostic plots of Model 1

Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
*|. |	*|. |	1	-0.099	-0.099	1.0894	0.297
.|* |	.|* |	2	0.088	0.079	1.9658	0.374
.|* |	.|* |	3	0.111	0.129	3.3688	0.338
.|. |	.|. |	4	0.001	0.017	3.3689	0.498
.|* |	.|* |	5	0.133	0.117	5.4181	0.367
.|. |	.|. |	6	-0.048	-0.041	5.6941	0.458
.|* |	.|* |	7	0.120	0.092	7.3974	0.389
*|. |	*|. |	8	-0.089	-0.096	8.3492	0.400
*|. |	*|. |	9	-0.135	-0.172	10.560	0.307
.|* |	.|. |	10	0.098	0.048	11.724	0.304
*|. |	.|. |	11	-0.108	-0.048	13.169	0.282
.|. |	.|. |	12	0.003	-0.011	13.170	0.357
*|. |	.|. |	13	-0.073	-0.047	13.849	0.385
.|. |	.|. |	14	-0.020	0.004	13.903	0.457
.|. |	.|. |	15	0.034	0.048	14.054	0.521
.|. |	.|. |	16	-0.016	0.054	14.089	0.592
.|* |	.|* |	17	0.156	0.129	17.308	0.434
*|. |	*|. |	18	-0.127	-0.101	19.448	0.365
.|* |	.|* |	19	0.137	0.113	21.967	0.286
.|. |	.|. |	20	0.022	0.001	22.035	0.339
.|. |	.|. |	21	-0.012	-0.021	22.056	0.396
.|* |	.|. |	22	0.127	0.046	24.299	0.332
.|. |	.|. |	23	0.009	0.049	24.310	0.387
.|. |	.|. |	24	0.016	-0.043	24.345	0.442
.|. |	.|. |	25	0.039	0.061	24.559	0.487
.|. |	.|. |	26	-0.031	-0.043	24.701	0.536
.|. |	*|. |	27	-0.024	-0.096	24.784	0.587
.|. |	.|. |	28	0.003	0.073	24.786	0.639
.|. |	.|. |	29	0.023	0.008	24.865	0.685
*|. |	*|. |	30	-0.099	-0.087	26.359	0.657
*|. |	.|. |	31	-0.073	-0.057	27.191	0.663
.|. |	.|. |	32	0.008	0.005	27.202	0.708
.|. |	.|. |	33	-0.017	0.024	27.249	0.749
.|. |	.|. |	34	0.014	0.053	27.281	0.786
.|. |	.|. |	35	0.000	0.028	27.281	0.821
.|. |	*|. |	36	-0.028	-0.072	27.407	0.848

Figure 4.3: ACF and PACF correlogram of residuals from the best fitting Model 1

Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
.|* |	.|* |	1	0.154	0.154	2.6469	0.104
.|. |	.|. |	2	0.053	0.030	2.9678	0.227
.|. |	.|. |	3	-0.010	-0.023	2.9792	0.395
*|. |	*|. |	4	-0.097	-0.096	4.0622	0.398
.|. |	.|. |	5	0.007	0.038	4.0683	0.540
*|. |	*|. |	6	-0.083	-0.084	4.8822	0.559
.|. |	.|. |	7	0.033	0.056	5.0076	0.659
.|* |	.|* |	8	0.100	0.088	6.1961	0.625
.|. |	.|. |	9	0.006	-0.024	6.1998	0.720
.|* |	.|. |	10	0.089	0.071	7.1710	0.709
.|. |	.|. |	11	-0.046	-0.059	7.4294	0.763
*|. |	*|. |	12	-0.100	-0.088	8.6766	0.730
*|. |	.|. |	13	-0.082	-0.051	9.5138	0.733
*|. |	.|. |	14	-0.105	-0.053	10.921	0.692
*|. |	*|. |	15	-0.114	-0.117	12.598	0.633
*|. |	.|. |	16	-0.067	-0.037	13.175	0.660
.|* |	.|* |	17	0.105	0.122	14.638	0.622
.|* |	.|* |	18	0.145	0.089	17.444	0.493
.|. |	.|. |	19	0.020	-0.029	17.497	0.556
.|. |	.|. |	20	0.001	-0.006	17.497	0.621
*|. |	.|. |	21	-0.083	-0.064	18.437	0.621
.|. |	.|. |	22	0.004	0.059	18.440	0.680
.|. |	.|. |	23	-0.061	-0.039	18.963	0.703
.|. |	.|. |	24	-0.006	0.018	18.969	0.754
.|. |	*|. |	25	-0.022	-0.068	19.041	0.795
.|. |	*|. |	26	-0.058	-0.076	19.536	0.813
.|. |	*|. |	27	-0.038	-0.088	19.750	0.841
.|. |	.|. |	28	-0.011	-0.010	19.769	0.873
.|. |	.|. |	29	-0.036	-0.024	19.962	0.894
.|. |	.|. |	30	0.008	0.022	19.972	0.917
.|. |	.|. |	31	-0.039	-0.020	20.204	0.931
.|. |	.|. |	32	-0.052	-0.037	20.628	0.939
.|. |	.|. |	33	-0.011	0.040	20.649	0.954
.|. |	.|. |	34	-0.035	-0.027	20.852	0.962
.|. |	.|. |	35	-0.021	-0.057	20.924	0.971
.|. |	.|. |	36	-0.018	-0.035	20.977	0.978

Figure 4.4: ACF and PACF correlogram of squared residuals from the best fitting Model 1

Chapter 5

Extreme rainfall: Candidature probability distributions for mean annual rainfall data: An application to Zimbabwean data

5.1 Introduction

Dry spells, of varying severities, are regular occurrences in Zimbabwe usually resulting in drought. Drought is a calamity with severe impacts on society. It contributes to the loss of crops, animals and valuable property. Although knowledge of rainfall patterns over an area may be used for such disaster prevention purposes, it is one of the most difficult meteorological parameters to study because of lack of reliable data and large variations of rainfall in space and time. Developing methods that can provide a suitable prediction of meteorological events is always interesting for both meteorologists and statisticians, because of its role in infrastructure development, water resource management and agriculture. Knowledge of rainfall characteristics, its temporal and spatial distribution plays a major role in drought-prone southern Africa, where economies are mainly driven by rain-fed agriculture (Jury, 1996). To

improve on modelling rainfall processes, many researches have been searching for physical and statistical properties of rainfall using observational data. One area of interest is the parent probability distribution of rainfall amount (Deka et al., 2009; Cho et al., 2004). Dan'azumi et al. (2010), Husak et al. (2007), Suhaila and Jemain (2007) and Martins and Stedinger (2000) provide the most recent research on mathematical modelling of rainfall patterns. It is generally assumed that a meteorological variable follows a certain probability distribution. Many probability distributions have been considered, in many different situations. These distributions are gamma distribution (Stagge et al., 2015; Husak et al., 2007; Cho et al., 2004; Aksoy, 2000; Adiku et al., 1997; McKee at al., 1993), lognormal distribution (Suhaila et al., 2011; Deka et al., 2009; Cho et al., 2004), the generalised extreme value distribution (GEVD) (Roth et al., 2014; Coles et al., 2001; Aksoy, 2000; Martins and Stedinger, 2000; Bulu and Aksoy, 1998; Madsen et al., 1997;) and the log-logistic distribution (Fitzgerald, 2005; Ahmad et al., 1988).

Rainfall data have been modelled by several researchers from different regions of the world. Koutsoyiannis and Baloutsos (2000) analysed rainfall data from Greece. Sakulski et al. (2014) fitted the six probability distributions, namely; the log-logistic, Singh-Maddala, lognormal, generalised extreme value, Fréchet and Rayleigh distributions to spring, summer, autumn and winter rainfall data from the Eastern Cape province, South Africa. They found that the Singh-Maddala distribution was the best fitting distribution for all seasons' rainfall data. Rakhecha and Soman (1994) analysed the annual rainfall series from India covering over 80-years of rainfall data. Stagge et al. (2015) fitted seven candidate distributions to the standardised precipitation index (SPI) and the standardized evapo-transpiration index (SPEI) for Europe and recommended the two-parameter gamma distribution for modelling SPI and GEVD for modelling SPEI. Suhaila and Jemain (2007) found the mixed Weibull distribution to be the best fitting distribution over single distributions in modelling rainfall amounts in Peninsular Malaysia. Suhaila and Jemain (2007) found the gen-

eralised lambda distribution as the best fitting distribution for rainfall amounts in Peninsular Malaysia as well. These results differ from the results obtained by Zin et al. (2009). Therefore, each kind of probability distribution has its own applicability and limitations. A regionalised study on statistical modelling of annual rainfall is very essential as the statistical properties may vary according to the geographical location of the area considered and the length of the rainfall data series. In this chapter, an attempt is made to fit the gamma, lognormal and log-logistic distributions to mean annual rainfall data series for Zimbabwe. These distributions are referred to as parent distributions since they fit the main body of data. However, the tail of the theoretical parent distributions sometimes diverges in the extreme minima or maxima rainfall data region. Extreme value theory is an alternative to fit minima and maxima mean annual rainfall (Chikobvu and Chifurira, 2015). Berger et al., (1982) and Surman et al., (1987) showed that the two-parameter exponential distribution fits extreme weather and atmospheric data well. Therefore, the purpose of this Chapter is to: (i) compare the relative performance of the gamma, lognormal, log-logistic distributions and the two-parameter exponential distributions in fitting the mean annual rainfall for Zimbabwe, (ii) investigate the performance of the candidate distributions at 25, 50 and 75-year periods, (iii) select the most robust model using goodness-of-fit tests namely relative root mean square error (RRMSE), relative mean absolute error (RMAE) and probability plot correlation coefficient (PPCC) as discussed in Chapter 2 and (iv) estimate the mean return period for specific return levels. It should be noted that the normal distribution was considered/assumed in chapters 2, 3 and 4.

The rest of the chapter is organized as follows: Section 5.2 presents the research methodology. Section 5.3 presents the empirical results and discussion of the findings. Finally, Section 5.4 provides the concluding remarks.

5.2 Research methodology

In this section, we present some background theory on theoretical distributions namely; the two-parameter gamma, two-parameter lognormal, the two-parameter log-logistic and the two-parameter exponential distribution fitted to the mean annual rainfall data described in Chapter 2. The parameters are estimated by method of maximum likelihood. The procedure always gives the minimum variance estimate of parameters.

5.2.1 Two-parameter gamma distribution

The two-parameter gamma distribution is recommended for hydrological/meteorological frequency analysis (McKee et al., 1993; Hosking, 1990). The two-parameter gamma distribution is defined by the density function:

$$f_{\mathrm{g}}(x) = \frac{x^{\xi_{\mathrm{g}}-1}}{\sigma_{\mathrm{g}}^{\xi_{\mathrm{g}}}\Gamma(\xi_{\mathrm{g}})}\exp\left(-\frac{x}{\sigma_{\mathrm{g}}}\right), \quad x \geq 0 \tag{5.1}$$

where σ_{g}, $\xi_{\mathrm{g}} > 0$ and x represent mean annual rainfall. σ_{g} is the scale parameter and ξ_{g} is the shape parameter of the gamma distribution. The quantile function of the gamma distribution has no explicit form. In this thesis, we estimate the quantiles for the gamma distribution using R package "fitdistrplus".

Parameter estimation of the two-parameter gamma distribution

Under the assumption that the observed n independent data points $X_1, ..., X_n$ have a gamma distribution, the log-likelihood for the two-parameter gamma distribution is

$$l(\sigma_{\mathrm{g}}, \xi_{\mathrm{g}}) = (\xi_{\mathrm{g}} - 1)\sum_{i=1}^{n}\ln(x_i) - n\ln\Gamma(\xi_{\mathrm{g}}) - n\xi_{\mathrm{g}}\ln(\sigma_{\mathrm{g}}) - \frac{1}{\sigma_{\mathrm{g}}}\sum_{i=1}^{n}x_i \tag{5.2}$$

The estimate of σ_{g} is found to be

$$\hat{\sigma_{\mathrm{g}}} = \frac{\overline{x}}{\xi_{\mathrm{g}}} \tag{5.3}$$

and substituting (5.3) into the log-likelihood gives

$$l(\xi_{\mathrm{g}}) = n(\xi_{\mathrm{g}} - 1)\sum_{i=1}^{n}\overline{\ln(x_i)} - n\ln\Gamma(\xi_{\mathrm{g}}) - n\xi_{\mathrm{g}}\ln(\overline{x}) + n\xi_{\mathrm{g}}\ln(\xi_{\mathrm{g}}) - n\xi_{\mathrm{g}} \tag{5.4}$$

which is maximised via the generalised Newton algorithm to obtain the estimate of ξ_{g}. For details of the algorithm see Minka (2002). The estimate of ξ_{g} is given by

$$\hat{\xi_{\mathrm{g}}} \approx \frac{0.5}{\ln\overline{x} - \overline{\ln x}} \tag{5.5}$$

where $\ln\overline{x} > \overline{\ln x}$.

5.2.2 Two-parameter lognormal distribution

Another distribution that is commonly used to model rainfall amounts is the two-parameter lognormal distribution (Cho et al., 2004). The two-parameter lognormal distribution is similar in appearance to the gamma distribution. An assumption of the lognormal distribution is that the logarithms of the data are normally distributed. The two-parameter lognormal distribution is defined by the density function

$$f_{\mathrm{l}}(x) = \frac{1}{x\xi_{\mathrm{l}}\sqrt{2\pi}}\exp\left[-\frac{(\ln(x) - \sigma_{\mathrm{l}})^2}{2\xi_{\mathrm{l}}^2}\right], \quad x > 0 \tag{5.6}$$

where $-\infty < \sigma_{\mathrm{l}} < \infty$ and $\xi_{\mathrm{l}} > 0$. σ_{l} and ξ_{l} are the scale and shape parameters of the lognormal density function respectively. Due to a close relationship with the normal distribution, the scale parameter σ_{l}, may be interpreted as the mean of the logarithm of the random variable, while the shape parameter ξ_{l}, maybe interpreted as the standard deviation of the logarithmically transformed variables. Modelling with the lognormal distribution allows the use of normal-theory statistics on a logarithmic scale, and parameter estimation is then straightforward (Manning and Mullahy, 2001).

The quantile function of the lognormal distribution is given by

$$Q(p) = \exp(\sigma_{\text{l}} + \xi_{\text{l}}\Phi^{-1}(p)) \tag{5.7}$$

where $\Phi^{-1}(.)$ has a standard normal distribution and $0 < p < 1$.

Parameter estimation of the two-parameter lognormal distribution

Under the assumption that the observed n independent data points $X_1, ..., X_n$ have a lognormal distribution, the log-likelihood for the two-parameter lognormal distribution is derived by taking the product of the probability densities of the individual X_is:

$$l(\sigma_{\text{l}}, \xi_{\text{l}}) = \ln\left((2\pi\xi_{\text{l}}^2)^{-\frac{n}{2}} \prod_{i=1}^{n} X_i^{-1} \exp\left[\frac{(\ln(x_i) - \sigma_{\text{l}})^2}{\xi_{\text{l}}}\right]\right) \tag{5.8}$$

By maximising the log-likelihood the estimates of σ_{l} and ξ_{l} are obtained. The maximum likelihood parameter estimates are:

$$\hat{\sigma_{\text{l}}} = \frac{\sum_{i=1}^{n} \ln(x_i)}{n} \tag{5.9}$$

and

$$\hat{\xi_{\text{l}}} = \frac{\sum_{i=1}^{n} \left(\ln(x_i) - \frac{\sum_{i=1}^{n} \ln(x_i)}{n}\right)^2}{n} \tag{5.10}$$

5.2.3 Two-parameter log-logistic distribution

The log-logistic distribution is related to the logistic distribution in the same manner as the lognormal distribution is related to the normal distribution. A logarithmic transformation of the logistic distribution generates the log-logistic distribution. The

log-logistic distribution is defined by the density function:

$$f_{\text{ll}}(x) = \frac{1}{\left[1 + \left(\frac{x}{\sigma_{\text{ll}}}\right)^{\xi_{\text{ll}}}\right]^2} \left(\frac{\xi_{\text{ll}}}{\sigma_{\text{ll}}}\right) \left(\frac{x}{\sigma_{\text{ll}}}\right)^{\xi_{\text{ll}}-1}, \quad x > 0 \tag{5.11}$$

where $\sigma_{\text{ll}} > 0$ and $\xi_{\text{ll}} > 0$. σ_{ll} and ξ_{ll} are the scale parameter and shape parameter of the log-logistic distribution respectively. The log-logistic distribution has different shapes, it can be strictly decreasing, right skewed, or unimodal. This property of flexibility enables the log-logistic distribution to fit data from many different fields, including engineering, economics, hydrology/meteorology and survival analysis.

The quantile function of the log-logistic distribution is given by:

$$Q(F_{\text{ll}}) = \sigma_{\text{ll}} \left(\frac{p}{1-p}\right)^{\frac{1}{\xi_{\text{ll}}}} \tag{5.12}$$

Parameter estimation of the log-logistic distribution

Under the assumption that n observations, denoted by $X_1, ..., X_n$ are from a log-logistic distribution, the log-likelihood function is:

$$l(\sigma_{\text{ll}}, \xi_{\text{ll}}) = n\ln(\xi_{\text{ll}}) - n\xi_{\text{ll}}\ln(\sigma_{\text{ll}}) + (\xi_{\text{ll}} - 1)\sum_{i=1}^{n}\ln(x_i) - 2\sum_{i=1}^{n}\ln\left[1 + \left(\frac{x_i}{\sigma_{\text{ll}}}\right)^{\sigma_{\text{ll}}}\right] \tag{5.13}$$

The parameter estimates are obtained by differentiating the log-likelihood function with respect to σ_{ll} and ξ_{ll}, and equating to zero.

In this thesis, we use the R package "fitdistrplus" to obtain the maximum likelihood parameter estimates of the gamma, lognormal and log-logistic distributions.

5.2.4 Two-parameter exponential distribution derived from extreme value theory

Extreme value theory has emerged as one of the most important statistical disciplines for meteorological sciences over the last 50 years (Li et al., 2015). The two-parameter exponential distribution is recommended to model extreme events or heavy tailed data (Lu, 2004). A random variable X is said to have a two-parameter exponential distribution ($\mathrm{Exp}(\mu_\mathrm{e}, \sigma_\mathrm{e})$) if its probability density function is of the form

$$f_{rme}(x) = \frac{1}{\sigma_\mathrm{e}} \exp\left[-\left(\frac{x-\mu_\mathrm{e}}{\sigma_\mathrm{e}}\right)\right], \quad -\infty < x < \infty, \sigma_\mathrm{e} > 0, \tag{5.14}$$

where μ_e is the location parameter and σ_e is the scale parameter. Berger et al. (1982) derived the two-parameter exponential distribution, F_e, from extreme value theory to represent the cumulative frequency distribution of maxima random variable over a specific percentile,

$$F_\mathrm{e} = 1 - \exp[-y] \tag{5.15}$$

where $y = \frac{x-\mu_\mathrm{e}}{\sigma_\mathrm{e}}$. The mean annual rainfall below a specific drought threshold was chosen from the complete set of mean annual rainfall to fit the two-parameter exponential distribution. The estimated cumulative probability $\hat{F}_\mathrm{e}$ can be calculated using

$$\hat{F}_\mathrm{e} = \frac{N-r+1}{N+1} = P_i \tag{5.16}$$

where N is the size of the chosen maxima mean annual rainfall. P_i is the probability of a value that is ranked i out of N values. Therefore, the relationship between the variate y and P_i is given by

$$y(i) = \ln(1 - P_i). \tag{5.17}$$

The estimates of μ_{e} and σ_{e} can be estimated by the least-squares method since y and x are linearly related. The quantile function of the two-parameter exponential distribution is given by

$$Q(F_{\mathrm{e}}) = \mu_{\mathrm{e}} - \sigma_{\mathrm{e}}\ln(1 - F_i) \tag{5.18}$$

Return period

In order to develop an effective early warning drought monitoring strategy, we can estimate how often the extreme quantiles occur with a certain return level. The cumulative two-parameter exponential distribution is used to calculate return period as suggested by Berger et al. (1982). The return period is given by

$$R(x) = \frac{1}{(1 - f)(1 - F_{\mathrm{e}}(x))} \tag{5.19}$$

where $R(x)$ is the return period in years of the mean annual rainfall, x and f is the chosen specific percentile.

5.3 Empirical results

The three parent distributions were fitted to the data described in **Chapter 2**. The performance of the best fitting parent distribution was compared to the performance of the two-parameter exponential distribution derived from extreme value theory. Figure 5.2 shows the c.d.f. of the three theoretical parent distributions and mean annual rainfall for Zimbabwe.

From Figure 5.1, the c.d.f. of the three fitted theoretical distributions seems similar to the frequency distribution of the data. The distribution parameters i.e. scale and shape parameters can be estimated by the maximum likelihood procedure. Figure 5.2 to Figure 5.4 (a) (Top right panel) the empirical and theoretical densities, (b)

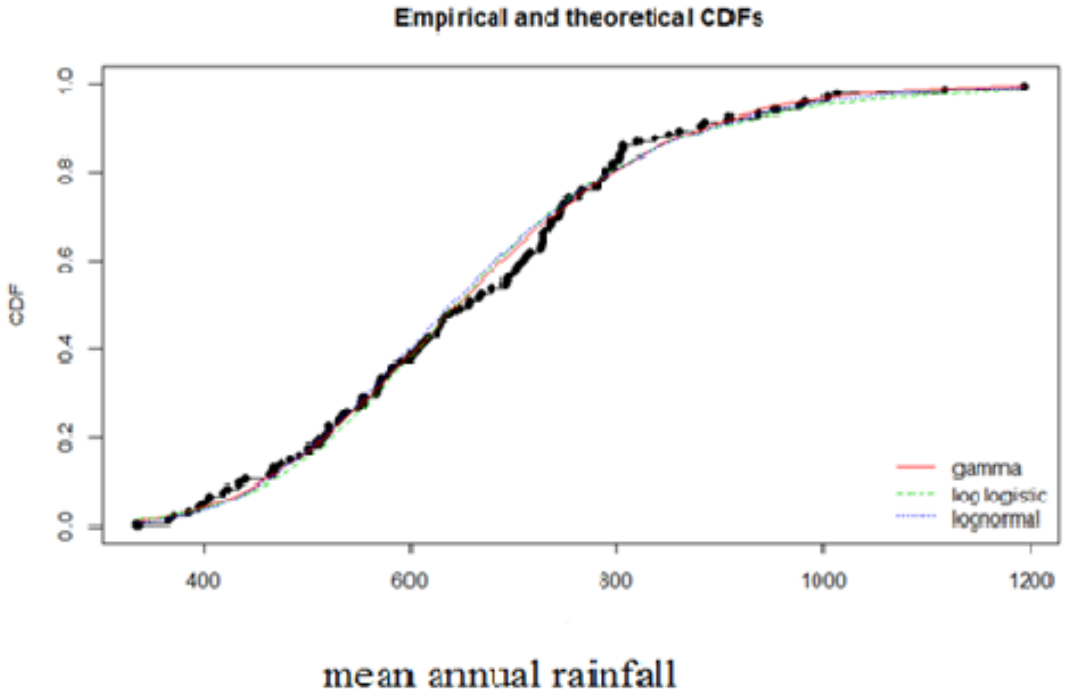

Figure 5.1: The c.d.f. of three theoretical parent distributions and mean annual rainfall for Zimbabwe

(top right panel) shows the QQ-plot, (c) (bottom left panel) shows the empirical and theoretical c.d.fs and (d) (bottom right panel) the empirical and theoretical c.d.f.s and (iv) the PP-plot of the fitted gamma, lognormal and log-logistic distributions respectively.

From Figure 5.2 (a) (Top right panel), it seems the empirical and theoretical densities are similar. The QQ-plot shown in Figure 5.2 (b) (top right panel) suggests that quantiles of the gamma distribution seems to match the quantile of the data (there is no serious divergence from the reference line), indicating that the gamma distribution may fit the data well.

From Figure 5.3 (a) (Top right panel), it seems the empirical and theoretical densities are similar. The QQ-plot shown in Figure 5.3(b) (top right panel) suggests that quan-

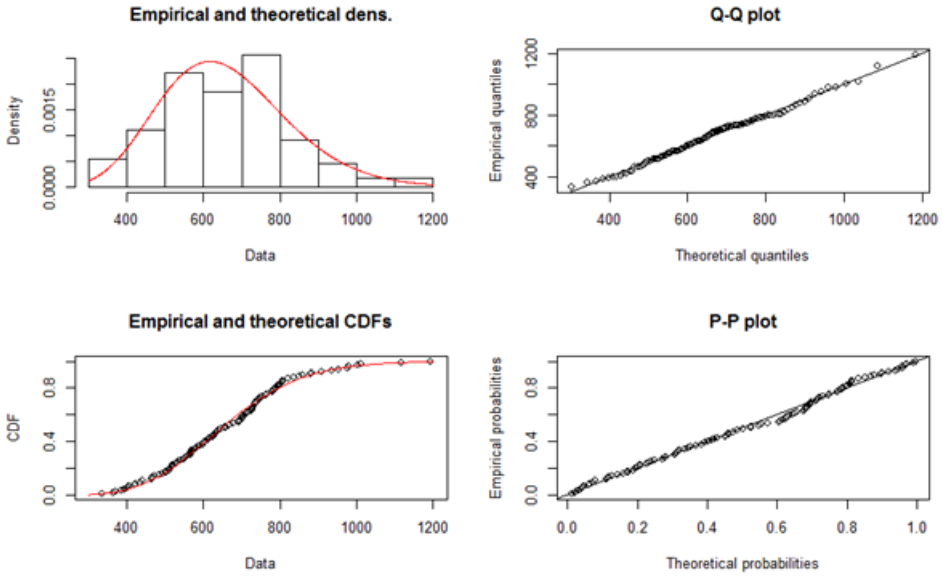

Figure 5.2: Diagnostic plots illustrating the fit of the mean annual rainfall data for Zimbabwe to the gamma distribution, (a) Empirical and gamma densities plot (top left panel), (b) QQ-plot (top right panel), (c) Empirical and gamma's c.d.f. plot (Bottom left panel) and (d) PP-plot (Bottom right panel)

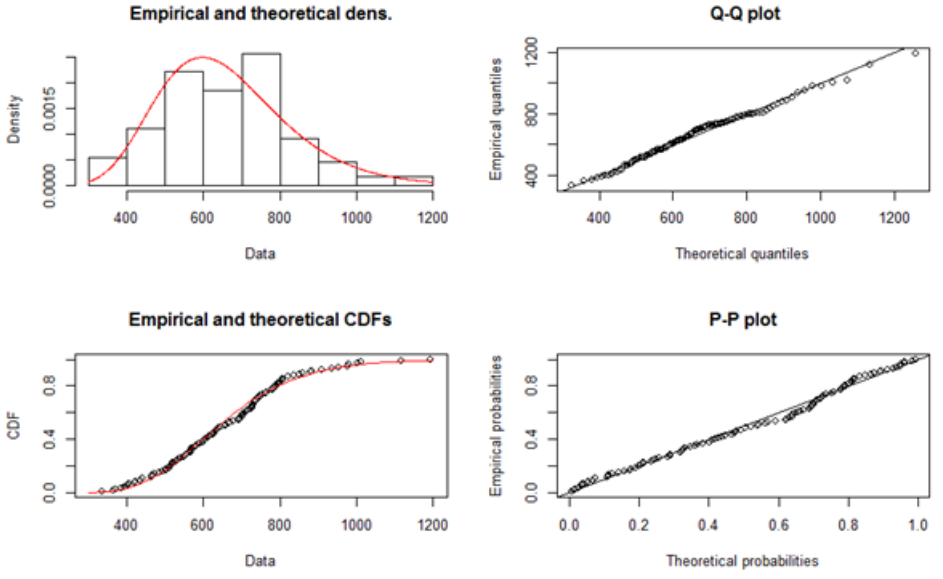

Figure 5.3: Diagnostic plots illustrating the fit of the mean annual rainfall data for Zimbabwe to the lognormal distribution, (a) Empirical and gamma densities plot (top left panel), (b) QQ-plot (top right panel), (c) Empirical and lognormal's c.d.f. plot (Bottom left panel) and (d) PP-plot (Bottom right panel)

tiles of the lognormal distribution seems to match the quantile of the data (there is no serious divergence from the reference), indicating that the lognormal distribution may fit well the data well.

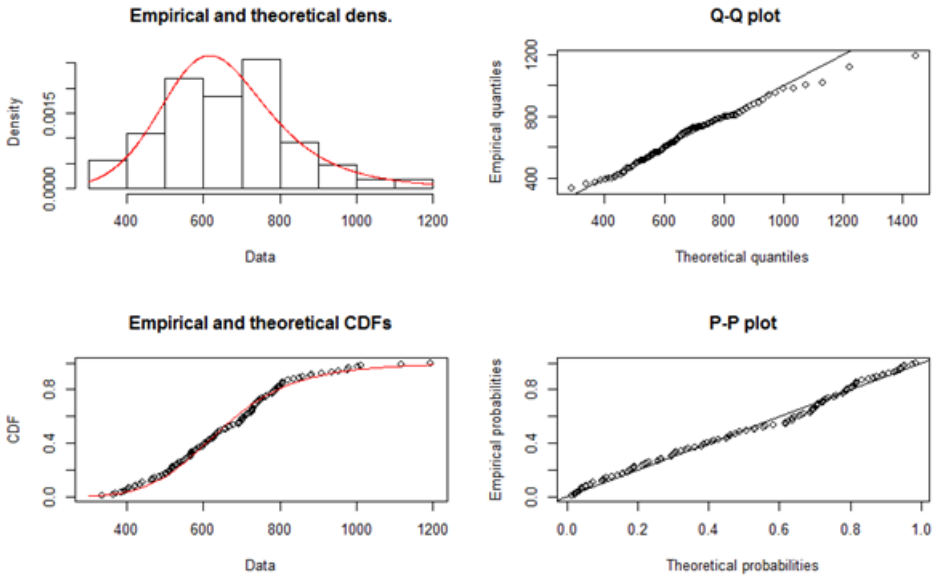

Figure 5.4: Diagnostic plots illustrating the fit of the mean annual rainfall data for Zimbabwe to the log-logistic distribution, (a) Empirical and gamma densities plot (top left panel), (b) QQ-plot (top right panel), (c) Empirical and log-logistic's c.d.f. plot (Bottom left panel) and (d) PP-plot (Bottom right panel)

From Figure 5.4 (a) (Top right panel), it seems the empirical and theoretical densities are similar. The QQ-plot shown in Figure 5.4(b) (top right panel) suggests that quantiles of the log-logistic distribution seems to match the quantile of the data (there is no serious divergence from the reference except at the upper tail of the distribution), indicating that the log-logistic distribution may fit the data well.

The maximum likelihood parameter estimates with their standard errors and p-values of the AD statistic for the fitted distributions are shown in Table 5.1.

Table 5.1: Fitted distributions, parameter estimates with standard errors in brackets and p-values of AD statistic

Distribution	$\hat{\sigma}$	$\hat{\xi}$	p-value for AD statistic
Gamma	0.0232	15.2801	0.2835
	(0.0030)	(1.9752)	
Lognormal	6.4591	0.2598	0.4637
	(0.0249)	(0.0176)	
Log-logistic	644.0337	6.6773	0.5125
	(16.1974)	(0.5288)	

The p-values of the AD statistics are all greater than 0.05 as reported in Table 5.1, confirming that the three theoretical distributions fits the data well. Overall, the log-logistic distribution gives the best fit by having the largest p-value for the AD statistic.

5.3.1 Selecting the best fitting parent distribution

Subsequent analysis involves selection of the best fitting distribution out of the three candidate distributions using the goodness-of-fit tests discussed in Chapter 2. Results of the goodness-of-fit tests are presented in Table 5.2.

Table 5.2: Outcomes of goodness-of-fit tests

Distribution	RRMSE	RMAE	PPCC
Gamma	**0.0218**	**0.0156**	**0.9970**
Lognormal	0.0260	0.0200	0.9950
Log-logistic	0.0318	0.0266	0.9879

The distribution that is best at least twice out of the three goodness-of-fit tests was selected as the best fitting distribution. Results indicated that the gamma distribution was the best fitting parent distribution since it had the least RRMSE and RMAE

and the highest PPCC values. We compared the relative performance of the fitted distributions at 25, 50 and 75 year periods. Table 5.3 shows the goodness-of-fit test results at different periods.

Table 5.3: Outcomes of the goodness-of-fit tests at different statistical periods

Period	Distribution								
	Gamma			Lognormal			Log-logistic		
	RRMSE	RMAE	PPCC	RRMSE	RMAE	PPCC	RRMSE	RMAE	PPCC
25	0.5821	0.5147	0.2359	0.5864	0.5135	0.2317	0.6141	0.5311	0.2260
50	0.3061	0.2718	0.9559	0.3089	0.2713	0.9389	0.3168	0.2739	0.9191
75	0.1774	0.1545	0.9657	0.1825	0.1537	0.9500	0.1998	0.1627	0.9257

From Table 5.3, the gamma distribution was found to be the best fitting period because it had the lowest RRMSE values and had the highest PPCC value which is closer to one. The results also showed that by increasing the period (hence data size) the performance of the fitted distribution improved. The PPCC value was greater than 0.9 for all the distributions when the period is 50 years or more. This suggests that analysing mean annual rainfall data using parent distributions, requires data of length at least 50 or longer.

Parent distributions are known to diverge in predicting high or low rainfall amounts. We fitted a two-parameter exponential distribution and compared their relative performance against the best fitting gamma distribution. To fit a two-parameter exponential distribution to extreme maxima rainfall, we selected the 90^{th} percentile which corresponds to 473 mm. Rainfall amounts above 473 mm were selected. We fitted the two-parameter exponential distributions to the selected data using the least-squares method since $y(i)$ and x are linearly related. Figure 5.5 shows the theoretical line of the variate, $y(i)$, and mean annual rainfall over the selected threshold of 473 mm.

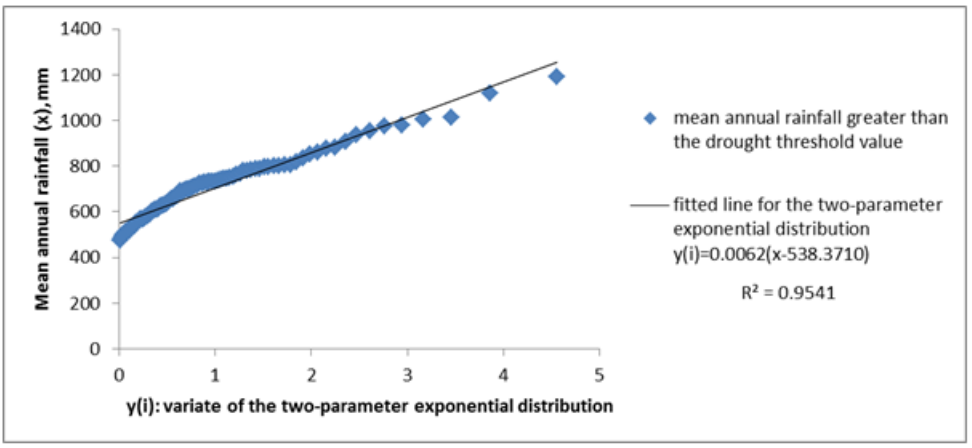

Figure 5.5: The fitted theoretical line of variate and mean annual rainfall above the selected threshold of 473 mm by the two-parameter exponential distribution

The regression model result showed that the coefficient of determination was greater than 0.95 which indicates that the two-parameter exponential distribution can fit maxima mean annual rainfall well. The F_{e} can be calculated from (5.15). The fitted two-parameter exponential distribution is $F_{\mathrm{e}}(x) = 1 - \exp[-0.0062(x - 538.3710)]$. The goodness-of-fit tests are used to compare the relative performance of the two-parameter distribution against the best fitting parent distribution. Table 5.4 shows the goodness-of-fit tests for the gamma and two-parameter exponential distributions.

Table 5.4: Outcomes of the goodness-of-fit tests for gamma and two-parameter exponential distributions

Distribution	RRMSE	RMAE	PPCC
Gamma	0.0323	0.0220	0.9887
Exponential	**0.0136**	**0.0012**	**0.9998**

From Table 5.4, the goodness-of-fit results shows that the two-parameter exponential distribution fits the data better at the tail than the best fitting parent distribution, the gamma distribution. We also compared the two distributions at different periods.

Table 5.5 shows the goodness-of-fit tests results at different periods.

Table 5.5: Outcomes of the goodness-of-fit tests for gamma and two-parameter exponential distributions at different periods.

Period	Distribution					
	Gamma			Exponential		
	RRMSE	RMAE	PPCC	RRMSE	RMAE	PPCC
25	0.5821	0.5147	0.2359	0.0230	0.0136	0.9892
50	0.3061	0.2718	0.9559	0.0254	0.0124	0.9887
75	0.1774	0.1545	0.9657	0.0262	0.0117	0.9900

From Table 5.5, the two-parameter exponential distribution was found to be the best performing distribution in fitting mean annual rainfall data series at all given periods. It can also be seen that the PPCC value for the two-parameter exponential distribution was closer to one at all periods. This indicates that the two-parameter exponential distribution fits the data well with even small data length. Thus, the distribution is a good candidate distribution for fitting and modelling extreme mean annual rainfall data regardless of the sample size of the data.

The F_{e} and return period of mean annual rainfall is calculated from (5.15) and (5.20). For example, $F_{\mathrm{e}}(473mm) = 1-\exp)[-0.0062(473-538.3701)] = -0.4997$. The chosen data above the drought threshold value of 473 mm corresponds to 90th percentile, i.e. $f = 0.90$. Then the return period of 473 mm is $R(x_c) = 1/[(1-0.9)(1+0.4997)] =$ 7 years, so a mean annual rainfall amount of 473 mm is expected to return in every 7 years' time. The maximum mean annual rainfall for Zimbabwe is 1192.6 mm recorded the in 1923/24 rainfall season. The mean return period associated with a return level estimate of 1193 mm is approximately 579 years. This suggests that an extreme flood of this magnitude is likely to return in every 579 years on average. This return period is very high, thus we explored analysis of extreme annual rainfall using the generalised extreme value and the generalised Pareto distributions in the

chapters that follow.

5.4 Concluding remarks

This chapter mainly investigated the relative performance of three commonly used probability distributions for mean annual rainfall, with the purpose of providing recommendation in the selection of suitable distribution for frequency analysis of mean annual rainfall. The results showed that the gamma distribution was the most suitable parent distribution. The worst performing was the log-logistic distribution. These results are consistent with the findings of Stagge et al. (2015), Cho et al. (2004) and Husak et al. (2007). This suggests that the statistical properties of daily rainfall may be similar to mean annual rainfall. Thus, in the absence of daily rainfall data set, the Department of Meteorological Services in Zimbabwe can use mean annual rainfall data to calculate drought indicators such as SPI. In addition, this study also established that when the length of the data is 50 or more, the gamma, lognormal and the logistic distributions fits well to the mean annual rainfall data. The performance of the best fitting parent distribution was compared to the performance of the two-parameter exponential distribution in fitting high rainfall extremes. This study established that the fitted two-parameter exponential agrees better with the actual data than the best fitting parent distribution at all different lengths of data outperforming the gamma distribution. This leads to the recommendation that the two-parameter exponential distribution is best suited to model mean annual rainfall. The return level estimates, which is the return level expected to be exceeded in a certain period of time T in years are calculated for Zimbabwe rainfall using the two-parameter exponential distributions. The highest mean annual rainfall amount recorded for the country was 1192.6 mm. A return level of 1193 mm is associated with a mean return period of 579 years, on average. This is too high and calls for further analysis of extreme rainfall for Zimbabwe using other extreme value distributions. Although, national data are analysed, the results of this study can be extended to station data in Zimbabwe. The findings of this study provided useful

information for early extreme rainfall monitoring management and provide a good alternative candidate for modelling mean annual rainfall extremes.

Chapter 6

Modelling of extreme maximum rainfall using generalised extreme value distribution for Zimbabwe

6.1 Introduction

Floods, although rare, present a problem in many parts of Southern Africa. The impact of such floods is felt on social and economic scales. The extent and severity of the damage resulting from such floods has been measured in many ways including loss of property, loss of lives, an agricultural crisis and high insurance claims (Gaioni et al., 2009). There are structural and non-structural flood mitigation measures in place in Zimbabwe. In Zimbabwe, dams were put in place to improve water security, they also serve as flood mitigation structures. Apart from the Kariba dam, Zimbabwe being in a semi-arid region, has 12 dams that play a part in flood mitigation. With regard to the non-structural flood mitigation measures, these include from early warning flood forecasting by the Department of Meteorological Services of Zimbabwe to rescue operations of the Civil Protection Organization of Zimbabwe. However, developing methods that can give an accurate prediction of meteorological events such as floods, is always a challenge to meteorologist and statisticians.

The use of standard statistical techniques which model average rainfall in forecasting and prediction of extreme events is less prudent due to gross under estimation (Schmidli et al., 2007). At the most extreme levels, data is scarce, whereas for structural design purposes, it is the behavior at these levels which is of greatest interest. The use of Extreme Value Theory (EVT) which deals with the stochastic behavior of extreme values in a process, has a long history (Nadarajah and Choi, 2007).

Recent research provides evidence of the importance of classical methods of modelling extreme rainfall amounts from different regions of the world; see Nadarajah and Choi (2007) for an application to a rainfall data set from South Korea. Koutsoyiannis (2004) applied extreme value theory to a rainfall data set from Europe and the USA. Koutsoyiannis and Baloutsous (2000) applied the theory to Greece's rainfall data. Coles and Tawn (2005) argues that classical methods are not efficient in terms of data usage since typically only the most extreme observation per year is modelled. Research has also focused on the development of inferential procedures that maximise the use of available data. To improve efficiency in the modelling of extreme values, it is important to include other sources of knowledge in the analysis. This may be information from known physical constraints on the data or derived from an understanding of related processes at different locations. This leads to the Bayesian inferential framework as a basis for undertaking extreme value analysis. Research on modelling extreme rainfall amounts incorporating Bayesian methodology is limited. Notable ones are Gaioni et al. (2009) who modelled flash floods with prior elicitation using one historic US river data while Coles and Tawn (2005) analysed extreme rainfall in South-West England using expert prior information. Smith and Naylor (1987) examined the effect of different prior assumptions on the distributions of the parameters of a Weibull distribution. There is no work known to us on extremes of Zimbabwean rainfall using the Bayesian framework. The aim of this study is to propose an early warning system to predict the frequency of floods in Zimbabwe. In this study, we provide an application of extreme value distribution to model rainfall

data from Zimbabwe using the Bayesian statistics approach. In March 1996, February 2000 and March 2003, cyclones hit Zimbabwe bringing with it intense storms which caused flooding in many parts of the country (Gwimbi, 2009; Madamombe, 2004). In February 2005 and December 2007, the Department of Meteorological Services in Zimbabwe warned of flooding in the Muzarabani and Chadereka districts (Chingombe et al., 2015). The mean annual rainfall for the years 1996, 2000, 2003, 2005 and 2007 for which recent flooding in many parts of the country was reported was 801.6 mm, 728.6 mm, 712.3 mm, 835.7 mm and 796.2 mm respectively. It is our assumption that a higher mean annual rainfall (above 775 mm) may lead to flooding. In this chapter, we model maxima annual rainfall assuming that it may lead to floods. Knowledge of floods, especially their return periods, will assist in developing flooding preparedness and coping strategies for the people of Zimbabwe. There is no known work to us on rainfall extremes in Zimbabwe. In this chapter, we provide the first application of extreme value distributions to model maxima annual rainfall in Zimbabwe.

In Chapter 4 we discussed the modelling of mean annual rainfall using parent distributions. Parent distributions covers the main body of the data. At times there is need to depart from focusing on the main body and exploit information provided by the extremes (tails of distribution). Parent distributions are known to diverge in predicting extreme rainfall amounts. In order to fully capture the statistical properties of mean annual rainfall, it is important to model the tail distribution of the rainfall data. The objective of this Chapter is to quantify and describe the behaviour of extreme high rainfall in Zimbabwe. In particular, the aim is to model (i) extreme high rainfall using GEVD by using the maximum likelihood estimation method and the Bayesian statistics approach. (ii) calculate the mean return period of high rainfall i.e. the number of years on average before another flood of equal or greater intensity. (iii) investigate the influence of time and weather/climate change drivers, namely; SOI and SDSLP anomalies on extreme maxima rainfall.

The rest of the chapter is organized as follows: Section 6.2 presents the research methodology. Section 6.3 presents the empirical results and discussion of the findings. Section 6.4 gives the general remarks of the results. Section 6.5 provides the concluding remarks. Finally, derivations of important results and diagnostic plots for the fitted non-stationary models are shown in the Section 6.6.

6.2 Research methodology

In this section we discuss the background theory on stationary and non-stationary GEVD models. The data used in this chapter was described in Chapter 3. The parameters of the fitted models are estimated by the method of maximum likelihood and method of Bayesian statistics.

6.2.1 Extreme value theory for block maxima

Suppose $X_1, ..., X_n$ is a sequence of independent and identically distributed (i.i.d.) mean annual rainfall amounts with distribution function G. Then, the behaviour of

$$M_n = \max\{X_1, ..., X_n\} \tag{6.1}$$

where M_n is the maximum mean annual rainfall amounts over an n-observation period. If the exact statistical behaviour of the X_i were known, the corresponding behaviour of M_n could also be calculated exactly. But, in practice, the behaviour of X_i is unknown, making it difficulty to determine the behaviour of M_n. However, under suitable assumptions, for large n the approximate behaviour of M_n can be determined. Thus, properties of M_n as $n \to \infty$ are of particular importance. The behaviour of selected M_n values in every consecutive period is of great interest. The selected observations are named block maxima. Figure 6.1 illustrates how 4 maxima were selected over a period covering 4 blocks. Each block has 5 observations.

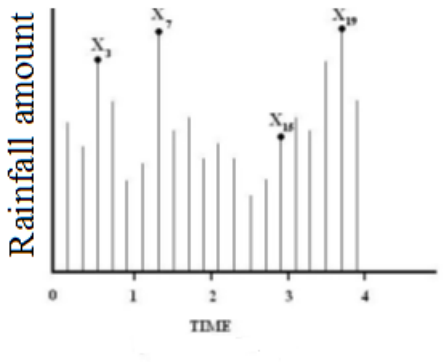

Figure 6.1: Illustration of selecting variables for block maxima approach

The block maxima approach in EVT, consists of dividing the observation period into non-overlapping periods of equal size (blocks) and restricts attention to the maximum observation in each period. The new observations (maxima from each block) under extreme value conditions, follows approximately an extreme value distribution (Ferreira and de Haan, 2015). Parametric statistical methods for the extreme value distributions (EVD) are then applied to these new observations. In theory the distribution of M_n in the case of a maxima can be derived exactly for all values of n as:

$$P_r\{M_n \leq x\} = P_r\{X_1 \leq x, ..., X_n \leq x\} \tag{6.2}$$

$$= P_r\{X_1 \leq x\} \times ... \times P_r\{X_n \leq x\}$$

$$= \prod_{i=1}^{n} P_r\{X \leq x\}$$

$$= [P_r(X \leq x)]^n$$

$$= G^n(x)$$

M_n converges almost surely to $\omega(G)$ whether it is finite or infinite. To estimate the value of G^n from observed data, one possibility is to use standard statistical techniques. However, For any $x < x_+$ where x_+ is the upper point of G, two problems can be encountered from using standard statistical techniques. The first problem is that, G is not the exact distribution for these new observations, a bias may appear. The second one, $FG^n \to 0$ as $n \to \infty$ since $G(x) < 1$ (Pocernich, 2002). The first problem is avoided by accepting that F is unknown and to look for appropriate families of models for G^n which can be estimated on the basis of extreme data only. The second problem is avoided by allowing a linear renormalization of the variable M_n. The limit theory in univariate extremes seeks normalising constants $a_n > 0$ and b_n and a nondegenerate function F such that the c.d.f. of a normalized version of M_n converges to G, i.e.

$$P_r\left(\frac{M_n - b_n}{a_n} \leq x\right) = G^n(a_n x + b_n) \to F(x) \tag{6.3}$$

as $n \to \infty$. If this holds for suitable choices of $\{a_n\}$ and $\{b_n\}$, then we say that F is an extreme value c.d.f. and G is in the domain attraction of F, written as $G \in D(F)$. We say further that two extreme value cdfs F and F^* are of the same type if $F^*(x) = F(ax + b)$ for some $a > 0$, b and all x (Fisher and Tippet, 1928 and Gnedenko, 1943).

Theorem 6.1 (Extremal types theorem)

If there exist sequences of constants $\{a_n > 0\}$ and $\{b_n\}$ such that

$$P_r\{(M_n - b_n)/a_n \leq x\} \to F(x), \quad \text{as } n \to \infty \tag{6.4}$$

where F is a non-degenerate distribution function, then F belongs to one of the following families:

$$I : F(x) = \exp\left\{-\exp\left[-\left(\frac{x-b}{a}\right)\right]\right\}, \text{if } -\infty < x < \infty; \tag{6.5}$$

$$II : F(x) = \begin{cases} 0, & \text{if } x \leq b, \\ \exp\left\{-\left(\frac{x-b}{a}\right)^{-\alpha}\right\}, & \text{if } x > b; \end{cases} \tag{6.6}$$

$$III : F(x) = \begin{cases} \exp\left\{-\left[-\left(\frac{x-b}{a}\right)^{\alpha}\right]\right\}, & \text{if } x < b; \\ 1, & \text{if } x \geq b, \end{cases} \tag{6.7}$$

For parameters $a > 0$, b, and $\alpha > 0$.

The theorem says if there are suitable sequences $\{a_n\}$ and $\{b_n\}$ to stabilize M_n, then F has a limiting distribution which is one of the types I, II and III. These three families of distributions are known as the Gumbel, Fréchet and Weibull distributions respectively and are collectively known as the extreme value distributions.

The family of EVDs can be equivalently characterised via the class of max-stable distributions. Two distributions F^* and F are equivalent if $F^*(ax+b) = F(x)$. While a distribution is max-stable if for $n = 2, 3, ...$ there are constants σ_n and ξ_n such that $F^n(\sigma_n x + \xi_n) = F(x)$ (de Haan and Lin, 2001; Leadbetter et al., 2012). Leadbetter et al. (2012) gave a comprehensive account of necessary and sufficient conditions for $G \in D(F)$ and characterizations of a_n and b_n, when F is one of the three extreme value cdfs. The necessary and sufficient conditions for the three cdfs in (6.5), (6.6) and (6.7) are:

$$I : \exists \gamma(t) > 0 \text{ s.t. } \lim_{t \uparrow \omega(G)} \frac{1 - G(t + x\gamma(t)}{1 - G(t)} = \exp(-x), \qquad x \in \Re, \tag{6.8}$$

$$II : \omega(G) = \infty \text{ and } \lim_{t \uparrow \infty} \frac{1 - G(tx)}{1 - G(t)} = x^{-\xi}, \qquad x > 0, \tag{6.9}$$

$$III : \omega(G) < \infty \text{ and } \lim_{t \downarrow 0} \frac{1 - G(\omega(G) - tx)}{1 - G(\omega(G) - t)} = x^{\xi}, \qquad x > 0. \tag{6.10}$$

The corresponding characterizations of a_n and b_n are:

$$
\begin{aligned}
I:\ a_n &= \gamma(G^{\leftarrow}(1-n^{-1}))\ \text{ and }\ b_n = G^{\leftarrow}(1-n^{-1}),\\
II:\ a_n &= G^{\leftarrow}(1-n^{-1})\ \text{ and }\ b_n = 0,\\
III:\ a_n &= \omega(G) - G^{\leftarrow}(1-n^{-1})\ \text{ and }\ b_n = \omega(G),
\end{aligned}
$$

where $G^{\leftarrow}$ denotes the inverse function of G.

An equivalent characterization of F is by means of the definition of max-stability:

Theorem 6.2 (Max-stable distributions) *A distribution F is max-stable if and only if, it is of the same type as an extreme value distribution.*
From Theorem 6.2, it implies that the distribution of maxima of independent samples is of the same type as that of the underlying population if and only if, the underlying population itself has a distribution of extreme value type.

Leadbetter et al. (2012) showed that the class of extreme value distributions are precisely the class of nondegenerate max-stable distributions. Resnick (2013) showed that the α_n and b_n for the distributions in (6.5), (6.6) and (6.7) are:

$$
\begin{aligned}
I:&\quad \sigma_n = 1\ \text{ and }\ \xi_n = -\log n,\\
II:&\quad \sigma_n = n^{-1/\sigma}\ \text{ and }\ \sigma_n = 0,\\
III:&\quad \sigma_n = n^{1/\sigma}\ \text{ and }\ \xi_n = 0.
\end{aligned}
$$

6.2.2 Generalised extreme value distribution (GEVD) for block maxima and minima

An approach to unify the three families of extreme value distributions was proposed by von Mises (1954) and Jenkinson (1955), into a single family to allow a continu-

ous range of possible shapes known as the "generalised extreme value distribution" (GEVD)(Kots and Nadarajah, 2000; Gill and Kllezi, 2006). Based on Theorem 6.1, the GEVD is the limiting distribution of properly normalized maxima of a sequence of independent and identically distributed random variables (Beirlant et al., 2004). Thus, the GEVD is used to model the maxima of a long (finite) sequence of random variables. The unified GEVD for modelling maxima is given by:

$$F_{\sigma,\mu,\xi}(x) = \exp\left\{-\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\} \tag{6.11}$$

defined on $\{x : 1+\xi(x-\mu)/\sigma > 0\}$ with $-\infty < \mu < \infty$, $\sigma > 0$ and $-\infty < \xi < \infty$. where μ, σ and ξ are location, scale and shape parameters, respectively. The value of the shape parameter ξ distinguishes the type of extreme value distribution (Coles et al., 2001).

In general form, the Gumbel distribution corresponds to the case of $\xi = 0$. The case of $\xi = 0$ is interpreted as the limit as $\xi \to 0$ of (6.11). The cumulative distribution function (c.d.f.) of the Gumbel distribution function is given by:

$$F(x) = \exp\left\{-\exp\left[-\left(\frac{x-\mu}{\sigma}\right)\right]\right\}, \quad -\infty < x < \infty. \tag{6.12}$$

for $-\infty < x < \infty$, $-\infty < \mu < \infty$ and $\sigma > 0$. The Gumbel distribution has two main parameters μ and σ as $\xi \to 0$. The probability density function is obtained as the derivative of (6.12). The probability density function of the Gumbel distribution is given by:

$$f(x) = \frac{1}{\sigma}\exp\left[-\exp\left[-\left(\frac{x-\mu}{\sigma}\right)\right]\right]\exp\left[-\left(\frac{x-\mu}{\sigma}\right)\right], \tag{6.13}$$

and the quantile function is given by:

$$Q(p) = \mu - \sigma\ln(-\ln(p)) \tag{6.14}$$

The Gumbel distribution comprises distributions such as:

(i) Exponential distribution: $F(x) = \left[1 - \frac{e^{-x}}{n}\right]^n \rightarrow \exp(-e^{-x})$, which is the Gumbel distribution with $\mu = 0$ and $\sigma = 1$.

(ii) Normal, logistic and gamma distributions.

These distributions are identified as light-tailed distributions, thus $\xi \rightarrow 0$ is related to the light-tailed Gumbel order of distribution. The Gumbel distribution is also called the double exponential distribution and is bounded, which means every moment exists.

The characteristic function of the Gumbel distribution is obtained by first transforming the Gumbel random variable X into a standard random variable using $Z = \frac{X-\mu}{\alpha}$. The density of Z is then given by:

$$f(z) = e^{-e^{-z}}e^{-z}, \quad \text{for} - \infty < z < \infty. \tag{6.15}$$

The characteristic function of Z is determined to be:

$$E(e^{itZ}) = \int_{-\infty}^{\infty} e^{itz}e^{-e^{-z}}e^{-z}dz$$

This solves to give:

$$E(e^{itZ}) = \Gamma(1 - it)$$

where $i = \sqrt{-1}$. Then the characteristic function of Gumbel distribution is therefore given by:

$$E(e^{itX}) = e^{it\mu}\Gamma(1 - it\sigma). \tag{6.16}$$

The first and second derivative of (6.16) at the interval $t = 0$, gives the first and

second moments of X. These moments are:

$$E(X) = \mu - \sigma\Gamma^{(1)} = \mu + \sigma\gamma \tag{6.17}$$

$$E(X^2) = \mu^2 - 2\sigma\mu\Gamma^{(1)} + \sigma^2\Gamma^{(2)}(1) = (\mu + \sigma\gamma)^2 + \sigma^2\frac{\pi^2}{6} \tag{6.18}$$

and the constant γ which is referred as the Euler-Mascheroni's constant and $\gamma \approx 0.577215664901532$. Thus the mean and variance of the Gumbel distribution are $\mu + \sigma\gamma$ and $\sigma^2\frac{\pi^2}{6}$ (Feng et al., 2007).

For the case $\xi \neq 0$ in (6.11), we have the type II when $\xi > 0$ and type III when $\xi < 0$. The type II EVD also known as the Fréchet distribution. The Fréchet distribution has heavy upper tails for all $\xi > 0$. Its upper tails become heavier with smaller values of ξ. The heavy tail distribution of the Fréchet family of distributions includes the Pareto, log-gamma, Student t and the Cauchy distributions. The Fréchet type of distribution is described by a polynomial tail decay, having moments only up to γ. The Fréchet distribution has a c.d.f. given by:

$$F(x) = \exp\left\{-\left(\frac{x-\mu}{\sigma}\right)^{-\alpha}\right\} \tag{6.19}$$

for $x > \mu$, $\alpha > 0$, $-\infty < \mu < \infty$ and $\sigma > 0$. The type III EVD is equivalent to the Weibull distribution. The Weibull distribution has heavy tails for $0 < \alpha < 1$. Its upper tails become heavier with smaller values of $\alpha \in (0, 1)$. The Weibull distribution has exponentially decaying light upper tails for all $\alpha \geq 1$. Its upper tails become lighter with larger values of $\alpha \in [1, \infty)$. The Weibull family of distributions includes the Uniform and Beta distributions. The Weibull type of distribution is characterized by a bounded tail with a finite endpoint, although not all of its moments are finite. The Weibull distribution has the cdf:

$$F(x) = 1 - \exp\left\{-\left(\frac{x-\mu}{\sigma}\right)^{\alpha}\right\} \tag{6.20}$$

for $x > \mu$, $\alpha > 0$, $-\infty < \mu < \infty$ and $\sigma > 0$.

The parameter, α in (6.19) and (6.20) controls the shape of the EVDs. Both the Fréchet and the Weibull distributions have three parameters. Since the EVDs of type I and type II accommodate heavy upper tails, they have received applications in numerous areas. These include: value at risk analysis (Chinhamu et al., 2015; Bali, 2007; Embrechts et al., 1999) and estimation of extreme rainfall return levels (Chikobvu and Chifurira, 2015; Mayooran and Laheetharan, 2014; Feng et al., 2007; Coles et al., 2001; Martins and Stedinger, 2000). For more examples: see Embrechts et al. (1999).

If we differentiate the cdf for the GEVD in (6.11), we get the probability density function (sometimes called the Fisher-Tippett distribution):

$$f(x) = \frac{1}{\sigma}\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-1-\frac{1}{\xi}} \exp\left\{-\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\}, \quad \xi \neq 0, \tag{6.21}$$

with $\mu \in \Re$,$\sigma > 0$ and $1+\xi\left(\frac{x-\mu}{\sigma}\right) > 0$. The shape parameter ξ is also known as the extreme value index and the parameter ξ^{-1} is the rate of decay of the GEVD. The quantile function of the GEVD is given by:

$$Q(p) = \mu + \frac{\sigma}{\xi}\left[(-\ln(1-p))^{-\xi} - 1\right], \xi \neq 0. \tag{6.22}$$

The closed form expressions for the characteristic function EVDs type II and III are unknown in the literature. Thus one way of calculating the moments for the GEVD (mainly the Fréchet and the Weibull distributions) is to first transform the generalised extreme value random variable X into a standard Gumbel random variable. If we let:

$$\left[1+\xi\left(\frac{X-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}} = e^{Z} \tag{6.23}$$

then

$$Z = \frac{1}{\xi}\ln\left[1+\xi\left(\frac{X-\mu}{\sigma}\right)\right] \tag{6.24}$$

and

$$X = \mu - \frac{\sigma}{\xi} + \frac{\sigma}{\xi}e^{Z\xi} \tag{6.25}$$

Z is a standard Gumbel random variable, thus the moment generating function of Z is of the form:

$$M_Z(t) = \Gamma(1-t). \tag{6.26}$$

We can easily calculate the first two moments of the generalised extreme value random variable X in (6.25) as:

$$\begin{aligned} E(X) =& E\left[\mu - \frac{\sigma}{\xi} + \frac{\sigma}{\xi}e^{Z\xi}\right] \\ =& \mu - \frac{\sigma}{\xi} + \frac{\sigma}{\xi}M_Z(\xi) \\ =& \mu - \frac{\sigma}{\xi} + \frac{\sigma}{\xi}\Gamma(1-\xi) \end{aligned} \tag{6.27}$$

and also

$$\begin{aligned} E(X^2) =& E\left[\left(\mu - \frac{\sigma}{\xi}\right)^2 + 2\left(\mu - \frac{\sigma}{\xi}\right)\frac{\sigma}{\xi}e^{Z\xi} + \frac{\sigma^2}{\xi^2}e^{2Z\xi}\right] \\ =& \left(\mu - \frac{\sigma}{\xi}\right)^2 + 2\left(\mu - \frac{\sigma}{\beta}\right)\frac{\sigma}{\xi}M_Z(\xi) + \frac{\sigma^2}{\xi^2}M_Z(2\xi) \\ & \left(\mu - \frac{\sigma}{\xi}\right)^2 + 2\left(\mu - \frac{\sigma}{\xi}\right)\frac{\sigma}{\xi}\Gamma(1-\xi) + \frac{\sigma^2}{\xi^2}\Gamma(1-2\xi) \end{aligned} \tag{6.28}$$

(6.27) and (6.28) only exists when $\xi < 1$ and $\xi < \frac{1}{2}$. Thus, the n^{th} moment $E(X^n)$ will only exist if $\xi < \frac{1}{n}$. Using the first and the second moments, the mean and the variance of the generalised extreme value random variable X is given by $\mu - \frac{\sigma}{\xi} + \frac{\sigma}{\xi}g_1$ and $\frac{\sigma^2}{\xi^2}(g_2 - g_1)$ respectively, where $g_k = \Gamma(1-k\xi)$.

6.2.3 Estimation procedure of parameters for the GEVD

There is a wide variety of methods used to estimate the GEVD parameters in the independent and identically distributed settings (Diebolt et al., 2007). The three parameters are estimated by method of moments, maximum likelihood and probability weighted moments. In this thesis we used the method of maximum likelihood, this procedure gives consistent estimates.

Method of maximum likelihood

Under the assumption that $x_i, x_2, ..., x_m$ are observed block maxima which are independent realizations from a generalised extreme value random variable, the log-likelihood function for the GEVD parameters when $\xi \neq 0$ is:

$$l(\mu, \sigma, \xi) = -m\ln(\sigma) - \left(1 + \frac{1}{\xi}\right) \sum_{i=1}^{m} \ln\left[1 + \xi\left(\frac{x_i - \mu}{\sigma}\right)\right] - \sum_{i=1}^{m} \left[1 + \xi\left(\frac{x_i - \mu}{\sigma}\right)\right]^{-\frac{1}{\xi}} \tag{6.29}$$

provided that $1+\xi\left(\frac{x_i-\mu}{\sigma}\right) > 0$ for $i = 1, ..., m$ (Coles et al., 2001). We differentiate the log-likelihood of GEVD with respect to each parameter to obtain a set of equations. By differentiating (6.29) with respect to ξ we obtain:

$$\begin{aligned}\frac{dl}{d\xi} =& \frac{1}{\xi^2} \sum_{i=1}^{m} \ln\left[1 + \xi\left(\frac{x_i - \mu}{\sigma}\right)\right] - \left(1 + \frac{1}{\xi}\right) \sum_{i=1}^{m} \frac{x_i - \mu}{\sigma + \xi(x_i - \mu)} \\ &- \frac{1}{\xi^2} \sum_{i=1}^{m} \sum_{i=1}^{m} \left[1 + \xi\left(\frac{x_i - \mu}{\sigma}\right)\right]^{-\frac{1}{\xi}} \ln\left[1 + \xi\left(\frac{x_i - \mu}{\sigma}\right)\right] \\ &+ \frac{1}{\sigma} \sum_{i=1}^{m} \sum_{i=1}^{m} \left[1 + \xi\left(\frac{x_i - \mu}{\sigma}\right)\right]^{-\frac{1}{\xi}} \left[\frac{x_i - \mu}{\sigma + \xi(x_i - \mu)}\right]\end{aligned} \tag{6.30}$$

By differentiating (6.29) with respect to σ we obtain:

$$\begin{aligned}\frac{dl}{d\sigma} =& -\frac{m}{\sigma} - \left(1 + \frac{1}{\xi}\right) \sum_{i=1}^{m} \frac{\xi(x_i - \mu)}{\sigma[\sigma + \xi(x_i - \mu)]} \\ &- \sum_{i=1}^{m} \sum_{i=1}^{m} \left[1 + \xi\left(\frac{x_i - \mu}{\sigma}\right)\right]^{-\frac{1}{\xi}} \frac{\xi(x_i - \mu)}{\sigma[\sigma + \xi(x_i - \mu)]}\end{aligned} \tag{6.31}$$

By differentiating 6.29 with respect to μ we obtain:

$$\frac{dl}{d\mu} = \left(1 + \frac{1}{\xi}\right) \sum_{i=1}^{m} \frac{\xi}{\sigma + \xi(x_i - \mu)} - \left[1 + \xi\left(\frac{x_i - \mu}{\sigma}\right)\right]^{-\frac{1}{\xi}} \left[\frac{1}{\sigma + \xi(x_i - \mu)}\right] \quad (6.32)$$

Note: see Section 6.6 for the derivation of the (6.30) to (6.32).

(6.30), (6.31) and (6.32) have no analytical solutions. To obtain the maximum likelihood estimates of ξ, σ and μ (6.30), (6.31) and (6.32) are set to zero and standard numerical optimization algorithms such as Quasi-Newton numerical maximization are applied.

6.2.4 Properties of the GEVD log-likelihood

The following proposition gives us some basic properties of the GEVD log-likelihood. We note x^- and x^+ the left and right end point of the domain l, i.e.

$$(x^-, x^+) = \{x \in \Re; 1 + \xi x > 0\}.$$

Clearly, it is equal to $(-\infty, -\frac{1}{\xi})$, $\Re$ and $(-\frac{1}{\xi}, +\infty)$ when $\xi < 0$, $\xi = 0$ and $\xi > 0$ respectively.

Proposition 6.1 *The function l is smooth on its domain.*

1. *If $\xi \leq -1$, l is strictly increasing on its domain and*

$$\lim_{x \to x^-} l(x) = -\infty \qquad \lim_{x \to x^+} l(x) = \begin{cases} +\infty & \text{if } \xi < -1 \\ \\ 0 & \text{if } \xi = -1 \end{cases} \quad (6.33)$$

2. *If $\xi > -1$, l is increasing on $(x^-, x^+]$ and decreasing on $[x^-, x^+)$, where*

$$x^* = \frac{(1 + \xi)^{-\xi} - 1}{\xi} \quad (6.34)$$

Furthermore

$$\lim_{x \to x^-} l(x) = \lim_{x \to x^+} l(x) = -\infty$$

and l reaches its maximum $l(x^*) = (1+\xi)(\ln(1+\xi) - 1)$ uniquely.

This entails that the log-likelihood (6.29) has no local maxima in $(-\infty, -1] \times \Re \times (0, +\infty)$ and that any maximum likelihood estimate $(\hat{\mu}, \hat{\sigma}, \hat{\xi})$ satisfies $\hat{\xi} > -1$. Hence, no consistent maximum likelihood estimate does exist if $\xi < -1$ (Smith, 1985b).

Estimating the parameters of GEVD using the Bayesian approach

Estimating parameters of extreme value distributions using the method of maximum likelihood may produce some irregularities as outlined by Smith (1985). According to Smith and Naylor (1987), the problem of regularity assumptions in the limiting behaviour of the maximum likelihood estimation method is encountered if the estimated shape parameter, $\hat{\xi} < -\frac{1}{2}$. The Bayesian estimation method is independent of any regularity conditions (Beirlant et al., 2004). Bayesian inference allows additional information about the processes to be incorporated as prior information. This additional information is modelled through the introduction of the prior distribution $\pi(\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a vector of parameters with no reference to the data, i.e. $\boldsymbol{\theta} = \{\sigma, \mu, \xi\}$ in this thesis. For specification of the prior, the parameterisation $\phi = \log(\sigma)$ is easier to work with since σ is constrained to be positive. The parameters ϕ, μ and ξ are treated as random variables for which we specify prior distributions. Therefore, prior density is

$$\pi(\phi, \mu, \xi) = \pi_\phi(\phi)\pi_\mu(\mu)\pi_\xi(\xi), \tag{6.35}$$

where each marginal prior is normally distributed with large variances. The variances are chosen to be large enough to make the distributions almost flat, corresponding to prior ignorance. The benefits of using any information available are likely to be great, however not all available information can be formulated into prior

information. Coles and Powell (1996) acknowledges that scarce extreme data may pose a challenge to experts to formulate prior beliefs about the process. There are two types of priors that are commonly used namely; informative and non-informative priors. Priors based on expert information are called informative priors, while the priors formulated without the expert information are known as non-informative priors.

After formulating the prior distribution, the second step involves collecting data to form the likelihood function denoted by $L(\boldsymbol{\theta} \mid \mathbf{x})$, where $\mathbf{x} = \{x_i, ..., x_m\}$ is the vector of realizations of a random variable with density from the parametric family $F = \{f(x;\theta) : \theta \in \Theta\}$. The likelihood for $\boldsymbol{\theta}$ is

$$\begin{aligned} L(\boldsymbol{\theta} \mid \mathbf{x}) &= f(\mathbf{x} \mid \boldsymbol{\theta}) \\ &= \prod_{i=1}^{m} \frac{1}{\sigma}\left[1+\xi\left(\frac{x_i-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}-1} \exp\left\{-\left[1+\xi\left(\frac{x_i-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\} \\ &= \frac{1}{\sigma^m}\exp\left\{-\sum_{i=1}^{m}\left[1+\xi\left(\frac{x_i-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\}\prod_{i=1}^{m}\left[1+\xi\left(\frac{x_i-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}-1} \end{aligned} \tag{6.36}$$

The likelihood describes how vector $\mathbf{x}$ depends on $\boldsymbol{\theta}$ (Renard et al., 2006). The last step is combining the prior distribution with the likelihood to form the posterior distribution. The posterior distribution is computed using Bayes' Theorem

$$\pi(\boldsymbol{\theta} \mid \mathbf{x}) = \frac{\pi(\mathbf{x} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int_{\Theta} \pi(\mathbf{x} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta}) d\boldsymbol{\theta}} \tag{6.37}$$

where Θ is the space parameter and the posterior distribution is usually written as:

$$\pi(\boldsymbol{\theta} \mid \mathbf{x}) \propto \pi(\boldsymbol{\theta}) \times L(\boldsymbol{\theta} \mid \mathbf{x}) \tag{6.38}$$

The Markov chain Monte Carlo (MCMC) technique is used to simulate realisations of the posterior distribution. Estimates of the posterior distribution could then be obtained from the simulated sample.

MCMC techniques

The MCMC technique provides a way of simulating from complex distributions by simulating Markov chains which have the target distributions as their stationary distributions. There are many MCMC techniques available (see Gilks et al., 1996; Gamerman, 1997; Besag, 2001), however, in this thesis we discuss the Gibbs sampler and Metropolis-Hastings sampling.

The Gibbs sampler

The Gibbs sampler was used by Geman and Geman (1984) for models with Gibbs distribution and was extended to the general form by Gelfand and Smith (1990). Suppose the density of interest is $\pi(\boldsymbol{\theta})$, where $\boldsymbol{\theta} = (\theta_1, ..., \theta_d)'$ and the full conditionals are given by

$$\pi(\theta_i \mid \theta_1, ..., \theta_{i-1}, \theta_{i+1}, ..., \theta_d) = \pi(\theta_i \mid \theta_{-i}), \quad i = 1, ..., d. \tag{6.39}$$

If it is possible to simulate from the full conditionals then the Gibbs sampler can be obtained using the following algorithm:

(i) Initialize the counter to $k = 1$ and the state of the chain to $\theta^{(0)} = (\theta_1^{(0)}, ..., \theta_d^{(0)})'$.

(ii) Obtain a new value $\theta^{(k)}$ from $\theta^{(k-1)}$ by successive simulation from the full conditionals

$$\theta_1^{(k)} \sim \pi\left(\theta_1 \mid \theta_2^{(k-1)}, ...\theta_d^{(k-1)}\right)$$

$$\theta_2^{(k)} \sim \pi\left(\theta_2 \mid \theta_1^{(k)}\theta_3^{(k-1)}, ...\theta_d^{(k-1)}\right)$$

.

.

.

$$\theta_d^{(k)} \sim \pi\left(\theta_d \mid \theta_1^{(k)}\theta_3^{(k-1)}, ...\theta_{d-1}^{(k-1)}\right)$$

(iii) Increase counter from k to $k+1$ and return to step (ii).

The Gibbs sampler should be run after initialising the sampler somewhere in the support of θ. The resulting chain will converge, after an initial "burn-in" period, to the posterior distribution (Tosh and Dasgupta, 2014).

Metropolis-Hastings sampling

The Gibbs sampler simulate from multivariate distributions provided that the full conditionals can be simulated. In many cases it is not straightforward to simulate from full conditionals. The Metropolis-Hastings sampling techniques provides a straightforward way of simulating from full conditionals (Korattikara et al., 2014; Hastings, 1970; Metropolis et al., 1953). Given a distribution of interest, π, a reversible Markov chain, which has this distribution as its stationary distribution, can be constructed.

The procedure is to construct a transition kernel $p(\theta, \phi)$ such that the equilibrium distribution of the chain is π. The transition kernel is made up of two elements namely the arbitrary transition kernel $q(\theta, \phi)$ and the acceptance probability $a(\theta, \phi)$. The Metropolis-hastings algorithm is

(i) Initialize the counter to $k = 1$ and the state of the chain to $\theta^{(0)}$.

(ii) Simulate a proposed value ϕ using the kernel $q(\theta, \phi)$.

(iii) Find the acceptance probability of proposed value $a(\theta^{(k-1)}, \phi)$.

(iv) Accept $\theta^{(k)} = \phi$ with probability $a(\theta^{(k-1)}, \phi)$ and take $\theta^{(k)} = \theta^{(k-1)}$ otherwise.

(v) Increase the counter from k to $k+1$ and return to step (ii).

MCMC algorithm

In this thesis, the density of interest is the posterior

$$\pi(\sigma, \mu, \xi \mid \mathbf{x}) \propto \pi_\phi(\phi)\pi_\mu(\mu)\pi_\xi(\xi)L(\boldsymbol{\theta} \mid \mathbf{x}) \tag{6.40}$$

with σ replaced by e^ϕ. So the full conditionals are of the form:

$$\pi(\phi \mid \mu, \xi) = \pi_\phi(\phi)L(\boldsymbol{\theta} \mid \mathbf{x})$$

$$\pi(\mu \mid \phi, \xi) = \pi_\mu(\mu)L(\boldsymbol{\theta} \mid \mathbf{x})$$

$$\pi(\xi \mid \mu, \sigma) = \pi_\xi(\xi)L(\boldsymbol{\theta} \mid \mathbf{x})$$

The three transition densities are denoted by q_ϕ, q_μ and q_ξ, and the proposed values for each variable are given by

$$\phi^* = \phi + \omega_\phi,$$

$$\mu^* = \mu + \omega_\mu,$$

$$\xi^* = \mu + \omega_\xi,$$

where ω_ϕ, ω_μ and ω_ξ are normally distributed with mean zero and high variances. The variances are high to make the distributions almost flat, corresponding to prior ignorance. For more details of the MCMC algorithm refer to R package: evdbayes version1.1-1.

6.2.5 Non-stationary GEVD model

Since the original work of Dalrymple (1960), almost all regional extreme rainfall analysis assumed that observations are independent, homogeneous and stationary (Nguyen et al., 2014). In the context of meteorological processes, non-stationarity is often apparent because of seasonal effects, perhaps due to long-changing climatic conditions. The models that were introduced in Section 6.2 and 6.3 assumed that

the observations used were independent and identically distributed. In Chapter 3, we found out that the mean annual rainfall for Zimbabwe is influenced by SOI values for September of the previous year and a component of SDSLP values for March which are not explained by SOI values. In this section we discuss the theory of non-stationary GEVD models. In the non-stationary case, the parameters are expressed as a function of covariates such as time: GEVD(μ_t, σ_t, ξ_t) (Coles et al., 2001). To ensure a positive value for the scale parameter σ, a transformation such that $\varphi_t = \log(\sigma_t)$ is used when estimating the parameters. Using the notation GEVD(μ, σ, ξ) to denote the GEVD with parameters μ,σ and ξ, it follows that a suitable model for X_t, maxima mean annual rainfall where the location parameter μ and σ are allowed to vary linearly with time t is given by:

$$X(t) \sim GEVD(\mu_t, \sigma_t, \xi) \tag{6.41}$$

where

$$\begin{aligned} \mu_t &= \theta_0 + \theta_1 t, \\ \text{with } \sigma_t &= \exp(\sigma_{01} + \sigma_{01} t), \quad \text{and } \xi = \text{constant} \end{aligned} \tag{6.42}$$

The log-likelihood of the non-stationary GEVD is:

$$\begin{aligned} l_{ns}(\mu_t, \sigma_t, \xi) = &- m\ln(\sigma_t) - \left(1 + \frac{1}{\xi}\right) \sum_{i=1}^{m} \ln\left[1 + \xi\left(\frac{x_i - \mu_t}{\sigma_t}\right)\right] \\ &- \sum_{i=1}^{m} \left[1 + \xi\left(\frac{x_i - \mu_t}{\sigma_t}\right)\right]^{-\frac{1}{\xi}} \end{aligned} \tag{6.43}$$

The maximum likelihood estimators are the solution of an equation system formed by setting the derivatives of l_{ns} with respect to each parameter to zero. For the non-stationary GEVD with only the location parameter μ allowed to vary linearly with time and $\sigma_t = \sigma$ (constant), the derivative of the log-likelihood with respect to each

parameter $\xi, \sigma, \theta_0, \theta_1$ are:

$$\begin{aligned}\frac{dl_{ns}}{d\xi} &= \frac{1}{\xi^2}\sum_{i=1}^{m}\ln\left[1+\xi\left(\frac{x_i-\mu_t}{\sigma}\right)\right]-\left(1+\frac{1}{\xi}\right)\sum_{i=1}^{m}\frac{x_i-\mu_t}{\sigma+\xi(x_i-\mu_t)}\\ &\quad-\frac{1}{\xi^2}\sum_{i=1}^{m}\sum_{i=1}^{m}\left[1+\xi\left(\frac{x_i-\mu_t}{\sigma}\right)\right]^{-\frac{1}{\xi}}\ln\left[1+\xi\left(\frac{x_i-\mu_t}{\sigma}\right)\right]\\ &\quad+\frac{1}{\xi}\sum_{i=1}^{m}\sum_{i=1}^{m}\left[1+\xi\left(\frac{x_i-\mu_t}{\sigma}\right)\right]^{-\frac{1}{\xi}}\left[\frac{x_i-\mu_t}{\sigma+\xi(x_i-\mu_t)}\right] \qquad (6.44)\\ \frac{dl_{ns}}{d\sigma} &= -\frac{m}{\sigma}-\left(1+\frac{1}{\xi}\right)\sum_{i=1}^{m}\frac{\xi(x_i-\mu_t)}{\sigma[\sigma+\xi(x_i-\mu_t)]}\\ &\quad-\sum_{i=1}^{m}\sum_{i=1}^{m}\left[1+\xi\left(\frac{x_i-\mu_t}{\sigma}\right)\right]^{-\frac{1}{\xi}}\frac{\xi(x_i-\mu_t)}{\sigma[\sigma+\xi(x_i-\mu_t)]} \qquad (6.45)\\ \frac{dl_{ns}}{d\theta_0} &= \frac{\xi+1}{\sigma}\sum_{i=1}^{m}\left[1+\xi\left(\frac{x_i-\mu_t}{\sigma}\right)\right]^{-1}-\frac{1}{\sigma}\sum_{i=1}^{m}\left[1+\xi\left(\frac{x_i-\mu_t}{\sigma}\right)\right]^{-1-\frac{1}{\xi}} \qquad (6.46)\\ \frac{dl_{ns}}{d\theta_1} &= \frac{\xi+1}{\sigma}\sum_{i=1}^{m}t\left[1+\xi\left(\frac{x_i-\mu_t}{\sigma}\right)\right]^{-1}-\frac{1}{\sigma}\sum_{i=1}^{m}t\left[1+\xi\left(\frac{x_i-\mu_t}{\sigma}\right)\right]^{-1-\frac{1}{\xi}} \qquad (6.47)\end{aligned}$$

respectively (see Section 6.6 for the derivations). By setting the system of equations i.e. (6.44) to (6.47) to zero and applying the numerical optimization algorithms such as Quasi-Newton numerical maximization, the maximum likelihood estimates are obtained for $\xi, \sigma, \theta_0, \theta_1$.

If we allow the location parameter to vary with different climatic change drivers such as SOI and SDSLP values (6.41) becomes

$$X(k) \sim \mathrm{GEVD}(\mu_k, \sigma, \xi) \qquad (6.48)$$

where

$$\mu_k = \theta_0 + \theta_1 k, \qquad (6.49)$$

and k is the climatic change driver value i.e. $\mathrm{SOI}_{\mathrm{MAY}}$, $\mathrm{SOI}_{\mathrm{AUGUST}}$, $\mathrm{SDSLP}_{\mathrm{APRIL}}$ or $\mathrm{SDSLP}_{\mathrm{AUGUST}}$ (see Chapter 3 for details of correlation between the weather/climatic

change drivers and annual rainfall for Zimbabwe).

6.2.6 Modelling minima random variables

The classical GEVD for extremes is based on asymptotic approximations to the sampling behaviour of block maxima. If $\widetilde{M_n} = \min\{X_1, ..., X_n\}$, where X_i denote the independent and identically distributed random variable and if $Y_i = -X_i$ for $i = 1, ..., m$,

$$\widetilde{M}_n = \max\{Y_1, ..., Y_n\} \tag{6.50}$$

then $\widetilde{M}_n = -M_n$. Hence, for large n,

$$\begin{aligned} P_r\left\{\widetilde{M}_n \leq x\right\} &= P_r\left\{-M_n \leq x\right\} \\ &= P_r\{M_n \geq -x\} \\ &= 1 - P_r\{M_n \leq -x\} \\ &\approx 1 - \exp\left\{-\left[1 + \xi\left(\frac{-x-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\} \\ &= 1 - \exp\left\{-\left[1 - \xi\left(\frac{x-\widetilde{\mu}}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\} \end{aligned} \tag{6.51}$$

defined on $\{x : 1 - \xi(x - \widetilde{\mu})/\sigma > 0\}$, where $\widetilde{\mu} = -\mu$. This result is a GEVD for minima. Which we can state as a theorem (Extremal types theorem for block minima).

Theorem 6.3 *If there exist sequences of constants $\{a_n > 0\}$ and $\{b_n\}$ such that*

$$P_r\left\{\left(\frac{\widetilde{M}_n - b_n}{a_n}\right) \leq x\right\} \to \widetilde{F}(x) \quad \text{as} \quad n \to \infty \tag{6.52}$$

for a nondegenerate distribution function $\widetilde{F}$, then $\widetilde{F}$ is a member of the generalised extreme

value family of distributions for minima:

$$\widetilde{F} = 1 - \exp\left\{ - \left[1 - \xi \left(\frac{x - \widetilde{\mu}}{\sigma} \right) \right]^{-\frac{1}{\xi}} \right\} \tag{6.53}$$

defined on $\{x : 1 - \xi(x - \widetilde{\mu})/\sigma > 0\}$, *where* $\widetilde{\mu} \in \Re$, $\sigma > 0$ *and* $-\infty < \xi < \infty$.

The maximum likelihood estimates can then be obtained. An alternative method to model minima random variables is to use the duality between the distributions for maxima and minima. If $M_n = \min\{X_1, ..., X_n\}$ and if $Y_i = -X_i$ for $i = 1, ..., m$, then the change of sign means the small values of X_i corresponds to large values of Y_i. So if

$$M_n = \min\{X_1, ..., X_n\}, \tag{6.54}$$

and

$$\widetilde{M_n} = \max\{Y_1, ..., Y_n\} \tag{6.55}$$

then $M_n = -\widetilde{M_n}$. The minima becomes:

$$\widetilde{M_n} = \max\{-X_1, ..., -X_n\} \tag{6.56}$$

Extreme maxima theory and methods discussed in Section 6.2 are then used to model extreme minima (Coles et al., 2001; Gomes and Guillou, 2015).

6.2.7 The models

In this chapter, we model maxima mean annual rainfall. We fit the basic model (6.2.8) with μ, σ and ξ constant (to be referred to as Model 1). To investigate the existence of a trend in the extreme rainfall with respect to time, we fit a four-parameter GEVD model, which is the variation of Model 1 where the location parameter μ is allowed

to vary linearly with time (to be referred to as Model 2).

$$\text{Model } 2 : \mu_t = \theta_0 + \theta_1 t \text{ with } \sigma = \text{constant, and } \xi = \text{constant}, \tag{6.57}$$

where t is the year i.e. $t = 1, 2, ..., 109$.

In Chapter 3, we observed that mean annual rainfall for Zimbabwe correlates with $\text{SOI}_{\text{May}}, \text{SOI}_{\text{August}}, \text{SDSLP}_{\text{April}}$ and $\text{SDSLP}_{\text{August}}$. To investigate the influence of SOI and SDSLP on extreme rainfall, we fit a non-stationary GEVD model, which is the variation of Model 1 where the location parameter μ is allowed to vary linearly with SOI_{May} (to be referred to as Model 3), $\text{SOI}_{\text{August}}$ (to be referred to as Model 4), $\text{SDSLP}_{\text{April}}$ (to be referred to as Model 5) and $\text{SDSLP}_{\text{August}}$ (to be referred to as Model 6) respectively. That is:

$$\text{Model } i : \ \mu_i = \theta_0 + \theta_1 k \text{ with } \sigma = \text{constant, and } \xi = \text{constant}, \tag{6.58}$$

where k is the weather/climatic change variable i.e. $\text{SOI}_{\text{May}}, \text{SOI}_{\text{August}}, \text{SDSLP}_{\text{April}}$ and $\text{SDSLP}_{\text{August}}$ and $i = 3, 4, 5$ and 6.

6.2.8 Return level estimates

Estimates of extreme quantiles of the annual maxima or minima are of particular interest in meteorological extremes, as they give an estimate of the level the process is expected to exceed once, on average, in a given number of years (T) with a probability of p. The quantiles are obtained by inverting (6.11) and (6.12):

$$Q_p = \begin{cases} \mu + \frac{\sigma}{\xi}\left[[-\ln(1-p)]^{-\xi} - 1\right], & \xi \neq 0, \\ \mu - \sigma \ln\left[-\ln(1-p)\right], & \xi = 0, \end{cases} \tag{6.59}$$

The quantity Q_p is known as the return level associated with the T year return period (note that $p = 1/T$). To obtain estimates of return levels beyond the range of the data used, the relationships given in (6.59) is extrapolated. One way of checking

whether the extrapolated return levels are significant is to use the return level plots. A return level plot is a plot of the estimated quantiles (Q_p) against return period, usually shown on a logarithmic scale.

The confidence intervals for the return level estimates are obtained by employing the profile likelihood method. The $100(1-\alpha)\%$ confidence interval for Q_p is given by the set:

$$\left\{\theta : 2\log\left(\frac{L(\hat{\mu},\hat{\sigma},\hat{\xi})}{L_m(\theta)}\right) < \chi^2_{1,1-\alpha}\right\}, \tag{6.60}$$

where $\chi^2_{1,1-\alpha}$ denotes $100(1-\alpha)\%$ quantile of the chi-square distribution with one degree of freedom. $L_m(\theta)$ is obtained by re-parameterizing the likelihood (6.29) in terms of (Q_p, σ, ξ) and by calculating the profile likelihood $L_m(Q_p)$ given by:

$$L_m(Q_p) = \max_{\sigma,\xi} L(Q_p, \sigma, \xi). \tag{6.61}$$

6.2.9 Model diagnostics

After estimating the parameters of GEVD, we checked for model adequacy using the AD test statistic, the PP plots and the QQ plots. In Chapter 2 we discussed the AD test, PP and QQ plots for model adequacy checking. We now describe the PP and QQ-plots in more detail for checking the model validity of a GEVD model.

The PP plot is a comparison of the empirical and fitted distribution functions. With ordered block maxima data $x_1 \leq x_2 \leq ... \leq x_m$, the empirical function evaluated at x_i is

$$\widetilde{F}(x_i) = i/(m+1). \tag{6.62}$$

If we substitute parameter estimates into (6.29), the corresponding model is

$$\hat{F}(x_i) = \exp\left\{-\left[1+\hat{\xi}\left(\frac{x_i-\hat{\mu}}{\hat{\sigma}}\right)\right]^{-\frac{1}{\hat{\xi}}}\right\},$$

and if the GEVD model is working well, then

$$\hat{F}(x_i) \approx \widetilde{F}(x_i)$$

for each i. Therefore, a probability plot, consisting of the points

$$\left\{\left(\hat{F}(x_i), \widetilde{F}(x_i)\right), \;\; i = 1, ..., m\right\}, \tag{6.63}$$

should lie close to the unit diagonal. Any significant departures from linearity are indicative of the GEVD model not fitting well to the data.

The major weakness of the probability plot for extreme value models is that both $\hat{F}(x_i)$ and $\widetilde{F}(x_i)$ are bound to approach 1 as x_i increases, while it is usually the accuracy of the model for large values of x that is of greatest concern (Coles et al., 2001). The weakness of the probability plot is avoided by the quantile plot, consisting of the points

$$\left\{\left(\hat{F}^{-1}(i/(m+1)), x_i\right)\right) \quad i = 1, ..., m\right\},$$

where from (6.59),

$$\hat{F}^{-1}\left(\frac{i}{m+1}\right) = \hat{\mu} + \frac{\hat{\sigma}}{\hat{\xi}}\left[\left[-\ln\left(\frac{i}{m+1}\right)\right]^{-\hat{\xi}} - 1\right]. \tag{6.64}$$

Any significant departures from linearity in the quantile plot also indicates model inadequacy.

6.2.10 Model selection

The deviance (D) statistic will be used to select the best model. With models $M_1 \subset M_2$, then we will define the D statistic as:

$$D = 2[l_2(M_2) - l_1(M_1)] \tag{6.65}$$

where $l_2(M_2)$ and $l_1(M_1)$ are the maximised log-likelihood under models M_2 and M_1 respectively. The D statistic follows a chi-square distribution with k degrees of freedom, χ^2_k. k is the difference in dimensionality of models M_2 and M_1. Thus, the calculated values of the D statistic are compared to critical values from χ^2_k at 5% level of significance. Large values of the D statistic suggest that model M_2 explains more variation in the data than M_1.

Goodness-of-fit tests namely; RRMSE, RMAE and PPCC will also be used to check for goodness-of-fit and model selection. In (6.22), we substitute p with F_i, where Landwehr plotting position $F_i = \frac{i-0.35}{n}$, to calculate (2.21), (2.22) and (2.23) for the GEVD models.

In this chapter we use the AD statistic and the quantile plot to check for model adequacy. The D statistic, RRMSE, RMAE and PPCC statistics are used to check for goodness-of-fit and model selection.

6.3 Empirical results

In this section, the results of modelling maximum annual rainfall using the maximum likelihood method and Bayesian framework are discussed. To fit models 1 to 6, we need to test whether mean annual rainfall data is independent and identically distributed (i.i.d.). In Chapter 3 we described the source and the properties of the data set used in this chapter. In Chapter 3, it was also shown that the mean annual rainfall data are i.i.d., thus we can fit the data to the GEVD.

6.3.1 The Maximum likelihood estimation of the annual maxima rainfall data

Table 6.1 shows the maximum likelihood estimates of the GEVD with their corresponding standard errors in brackets and negative log-likelihood (NLL) value for Model 1.

Table 6.1: Maximum likelihood estimates (standard errors) of Model 1

$\hat{\mu}$	$\hat{\sigma}$	$\hat{\xi}$	NLL value
590.3810	152.8543	-0.1407	711.3708
(16.2924)	(11.4984)	(0.0630)	

The results show that data can be modelled using a Weibull class of distribution because $\hat{\xi} = -0.1407 < 0$ (bounded tail). Combining estimates and their corresponding standard errors, the 95% confidence intervals for μ, σ and ξ are $[558.4401; 606.0255]$, $[130.3174; 175.3912]$ and $[-0.2642; -0.0172]$ respectively.

It is important to confirm that the data adequately fits the stationary GEVD. Figure 6.2 shows the diagnostic plots for the goodness of fit of the GEVD fitted using maximum likelihood method to mean annual rainfall for Model 1.

The quantiles of maxima rainfall regressed against the quantiles of GEVD shows a straight line. The diagnostic plots show each set of plotted points to be near linear, validating use of the GEVD. Also plotted is a 95% confidence interval of the return level estimates, which are wider for long return periods. This finding suggests that the data does not deviate significantly from the assumption that they follow a GEVD. This is confirmed by the Anderson-Darling statistic (p-value$= 0.8725 > 0.05$) which fails to reject the null hypothesis. We therefore conclude that the maximum annual rainfall for Zimbabwe follows the specified GEVD (Model 1).

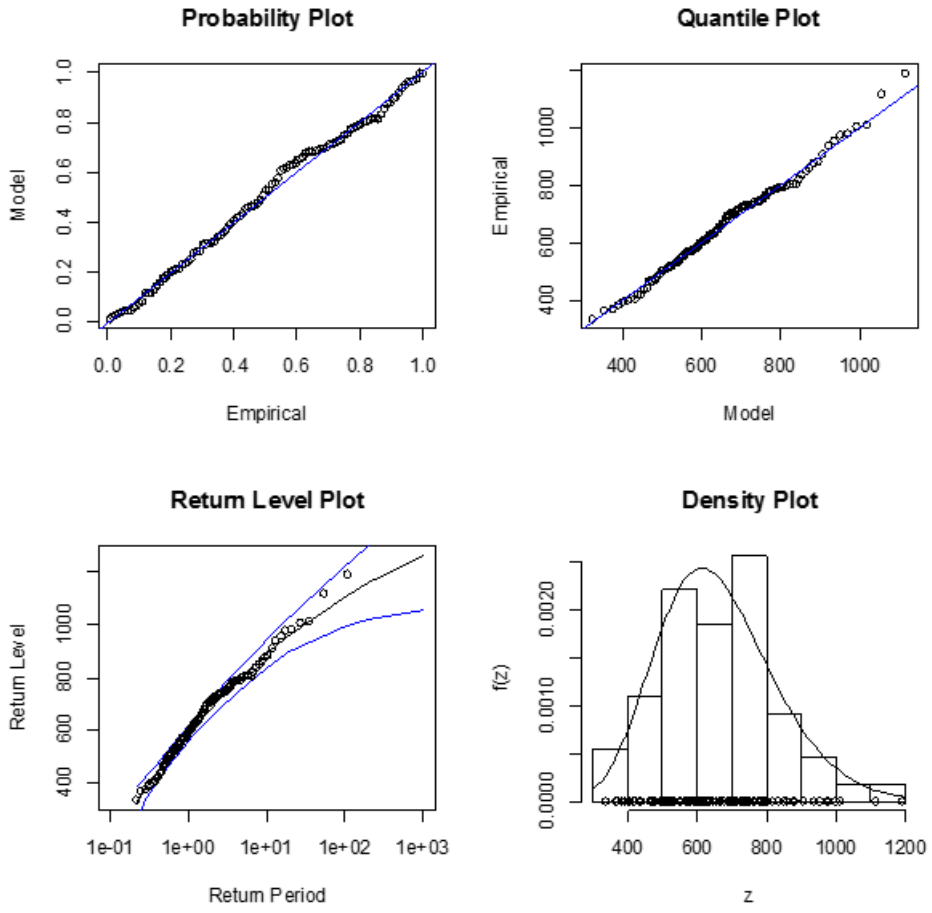

Figure 6.2: Diagnostic plots illustrating the fit of the mean annual rainfall data for Zimbabwe to the GEVD for Model 1, (a) Probability plot (top left panel), (b) Quantile plot (top right panel), (c) Return level plot (bottom left panel) and (d) Density plot (Bottom right panel)

The maximum likelihood for ξ is significantly different from zero because zero is not contained in the 95% confidence interval. Greater accuracy of the confidence interval is achieved by the use of the profile likelihood. Figure 6.3 shows the profile likelihood of the generalised extreme value parameter ξ, from which a 95% confidence interval for ξ is obtained as approximately $[-0.2400; 0]$ which is almost the same as the calculated 95% confidence interval. This suggests that the GEVD model 1 of the Weibull class of distribution is a good fit to the maxima mean annual rainfall.

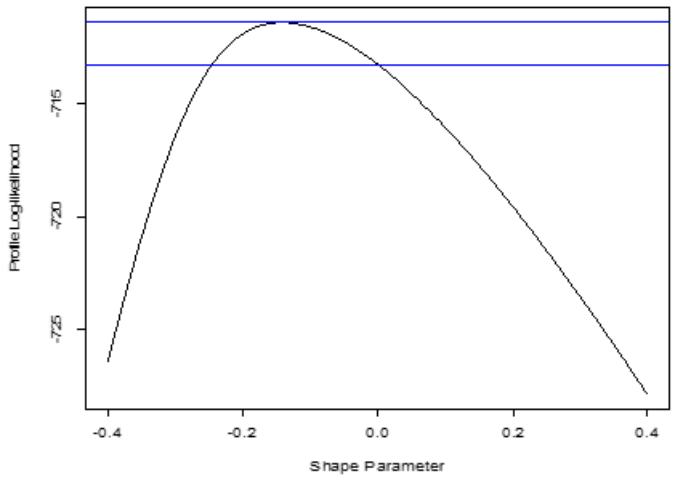

Figure 6.3: Profile likelihood for the generalised parameter shape for Model 1

Bayesian analysis of maxima annual rainfall data

We now seek more informative estimates of the GEVD model 1 based on the prior information. The Markov Chain Monte Carlo (MCMC) method is applied to the annual maxima rainfall data. In the absence of any expert information on the rainfall pattern for Zimbabwe, the only genuine prior information available in the GEVD parameters will be non-informative distributions with large variances. The GEVD scale parameter was re-parameterised as $\phi = \log(\sigma)$ to retain the positivity of this parameter. The prior density was chosen to be

$$\pi(\mu, \phi, \xi) = \pi_\mu(\mu)\pi_\phi(\phi)\pi_\xi(\xi),$$

where the marginal priors, $\pi_\mu(.)$, $\pi_\phi(.)$ and $\pi_\xi(.)$ are

$$\mu \sim N(0, 400000)$$

$$\phi \sim N(0, 400000)$$

$$\xi \sim N(0, 10000)$$

for the three parameters of the GEVD, where, for example, $N(0, 400000)$ denotes a Gaussian distribution with mean 0 and variance, 400000. These are independent normal priors with large variances. The variances are chosen large enough to make the distributions almost flat, corresponding to prior ignorance. Initializing the MCMC algorithm with maximum likelihood estimates as our initial vector $\theta_0 = (590, 152, -0.14)$, should produce a chain with small burn-in period. After some pilot runs, the proposed standard deviations were taken to be $\mathbf{s} = (0.25, 0.1, 0.1)$ and a Markov chain of 30000 iterations were carried out, of which the first 1000 were discarded. The remaining simulations were then regarded as realizations of the marginal distributions of the posterior. Figure 6.4 shows the MCMC trace plots.

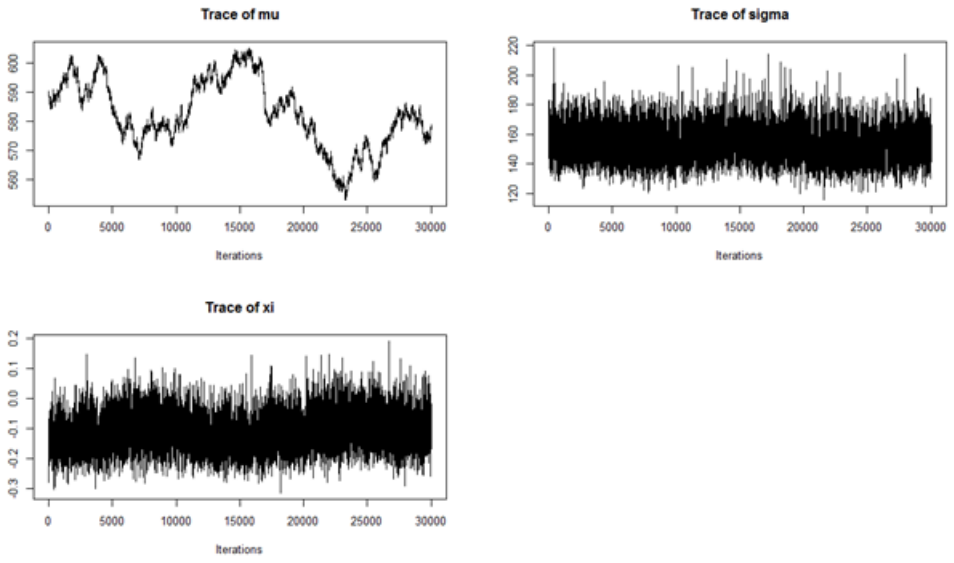

Figure 6.4: Trace plots of the GEVD parameters using non-informative priors for maxima annual rainfall.

To check that the chains had converged to the correct place, different starting points were used. The chains all converged. The estimated posterior densities for the GEVD parameters for Zimbabwe are given in Figure 6.5.

The posterior means and standard deviations for the GEVD parameters are given in

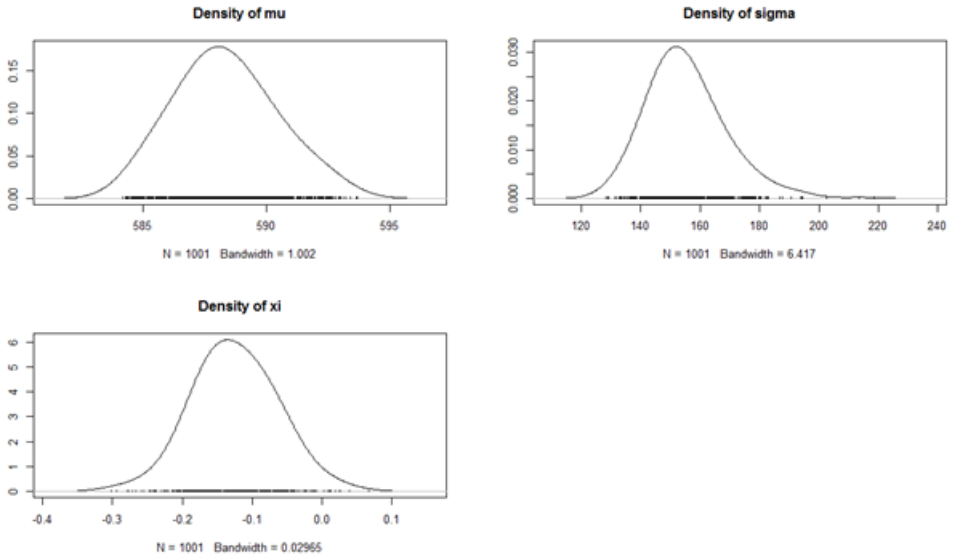

Figure 6.5: Posterior densities of the GEVD parameters using non-informative priors for maximum annual rainfall for Zimbabwe for the period 1901-2009.

Table 6.2. Using non-informative priors which are almost flat and add very little information to the likelihood, the posterior means are close to the maximum likelihood estimates of the GEVD parameters given in Table 6.1 but with smaller standard deviations. The frequentist properties are preserved by using non-informative priors in the Bayesian statistics approach.

Table 6.2: Posterior means (standard deviations) of the GEVD Model 1 parameters

$\hat{\mu}$	$\hat{\sigma}$	$\hat{\xi}$
588.3402	155.3500	-0.1213
(2.0100)	(12.8300)	(0.0621)

Return level estimates

The return levels/periods are estimated using the GEVD Model 1 parameters estimated by maximum likelihood estimation (ML) method and the Bayesian approach. In Table 6.3 we show the return level estimates (in mm of rainfall) at selected return

intervals T in years.

Table 6.3: Return level estimate from the GEVD model 1

Method	$T = 5$	$T = 10$	$T = 15$	$T = 20$	$T = 25$	$T = 30$	$T = 50$	$T = 100$
ML	797.0761	885.2228	930.9982	961.4641	984.0796	1001.9494	1049.3508	1108.0562
Bayesian	801.3873	894.2811	943.0781	975.7832	1000.1849	1019.5439	1071.2444	1136.0306

We observe that the return level estimates based on the Bayesian approach are slightly higher than their corresponding return level estimates based on the maximum likelihood estimation method. Using the parameter estimates based on the Bayesian approach, the return level of 1193 mm is associated with a mean return period of around 190 years. The maximum mean annual rainfall for Zimbabwe is 1192.6 mm recorded in 1923/24 rainfall season. This suggests that an extreme flood of this magnitude is likely to return once in every 190 years. Recent floods in many parts of Zimbabwe, as reported by Chingombe et al. (2015), Gwindi (2009) and Madamombe (2004), occurred when the annual rainfall are around 775 mm or above. The return level of 775 mm is associated with a mean return period of around 4 years. This also suggests that floods of this magnitude is expected once in every 4 years. The posterior return level plot in Figure 6.6 shows the upper 95% limit to be further from the mean than the lower limit. This is because of the heavier upper tail of the posterior distribution.

We explore the possibility of time (Model 2) and different natural drivers of climatic change (Models 3, 4, 5 and 6) influencing extreme rainfall for Zimbabwe. We fit Model 2, Model 3, Model 4, Model 5 and Model 6 to mean annual rainfall. Table 6.3 shows the maximum likelihood parameter estimates with standard errors in brackets and negative log-likelihood (NLL) values for the fitted models.

The residual probability plot of Model 2 presented in Figure 6.6 (left panel) (Appendix 6.2) shows some slight departure from the straight line. This suggests that the extreme annual rainfall for Zimbabwe does not trend with time. We use the

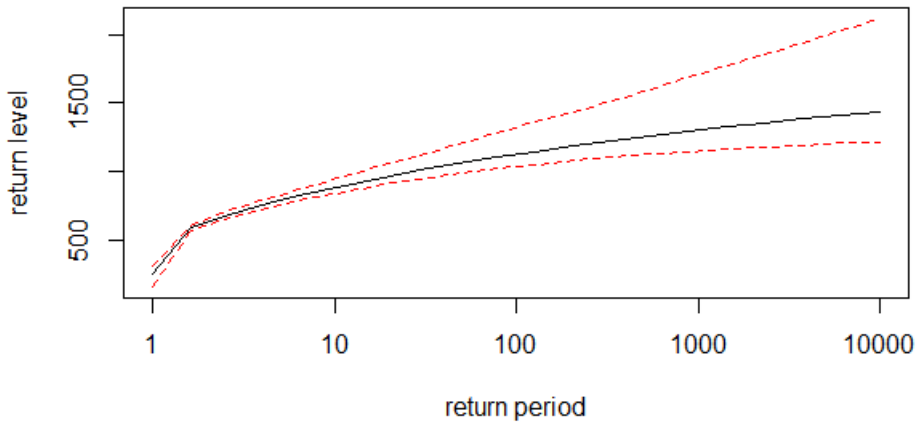

Figure 6.6: Posterior return level plot in a Bayesian analysis of the Zimbabwean rainfall data. The curves represent means (solid line) and intervals containing 95% of the posterior probability (dashed lines).

D statistic to check whether allowing the location parameter to trend with time is worthwhile. The model pair (M_1, M_2) had a D statistic of $2[-711.1775 + 711.3708] = 0.3866 < \chi^2_{1,0.05} = 3.841$, we conclude that there is no evidence of significant trends with respect to time. The residual probability and residual quantile plots for model 3 to Model 6, shown in Figure 6.7 to Figure 6.9 respectively (Appendix 6.2), shows no significant departure from the straight line. This suggests that climatic variables appear to influence the extreme annual rainfall for Zimbabwe. We check whether there is significant improvement to the stationary model (Model 1) if we allow the location parameter to trend with climatic variables using the D statistic. The model pair (M_1, M_3) had a D statistic of $2[-707.4624 + 711.3708] = 7.8168 > \chi^2_{1,0.05} = 3.841$, we conclude that there is a significant influence of SOI value for May on the extreme annual rainfall for Zimbabwe. The model pair (M_1, M_4) had a D statistic of $2[-701.1743 + 711.3708] = 20.3930 > \chi^2_{1,0.05} = 3.841$, we conclude that there is a significant influence of the SOI value for August on the extreme annual rainfall for Zimbabwe. The model pair (M_1, M_5) had a D statistic of $2[-707.9489 + 711.3708] = 6.8438 > \chi^2_{1,0.05} = 3.841$, we conclude that there is a significant influence of SDSLP

Table 6.4: The maximum likelihood parameter estimates (standard errors) and negative log-likelihood values of non-stationary GEVD Models

Model	$\hat{\theta}_0$	$\hat{\theta}_1$	$\hat{\sigma}$	$\hat{\xi}$	NLL value
2	607.7044	-0.3055	152.9264	-0.1455	711.1775
	(32.1209)	(0.4916)	(11.5560)	(0.0645)	
3	596.4754	4.2380	146.9625	-0.1433	707.4624
	(16.0620)	(1.5073)	(11.5544)	(0.0749)	
4	603.9008	6.6728	141.8688	-0.1681	701.1743
	(15.0597)	(1.4221)	(10.3929)	(0.0575)	
5	579.7127	-43.7269	148.3106	-0.1421	707.9489
	(16.3725)	(16.5153)	(11.1393)	0.0627	
6	570.4727	-55.6099	144.8190	-0.1330	705.9603
	(16.61930)	(16.3992)	(10.8993)	0.0632	

anomalies for April on the extreme annual rainfall for Zimbabwe. The model pair (M_1, M_6) had a D statistic of $2[-705.9603 + 711.3708] = 10.9603 > \chi^2_{1,0.05} = 3.841$, we conclude that there is a significant influence of SDSLP anomalies for August on the extreme annual rainfall for Zimbabwe.

It is important to select a robust model for modelling extreme maximum annual rainfall for Zimbabwe. The goodness-of-fit tests namely; RRMSE, RMAE and PPCC values are used to select the most robust model. We use the i^{th} Landwehr plotting position $F_i = \frac{i-0.35}{n}$, in place of p in (6.2.18) and

$$\hat{\mu} = \hat{\theta}_0 + \hat{\theta}_1 \times \text{value of the weather/climatic variable}$$

to estimate $Q(F_i)$ for the models. Table 6.6 shows the goodness-of-fit values of the non-stationary models.

The best model at least twice out of the three goodness-of-fit tests is selected as the best performing model. Results indicate that Model 5 is the best performing GEVD

Table 6.5: Goodness-of-fit test results for GEVD models with location parameter influenced by weather/climatic variable

Model	RRMSE	RMAE	PPCC
3	**0.0325**	0.0543	0.9077
4	0.1019	0.0758	0.9335
5	0.0662	**0.0503**	**0.9758**
6	0.0828	0.0622	0.9712

model since it has the least RMAE and the highest PPCC values. Thus, this study proposes that the best performing GEVD model for mean annual rainfall for Zimbabwe is:

$$F(x) = \exp\left\{-\left[1 - 0.1421\left(\frac{x - (579.7127 + 43.7269\text{SDSLP}_{\text{APRIL}})}{148.3106}\right)\right]^{\frac{1}{0.1421}}\right\}$$

where x is the mean annual rainfall for Zimbabwe. According to Cheng et al. (2014) estimating of return levels using non-stationary GEVD models is challenging since the behaviour of physically-based covariates is not constant. In the presence of $\text{SDSLP}_{\text{APRIL}}$ values of the next 100 or more years, the return level estimates of maxima mean annual rainfall for Zimbabwe can be calculated.

6.4 Concluding remarks

The purpose of this chapter was to fit the stationary generalised extreme value distributions using the maximum likelihood estimation method and Bayesian statistics approach to mean annual rainfall using annual rainfall data from year 1901 to 2009. Results obtained using the maximum likelihood estimation method were compared with those obtained using non-informative priors Bayesian approach. The distribution of the block maxima was explored by fitting the GEVD. The GEVD model with shape and scale parameters constant, but allowing the location parameter to vary with time, was found to be inadequate to model extreme maximum rainfall in Zimbabwe. Model diagnostics which include the probability plot, quantile plot and AD

tests show that the simple GEVD model (Model 1) fits the data well. The confidence interval of the shape parameter ξ using the profile likelihood contains a zero. This suggests that the GEVD of the Weibull class of distributions fits the maxima mean annual rainfall well. Return level estimates, which is the return level expected to be exceeded in a certain period of time T in years were calculated for Zimbabwe rainfall. The highest mean annual rainfall amount recorded for the country is 1192.6 mm. A return level of 1193 mm is associated with a mean return period of 190 years while a return level of 775 mm is associated with a mean return period of around 4 years using the Bayesian approach. Recent flooding has occurred in many parts of the country with a mean annual rainfall of 775 mm and above. The GEVD parameter estimates using the Bayesian approach, are close to the maximum likelihood estimates with smaller standard deviations. This improves the precision of the estimates. Thus, using non-informative priors, the expected benefit of the Bayesian analysis is the improvement in the precision of parameter estimates over maximum likelihood estimates is realised.

However, substantial evidence shows that climate is non-stationary, possibly due to anthropogenic climate change (Cheng, et al., 2014). We explored the possibility of extreme annual rainfall for Zimbabwe influenced by different natural and anthropogenic drivers. We allowed the location parameter to be influenced by SOI and SDSLP values. The GEVD models which allow the location parameter to be influenced by natural drivers perform better than the stationary GEVD model. Thus, this study proposes the GEVD model with location parameter influenced by SDSLP anomalies of April as the best performing model for mean annual rainfall for Zimbabwe. This proposed model should be used only as a early warning tool for floods in Zimbabwe. For climate change drivers-varying GEVD models to be accepted as a simple extrapolation of results from historical trends, one has to rely on the inertia of the climate system. In order to understand when the inertia stops, one has to rely on other tools such as dynamic models. Therefore, extrapolating into the future using

historical trends is risky. Furthermore, we have incorporated one natural driver as a covariate in our proposed model, but many different natural and anthropogenic weather/climate change drivers can be added as covariates to improve the model. Therefore, an area of further research is modelling maximum annual rainfall in Zimbabwe using many weather/climate change drivers.

6.5 Appendix

Proof of maximum likelihood estimates for stationary GEVD model parameters

$$f_{\mu,\sigma,\xi}(x) = \frac{1}{\sigma}\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-1-\frac{1}{\xi}} \exp\left\{-\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\}$$

the log-likelihood of the GEVD when $\xi \neq 0$ is:

$$l(\mu,\sigma,\xi) = -m\ln\sigma - \left(1+\frac{1}{\xi}\right)\sum_{i=1}^{m}\ln\left[1+\xi\left(\frac{x_i-\mu}{\sigma}\right)\right] - \sum_{i=1}^{m}\left[1+\xi\left(\frac{x_i-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}$$

Differentiating l with respect to ξ

To differentiate $\sum_{i=1}^{m}\left[1+\xi\left(\frac{x_i-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}$ we let

$$y = \sum_{i=1}^{m}\left[1+\xi\left(\frac{x_i-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}$$

taking natural logarithm on both sides

$$\begin{aligned}
\ln y &= \ln\sum_{i=1}^{m}\left[1+\xi\left(\frac{x_i-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}} \\
&= \sum_{i=1}^{m}\ln\left[1+\xi\left(\frac{x_i-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}} \\
&= -\frac{1}{\xi}\sum_{i=1}^{m}\ln\left[1+\xi\left(\frac{x_i-\mu}{\sigma}\right)\right]
\end{aligned}$$

Now we differentiate implicitly

$$\begin{aligned}
\frac{1}{y}\frac{dy}{d\xi} &= \frac{1}{\xi^2}\sum_{i=1}^{m}\ln\left[1+\xi\left(\frac{x_i-\mu}{\sigma}\right)\right] - \frac{1}{\xi}\sum_{i=1}^{m}\frac{x_i-\mu}{\sigma}\cdot\frac{1}{1+\xi\left(\frac{x_i-\mu}{\sigma}\right)} \\
&= \frac{1}{\xi^2}\sum_{i=1}^{m}\ln\left[1+\xi\left(\frac{x_i-\mu}{\sigma}\right)\right] - \frac{1}{\xi}\sum_{i=1}^{m}\frac{x_i-\mu}{\sigma+\xi(x_i-\mu)}
\end{aligned}$$

Therefore

$$\begin{aligned}\frac{dy}{d\xi} =& y\left[\frac{1}{\xi^2}\sum_{i=1}^{m}\ln\left[1+\xi\left(\frac{x_i-\mu}{\sigma}\right)\right]-\frac{1}{\xi}\sum_{i=1}^{m}\frac{x_i-\mu}{\sigma+\xi(x_i-\mu)}\right]\\ =&\sum_{i=1}^{m}\left[1+\xi\left(\frac{x_i-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\left[\frac{1}{\xi^2}\sum_{i=1}^{m}\ln\left[1+\xi\left(\frac{x_i-\mu}{\sigma}\right)\right]-\frac{1}{\xi}\sum_{i=1}^{m}\frac{x_i-\mu}{\sigma+\xi(x_i-\mu)}\right]\end{aligned}$$

Now

$$\begin{aligned}\frac{dl}{d\xi} =&\frac{1}{\xi^2}\sum_{i=1}^{m}\ln\left[1+\xi\left(\frac{x_i-\mu}{\sigma}\right)\right]-\left(1+\frac{1}{\xi}\right)\sum_{i=1}^{m}\frac{x_i-\mu}{\sigma}.\frac{1}{1+\xi\left(\frac{x_i-\mu}{\sigma}\right)}\\ &-\sum_{i=1}^{m}\left[1+\xi\left(\frac{x_i-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\left[\frac{1}{\xi^2}\sum_{i=1}^{m}\ln\left[1+\xi\left(\frac{x_i-\mu}{\sigma}\right)\right]-\frac{1}{\xi}\sum_{i=1}^{m}\frac{x_i-\mu}{\sigma}.\frac{1}{1+\xi\left(\frac{x_i-\mu}{\sigma}\right)}\right]\end{aligned}$$

$$\begin{aligned}\frac{dl}{d\xi} =&\frac{1}{\xi^2}\sum_{i=1}^{m}\ln\left[1+\xi\left(\frac{x_i-\mu}{\sigma}\right)\right]-\left(1+\frac{1}{\xi}\right)\sum_{i=1}^{m}\frac{x_i-\mu}{\sigma+\xi(x_i-\mu)}\\ &-\frac{1}{\xi^2}\sum_{i=1}^{m}\sum_{i=1}^{m}\left[1+\xi\left(\frac{x_i-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\ln\left[1+\xi\left(\frac{x_i-\mu}{\sigma}\right)\right]\\ &+\frac{1}{\sigma}\sum_{i=1}^{m}\sum_{i=1}^{m}\left[1+\xi\left(\frac{x_i-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\left[\frac{x_i-\mu}{\sigma+\xi(x_i-\mu)}\right]\end{aligned}\tag{6.66}$$

Differentiating l with respect to σ

$$\begin{aligned}\frac{dl}{d\sigma} =&-\frac{m}{\sigma}-\left(1+\frac{1}{\xi}\right)\sum_{i=1}^{m}\frac{1}{1+\xi\left(\frac{x_i-\mu}{\sigma}\right)}.(-\xi)\left(\frac{x_i-\mu}{\sigma^2}\right)\\ &-\sum_{i=1}^{m}\left[1+\xi\left(\frac{x_i-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\left[-\frac{1}{\xi}\sum_{i=1}^{m}\frac{1}{1+\xi\left(\frac{x_i-\mu}{\sigma}\right)}.(-\xi)\left(\frac{x_i-\mu}{\sigma^2}\right)\right]\\ \frac{dl}{d\sigma} =&-\frac{m}{\sigma}-\left(1+\frac{1}{\xi}\right)\sum_{i=1}^{m}\frac{\xi(x_i-\mu)}{\sigma[\sigma+\xi(x_i-\mu)]}\\ &-\sum_{i=1}^{m}\left[1+\xi\left(\frac{x_i-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}.\sum_{i=1}^{m}\frac{\xi(x_i-\mu)}{\sigma[\sigma+\xi(x_i-\mu)]}\\ \frac{dl}{d\sigma} =&-\frac{m}{\sigma}-\left(1+\frac{1}{\xi}\right)\sum_{i=1}^{m}\frac{\xi(x_i-\mu)}{\sigma[\sigma+\xi(x_i-\mu)]}\\ &-\sum_{i=1}^{m}\sum_{i=1}^{m}\left[1+\xi\left(\frac{x_i-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\frac{\xi(x_i-\mu)}{\sigma[\sigma+\xi(x_i-\mu)]}\end{aligned}\tag{6.67}$$

Differentiating l with respect to μ

$$\begin{aligned}\frac{dl}{d\mu} &= -\left(1+\frac{1}{\xi}\right)\sum_{i=1}^{m}\frac{1}{1+\xi\left(\frac{x_i-\mu}{\sigma}\right)}\cdot\left(-\frac{\xi}{\sigma}\right) \\ &\quad -\left[1+\xi\left(\frac{x_i-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\left[-\frac{1}{\xi}\frac{1}{1+\xi\left(\frac{x_i-\mu}{\sigma}\right)}\cdot\left(-\frac{\xi}{\sigma}\right)\right] \\ \frac{dl}{d\mu} &= \left(1+\frac{1}{\xi}\right)\sum_{i=1}^{m}\frac{\xi}{\sigma+\xi(x_i-\mu)} - \left[1+\xi\left(\frac{x_i-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\left[\frac{1}{\sigma+\xi(x_i-\mu)}\right] \qquad (6.68)\end{aligned}$$

To obtain maximum likelihood estimates of ξ, σ and μ (6.66), (6.67) and (6.68) are set to zero and standard numerical optimization algorithms are applied.

Proof of maximum likelihood estimates for non-stationary GEVD model parameters

$$f_{\mu_t,\sigma,\xi}(x) = \frac{1}{\sigma}\left[1+\xi\left(\frac{x-\mu_t}{\sigma}\right)\right]^{-1-\frac{1}{\xi}} \exp\left\{-\left[1+\xi\left(\frac{x-\mu_t}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\}$$

the log-likelihood of the GEVD when $\xi \neq 0$ is:

$$l(\mu_t,\sigma,\xi) = -m\ln\sigma - \left(1+\frac{1}{\xi}\right)\sum_{i=1}^{m}\ln\left[1+\xi\left(\frac{x_i-\mu_t}{\sigma}\right)\right] - \sum_{i=1}^{m}\left[1+\xi\left(\frac{x_i-\mu_t}{\sigma}\right)\right]^{-\frac{1}{\xi}}$$

where $\mu_t = \theta_0 + \theta_1 t$. It implies that

$$\begin{aligned} l(\theta_0+\theta_1 t,\sigma,\xi) = &- m\ln\sigma - \left(1+\frac{1}{\xi}\right)\sum_{i=1}^{m}\ln\left[1+\xi\left(\frac{x_i-(\theta_0+\theta_1 t)}{\sigma}\right)\right] \\ &- \sum_{i=1}^{m}\left[1+\xi\left(\frac{x_i-(\theta_0+\theta_1 t)}{\sigma}\right)\right]^{-\frac{1}{\xi}} \end{aligned}$$

set

$$h = -\left(1+\frac{1}{\xi}\right)\sum_{i=1}^{m}\ln\left[1+\xi\left(\frac{x_i-(\theta_0+\theta_1 t)}{\sigma}\right)\right]$$

and

$$g = -\sum_{i=1}^{m}\left[1+\xi\left(\frac{x_i-(\theta_0+\theta_1 t)}{\sigma}\right)\right]^{-\frac{1}{\xi}}$$

so that

$$l(\mu_t,\sigma,\xi) = -m\ln\sigma + h + g$$

so

$$\begin{aligned}\frac{dh}{d\theta_0} &= -\left(1+\frac{1}{\xi}\right)\sum_{i=1}^{m}\frac{1}{1+\xi\left(\frac{x_i-(\theta_0+\theta_1 t)}{\sigma}\right)}\frac{d}{d\theta_0}\left[1+\xi\left(\frac{x_i-(\theta_0+\theta_1 t)}{\sigma}\right)\right] \\ &= -\left(1+\frac{1}{\xi}\right)\sum_{i=1}^{m}\frac{\sigma}{\sigma\left[1+\xi\left(\frac{x_i-(\theta_0+\theta_1 t)}{\sigma}\right)\right]}.\left(-\frac{\xi}{\sigma}\right) \\ &= \left(\frac{\xi+1}{\sigma}\right)\sum_{i=1}^{m}\frac{1}{1+\xi\left(\frac{x_i-(\theta_0+\theta_1 t)}{\sigma}\right)}\end{aligned}$$

for g and taking natural logarithm both sides we obtain

$$\ln g = \frac{1}{\xi}\sum_{i=1}^{m}\ln\left[1+\xi\left(\frac{x_i-(\theta_0+\theta_1 t)}{\sigma}\right)\right]$$

Differentiating implicitly

$$\begin{aligned}\frac{1}{g}\frac{dg}{d\theta_0} &= \frac{1}{\xi}\sum_{i=1}^{m}\frac{1}{1+\xi\left(\frac{x_i-(\theta_0+\theta_1 t)}{\sigma}\right)}.\frac{d}{d\theta_0}\left[1+\xi\left(\frac{x_i-(\theta_0+\theta_1 t)}{\sigma}\right)\right] \\ &= \frac{1}{\xi}\sum_{i=1}^{m}\frac{\sigma}{\sigma+\xi(x_i-\theta_0-\theta_1 t)}.\frac{\xi}{\sigma} \\ \frac{1}{g}\frac{dg}{d\theta_0} &= \frac{1}{\sigma}\sum_{i=1}^{m}\frac{1}{1+\xi\left(\frac{x_i-(\theta_0+\theta_1 t)}{\sigma}\right)}\end{aligned}$$

Therefore

$$\begin{aligned}\frac{dg}{d\theta_0} &= \frac{1}{\sigma}\sum_{i=1}^{m}\frac{1}{1+\xi\left(\frac{x_i-(\theta_0+\theta_1 t)}{\sigma}\right)}.\left[\sum_{i=1}^{m}\left[1+\xi\left(\frac{x_i-(\theta_0+\theta_1 t)}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right] \\ &= -\frac{1}{\sigma}\sum_{i=1}^{m}\left[1+\xi\left(\frac{x_i-\theta_0-\theta_1 t}{\sigma}\right)\right]^{-1-\frac{1}{\xi}}\end{aligned}$$

Thus

$$\frac{dl}{d\theta_0} = \frac{\xi+1}{\sigma}\sum_{i=1}^{m}\left[1+\xi\left(\frac{x_i-\theta_0-\theta_1 t}{\sigma}\right)\right]^{-1} - \frac{1}{\sigma}\sum_{i=1}^{m}\left[1+\xi\left(\frac{x_i-\theta_0-\theta_1 t}{\sigma}\right)\right]^{-1-\frac{1}{\xi}} \tag{6.69}$$

and for $\frac{dl}{d\theta_1}$:

$$\begin{aligned}
\frac{dh}{d\theta_1} &= -\left(1+\frac{1}{\xi}\right)\sum_{i=1}^{m}\frac{1}{1+\xi\left(\frac{x_i-(\theta_0+\theta_1 t)}{\sigma}\right)}\frac{d}{d\theta_1}\left[1+\xi\left(\frac{x_i-(\theta_0+\theta_1 t)}{\sigma}\right)\right] \\
\frac{dh}{d\theta_1} &= -\left(1+\frac{1}{\xi}\right)\sum_{i=1}^{m}\frac{\sigma}{\sigma+\xi(x_i-\theta_0-\theta_1 t)}.\left(-\frac{\xi t}{\sigma}\right) \\
&= \left(\frac{\xi+1}{\sigma}\right)\sum_{i=1}^{m} t\left[1+\xi\left(\frac{x_i-(\theta_0+\theta_1 t)}{\sigma}\right)\right]^{-1}
\end{aligned}$$

and

$$\begin{aligned}
\frac{dg}{d\theta_1} &= \frac{1}{\sigma}\sum_{i=1}^{m} t\left[1+\frac{\xi(x_i-\theta_0-\theta_1 t)}{\sigma}\right]^{-1}\left[-\left[1+\frac{\xi(x_i-\theta_0-\theta_1 t)}{\sigma}\right]^{-\frac{1}{\xi}}\right] \\
\frac{dg}{d\theta_1} &= -\frac{1}{\sigma}\sum_{i=1}^{m} t\left[1+\xi\left(\frac{x_i-\theta_0-\theta_1 t}{\sigma}\right)\right]^{-1-\frac{1}{\xi}}
\end{aligned}$$

hence

$$\frac{dl}{d\theta_1} = \frac{\xi+1}{\sigma}\sum_{i=1}^{m} t\left[1+\xi\left(\frac{x_i-\theta_0-\theta_1 t}{\sigma}\right)\right]^{-1} - \frac{1}{\sigma}\sum_{i=1}^{m} t\left[1+\xi\left(\frac{x_i-\theta_0-\theta_1 t}{\sigma}\right)\right]^{-1-\frac{1}{\xi}} \tag{6.70}$$

$$\begin{aligned}
\frac{dl}{d\sigma} =& -\frac{m}{\sigma} - \left(1+\frac{1}{\xi}\right)\sum_{i=1}^{m}\frac{\xi(x_i-\mu_t)}{\sigma[\sigma+\xi(x_i-\mu_t)]} \\
& - \sum_{i=1}^{m}\sum_{i=1}^{m}\left[1+\xi\left(\frac{x_i-\mu_t}{\sigma}\right)\right]^{-\frac{1}{\xi}}\frac{\xi(x_i-\mu_t)}{\sigma[\sigma+\xi(x_i-\mu_t)]}
\end{aligned} \tag{6.71}$$

$$\begin{aligned}
\frac{dl}{d\xi} =& \frac{1}{\xi^2}\sum_{i=1}^{m}\ln\left[1+\xi\left(\frac{x_i-\mu_t}{\sigma}\right)\right] - \left(1+\frac{1}{\xi}\right)\sum_{i=1}^{m}\frac{x_i-\mu_t}{\sigma+\xi(x_i-\mu_t)} \\
& - \frac{1}{\xi^2}\sum_{i=1}^{m}\sum_{i=1}^{m}\left[1+\xi\left(\frac{x_i-\mu_t}{\sigma}\right)\right]^{-\frac{1}{\xi}}\ln\left[1+\xi\left(\frac{x_i-\mu_t}{\sigma}\right)\right] \\
& + \frac{1}{\sigma}\sum_{i=1}^{m}\sum_{i=1}^{m}\left[1+\xi\left(\frac{x_i-\mu_t}{\sigma}\right)\right]^{-\frac{1}{\xi}}\left[\frac{x_i-\mu_t}{\sigma+\xi(x_i-\mu_t)}\right]
\end{aligned} \tag{6.72}$$

To obtain maximum likelihood estimates of θ_0, θ_1, σ and ξ (6.69), (6.70), (6.71) and (6.72) are set to zero and standard numerical optimization algorithms are applied.

Diagnostic plots for non-stationary GEVD models of mean annual rainfall for Zimbabwe

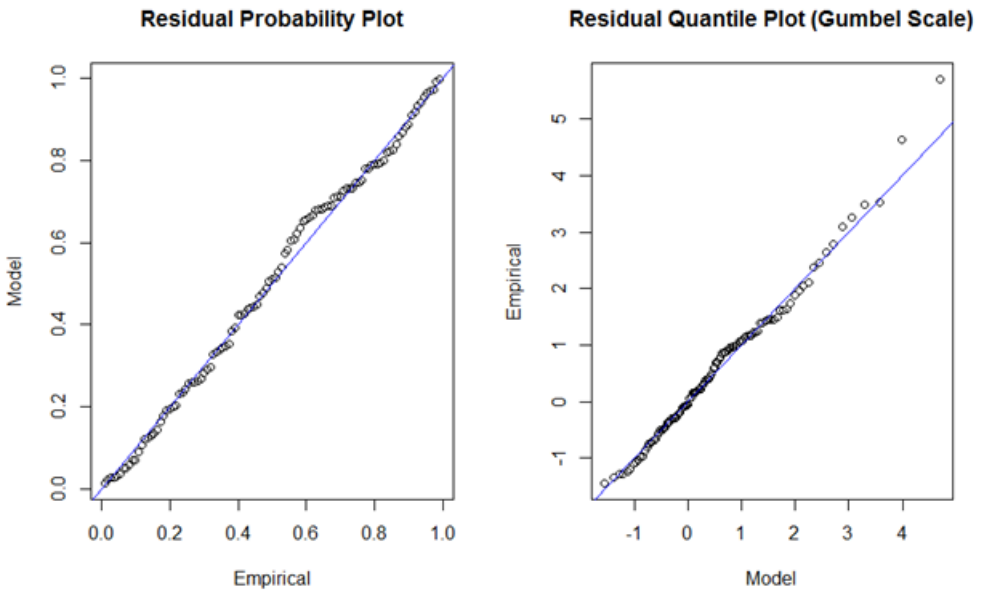

Figure 6.7: Diagnostic plot for GEVD Model 2

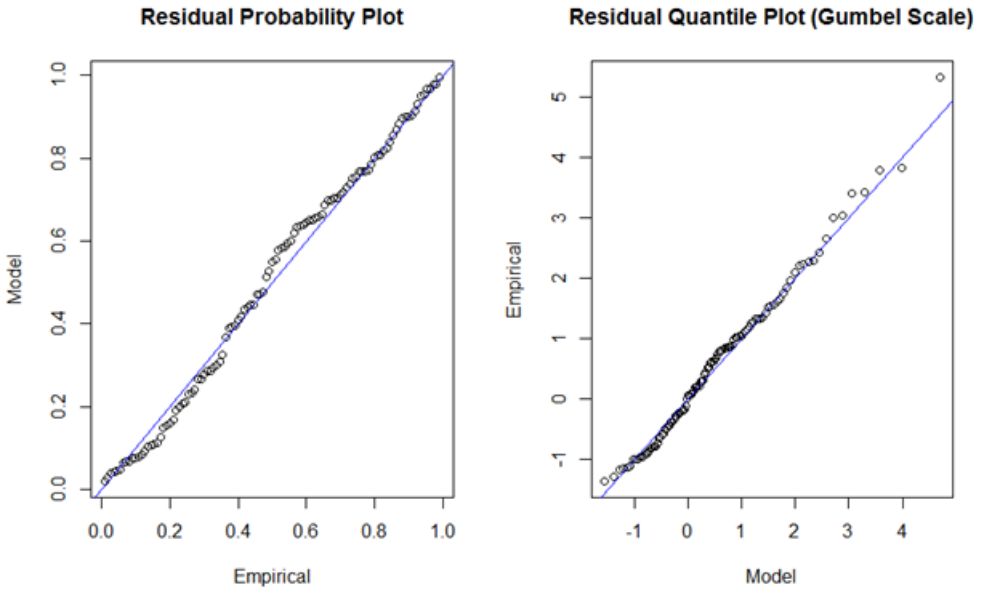

Figure 6.8: Diagnostic plot for GEVD Model 3

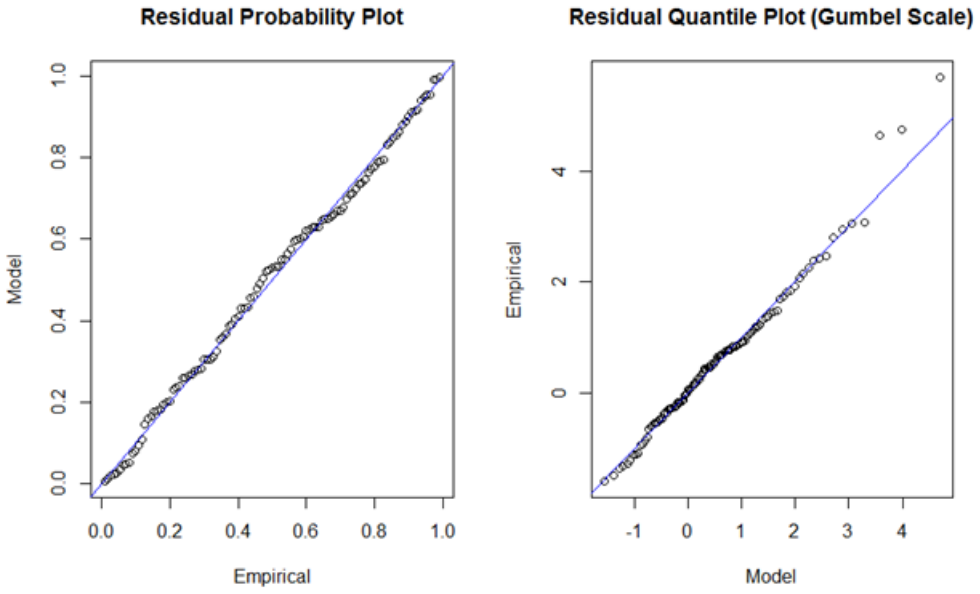

Figure 6.9: Diagnostic plot for GEVD Model 4

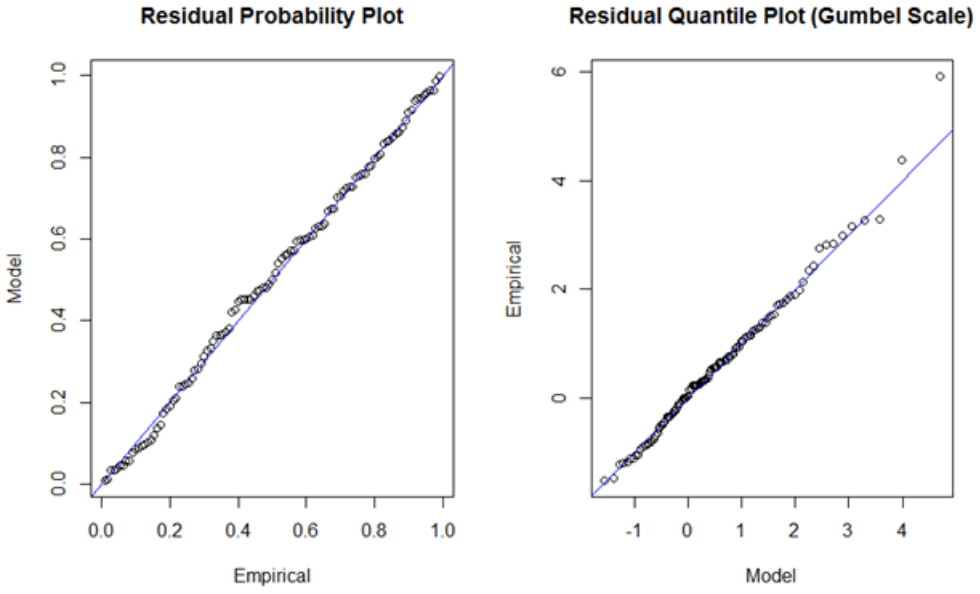

Figure 6.10: Diagnostic plot for GEVD Model 5

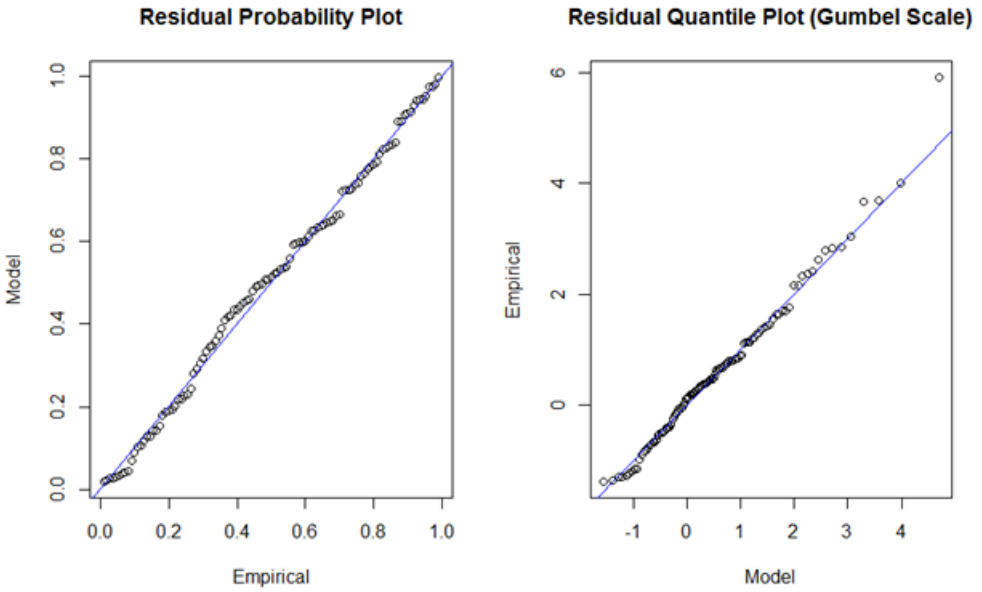

Figure 6.11: Diagnostic plot for GEVD Model 6

Chapter 7

Modelling of extreme minimum rainfall using generalised extreme value distribution for Zimbabwe

7.1 Introduction

Extreme low rainfall attributed to global warming, although rare, is a natural phenomenon that affects people's socio-economic activities worldwide. Extreme droughts occur from time to time in Zimbabwe and impact negatively on the country's economic performance. The drought of rainfall season year 1991/1992 was one of the worst in the recorded history of Zimbabwe. Its impact was felt even in the insurance industry which received high claims for crop failure (Makarau and Jury, 1997). Droughts can be viewed as extreme events outside of the normal rainfall occurrences, such as exceptionally lower amounts of mean annual rainfall (Panu and Sharma, 2002). In Zimbabwe, at least 50% of the gross domestic product is derived from rain-fed agriculture (Jury, 1996). With more low technology indigenous farmers entering commercial agriculture through the accelerated land-reform programme, modelling and prediction of extremely low annual rainfall and the associated probabilities of drought become more relevant.

Developing methods that can give a suitable prediction of meteorological events is always interesting for both meteorologists and statisticians. The use of standard statistical techniques in modelling, forecasting and prediction of extremes in average rainfall and rare events is less prudent because of gross under-estimation (Hasan et al., 2012). Extreme value theory is an alternative and superior approach to quantify the stochastic behaviour of a process at unusually large or small levels (Hasan et al., 2012). Extreme value theory provides the statistical framework to make inferences about the probability of very rare and extreme events. It is based on the analysis of the maximum (or minimum) value in a selected time period.

Recently there has been growing interest in modelling extreme events, especially in situations in which scientists underestimated the probabilities of extreme events that subsequently occurred and caused catastrophic damage (Coles et al., 2001). Research conducted provides evidence of the importance of modelling rainfall from different regions of the world: Nadarajah and Choi (2007) used extreme value theory for rainfall data from South Korea while Koutsoyiannis (2004) applied extreme value theory to rainfall data from Europe and the USA. Koutsoyiannis and Baloutsos (2000) applied extreme value theory to Greece's rainfall data; and Crisci et al. (2002) applied extreme value distributions to rainfall data from Italy. The use of extreme value distributions is not restricted to meteorological events and research occurs in such fields such as energy (Chikobvu and Sigauke, 2013); insurance (Smith and Goodman, 2000); fish management (Hilborn and Mangel, 1997) and ecology (Ludwig, 1996). There is no work known to us on rainfall extremes in Zimbabwe. In this research, we provide the first application of extreme value distributions to model minimum annual rainfall in Zimbabwe.

Rainfall in Zimbabwe is associated with the behaviour of the inter-tropical convergence zones whose oscillations are influenced by changing pressure patterns to the

north and south of the country (Makarau and Jury, 1997). Zimbabwe lies in the Southwest Indian Ocean zone, which is often affected by tropical cyclones. Tropical cyclones are low pressure systems that have well-defined clockwise (in the southern hemisphere) wind circulations which spiral toward the centre where the winds are strongest and rains are heaviest. Cyclones that develop over the western side of the Indian Ocean occasionally affect the rainy season. The amount and intensity of rainfall during a given wet spell is enhanced by the passage of upper westerly wind waves of mid-latitude origin (Buckle, 1996; Smith, 1985a).

Studies of extreme low rainfall are beneficial to decision-makers in government and non-governmental organisations involved in early warning systems and food security, poverty alleviation and disaster management and risk management. This study will also inform climatologists about the behaviour of extreme low rainfall. Appropriate decisions and plans can be made based on the results of this study to prepare the general public for changes brought on by extremely low rainfall. The objective of this chapter was to quantify and describe the behaviour of extreme minimum rainfall in Zimbabwe. The aim was to model (i) extreme minimum rainfall using GEVD by using the maximum likelihood estimation method and the Bayesian statistics approach, (ii) calculate the mean return period of low rainfall i.e. the number of years on average before another drought of equal or greater intensity occurs, (iii) investigate the influence of time and climatic change drivers, namely; SOI and SDSLP values on extreme minimum annual rainfall.

The rest of the chapter is organised as follows: Section 7.2 presents the research methodology. Section 7.3 presents the empirical results and discussion of the findings. Section 7.4 provides the concluding remarks. Finally, Section 7.6 (Appendix) presents some diagnostic plots of the fitted models.

7.2 Research methodology

In this section we discuss the background theory on the GEVD model. The data used in this chapter is described in Chapter 3. The parameters of the fitted models are estimated by the method of maximum likelihood and method of Bayesian statistics.

7.2.1 Normal distribution

A normal distribution is symmetrical and has a bell-shaped density curve with a single peak. The normal density function, which gives the height of the density at any value x is given by:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(x_i - \mu)^2\right\}, \text{ with } \sigma > 0, \tag{7.1}$$

where μ is the mean (where the peak of the density occurs) and σ is the standard deviation (which indicates the spread or girth of the bell curve).

7.2.2 Generalised extreme value distribution (GEVD)

In climatology, meteorology and hydrology, maxima of temperatures, precipitation and river discharges have been recorded for many decades (Hosking and Wallis, 1987). The extreme value theorem provides a theoretical framework to model the distribution of extreme events and the three-parameter GEVD was recommended for meteorology frequency analysis (Bunya et al., 2007). The three parameters are: location, scale and shape. The GEVD is a family of continuous probability distributions developed within extreme value theorem. The GEVD unites the Gumbel, Fréchet and Weibull family of distributions into a single family to allow for a continuous range of possible shapes. Based on the extreme value theorem, the GEVD is the limiting distribution of properly normalised maxima of a sequence of independent and identically distributed random variables (Beirlant et al., 2004). Thus, the GEVD is used to model the maxima of a long (finite) sequence of random variables. The

unified GEVD for modelling maxima is given by:

$$F_{\xi,\mu,\sigma}(x) = \exp\left\{-\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\}, \quad \text{with } \xi \neq 0, \tag{7.2}$$

with $\mu \in \Re$, $\sigma > 0$ and $1 + \xi\left(\frac{x-\mu}{\sigma}\right) > 0$.

Where μ, σ and ξ are the location, scale and shape parameters, respectively. The probability density function is sometimes called the Fisher-Tippett distribution and is obtained as the derivative of the distribution function:

$$f_{\xi,\mu,\sigma}(x) = \frac{1}{\sigma}\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-1-\frac{1}{\xi}} \exp\left\{-\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\}, \xi \neq 0 \tag{7.3}$$

The shape parameter ξ is also known as the extreme value index. The parameter ξ^{-1} is the rate of tail decay of the GEVD. If $\xi > 0$, F belongs to the heavy-tailed Fréchet class of distributions such as Pareto, Cauchy, student-t and mixture distributions. If $\xi < 0$, F belongs to the short-tailed with finite lower bounds Weibull class of distributions which includes distributions such as uniform and beta distributions. If $\xi = 0$, F then belongs to the light-tailed Gumbell class of distributions which includes distributions such as normal, exponential, gamma and log-normal distributions (Bali, 2003).

Modelling minima random variables

The classical GEVD for extremes is based on asymptotic approximations to the sampling behaviour of block maxima (discussed in Chapter 6). The block maxima size (hourly, daily, weekly, monthly or yearly) varies according to instrument constraints, seasonality and the application at hand. The only possible limiting form of a normalised maximum of a random sample (when a non-degenerate limit exists) is captured by the GEVD. The data set is partitioned into blocks of equal length and distribution and GEVD is fitted to the set of block maxima. In this study, minima rainfall was modelled using GEVD. In order to model minima random variables we use

the duality between the distributions for maxima and minima (see Chapter 6 for details). If $M_N = \min\{X_1, X_2, ..., X_N\}$ where $X_1, X_2, ..., X_N$ is a sequence of independent random variables having a common distribution function and $Y_i = -X_i$ for $i = 1, 2, ..., N$, the change of sign means that small values of X_i correspond to large values of Y_i . So if $M_N = \min\{X_1, X_2, ..., X_N\}$ and $\widetilde{M}_N = \max\{Y_1, Y_2, ..., Y_N\}$, then $M_N = -\widetilde{M}_N$. The minima becomes:

$$\widetilde{M}_n = \max\{-X_1, ..., -X_n\}$$

Extreme maxima theory and methods discussed in Section 6.2 are then used to model extreme minima $-X_1, ..., -X_n$ where X_i represents mean annual rainfall in period i (Coles et al., 2001). Extreme maxima theory and methods are then used to model extreme minima (Coles et al., 2001; Hosking, 1984). Based on the extreme value theorem that derives the GEVD, we can fit a sample of extremes to the GEVD to obtain the parameters that best explain the probability distribution of the extremes.

Parameter estimation

There is a wide variety of methods to estimate the GEVD parameters in the independent and identically distributed settings (Diebolt et al., 2007). The three parameters are estimated using methods of moments, maximum likelihood method and probability weighted moments or equivalent L-moments (Bunya et al., 2007). Hosking (1985) showed that the probability weighted moments quantile estimators for the GEVD are better than the maximum likelihood method for small samples ($n < 50$). Madsen et al. (1997), also showed that method of moments quantile estimators perform well when the sample size is modest. In this study, the maximum likelihood method was exploited because $n > 50$.

Maximum likelihood method

Under the assumption that $X_1, X_2, ..., X_m$ are independent random samples having a GEVD, the log-likelihood for the GEVD parameters when $\xi \neq 0$ is:

$$l(\mu, \sigma, \xi) = -m\ln(\sigma) - \left(1 + \frac{1}{\xi}\right) \sum_{i=1}^{m} \ln\left[1 + \xi\left(\frac{x_i - \mu}{\sigma}\right)\right] - \sum_{i=1}^{m}\left[1 + \xi\left(\frac{x_i - \mu}{\sigma}\right)\right]^{-\frac{1}{\xi}} \tag{7.4}$$

provided that $1 + \xi\left(\frac{x_i - \mu}{\sigma}\right) > 0$ for $i = 1, ..., m$ (Coles et al., 2001). We differentiated the log-likelihood of GEVD to find a set of equations which we solved using numerical optimization algorithms. For computational details, we refer to Chapter 6 and previous studies (Macleod, 1989; Hosking, 1985; Prescott and Walden, 1980). The support of F depends on the unknown parameter values therefore, the usual regularity conditions underlying the asymptotic properties of maximum likelihood estimators are not satisfied. This problem is studied in depth by Smith (1985b). In the case $\xi > -0.5$, the usual properties of consistency, asymptotic efficiency and asymptotic normality hold.

Test for stationarity

The augmented Dickey Fuller (ADF) stationarity test is performed on the data to test for stationarity. The null hypothesis of the ADF test is that there is no trend while the alternative hypothesis is that there is a trend in the data.

Goodness of fit

To access the quality of convergence of the GEVD, the Kolmogorov-Smirnov (KS) and Anderson-Darling (AD) goodness-of-fit tests were used. The KS test, based on the empirical cumulative distribution function, is used to decide if a sample comes from the hypothesized continuous distribution. The K-S test is less sensitive at the tails than at the centre of the distribution. The AD test, which is an improvement of the KS test, compares the fit of an observed cumulative distribution function to

an expected cumulative distribution function. The AD test gives more weight to the tails of a distribution than does the KS test (Stephens, 1974).

Return period or level estimates

We can estimate how often the extreme quantiles occur with a certain return level. The return level is defined as a level that is expected to be equalled or exceeded on average once every interval of time (T) with a probability of p . For the normal distribution we set:

$$F_{\mu,\sigma}x = \int_{-\infty}^{x_p} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(x_p-\mu)^2\right\} dx = 1-p \tag{7.5}$$

where x_p the return level. By setting the return period and solving the equation (Section 7.5) the return level, x_p , can be calculated:

$$x_p = \sigma\phi^{-1}(1-p) + \mu \tag{7.6}$$

which can be re-written as:

$$x_p = \sigma Z_{1-p} + \mu \tag{7.7}$$

Similarly, for the GEVD we set:

$$F_{\xi,\mu,\sigma}(x) = \exp\left\{-\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\} = 1-p, \tag{7.8}$$

The return level (see Chapter 6) is given by:

$$\mu + \frac{\sigma}{\xi}\left[\left[-\ln\left(1-\frac{1}{T}\right)\right]^{-\xi} - 1\right] \tag{7.9}$$

Return levels are important for prediction purposes and can be estimated from stationary models. The mean return period is the number of years we expect to wait on average before we observe another drought of equal or greater intensity. If the

exceedance probability of observing a drought of a given severity in any given year is p then the mean return period T is such that $T = \frac{1}{p}$.

7.2.3 Bayesian analysis of extreme values for GEVD

Inference on the extremes of environmental processes is important to meteorologists, civil engineers, agriculturalists and statisticians. Naturally, data at extreme levels are scarce. Bayesian inference allows any additional information about the processes to be incorporated as prior information. The basic theory of Bayesian analysis of extreme values is well documented (see Gamerman, 1997, Coles et al., 2001 and Coles and Tawn, 2005 for more information). The Markov Chain Monte Carlo techniques are applied in this research to give Bayesian analysis of the annual minima rainfall data for Zimbabwe. In this chapter, the prior is constructed by assuming there is no information available about the process (rainfall) apart from the data. The annual rainfall data have a GEVD, i.e. $X_i \sim GEVD(\mu, \sigma, \xi)$, and the parameters μ, σ and ξ are treated as random variables for which we specify prior distributions. For specification of the prior, the parameterisation $\phi = \log(\sigma)$ is easier to work with because it is constrained to be positive. The specification of priors enables us to supplement the information provided by the data. The prior density is:

$$\pi(\mu, \phi, \xi) = \pi_\mu(\mu)\pi_\phi(\phi)\pi_\xi(\xi) \tag{7.10}$$

where each marginal prior is normally distributed with large variances. The variances are chosen to be large enough to make the distributions almost flat, corresponding to prior ignorance. The joint posterior density is the product of the prior and the likelihood and is given as:

$$\pi(\sigma, \mu, \xi \mid \mathbf{x}) \propto \pi_\phi(\phi)\pi_\mu(\mu)\pi_\xi(\xi)L(\mu, \phi, \xi \mid \mathbf{x}) \tag{7.11}$$

where

$$L(\sigma,\mu,\xi \mid \mathbf{x}) = \frac{1}{\sigma^m}\exp\left\{-\sum_{i=1}^{m}\left[1+\xi\left(\frac{x_i-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\}\prod_{i=1}^{m}\left[1+\xi\left(\frac{x_i-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}-1} \tag{7.12}$$

is the likelihood with σ replaced by e^{ϕ}. The Gibbs sampler is used to simulate from each of the full conditionals. The posterior is of the form:

$$\pi(\sigma,\mu,\xi \mid \mathbf{x})L(\mu,\phi,\xi \mid \mathbf{x}) \propto \pi_{\mu}(\mu)\pi_{\phi}(\phi)\pi_{\xi}(\xi)L(\mu,\phi,\xi \mid \mathbf{x}) \tag{7.13}$$

So the full conditionals are of the form:

$$\pi(\phi \mid \mu,\xi) = \pi_{\phi}(\phi)L(\mu,\phi,\xi \mid \mathbf{x})$$

$$\pi(\mu \mid \phi,\xi) = \pi_{\mu}(\mu)L(\mu,\phi,\xi \mid \mathbf{x})$$

$$\pi(\xi \mid \mu,\sigma) = \pi_{\xi}(\xi)L(\mu,\phi,\xi \mid \mathbf{x})$$

For details of the Markov Chain Monte Carlo algorithm refer to R package (evdbayes version1.1-1).

7.3 Empirical results

In this section, the results of modelling minimum annual rainfall using the maximum likelihood method and Bayesian framework are discussed. We also briefly discuss the description of the rainfall data (Note: a full description of the data set is given in Chapter 3).

Description of data

The analyses was based on the historical mean annual rainfall data recorded from all 40 weather stations in Zimbabwe, dating from as far back as year 1901 to year 2009. A mean annual rainfall figure for the country was calculated. The mean data

were obtained from the Zimbabwe Department of Meteorological Services. For the description of the mean annual data set see Chapter 3. In fitting a 109-year data set to a GEVD, a block size had to be chosen so that individual block minima had a common distribution; yearly blocks were therefore used in this study. The duality principle between the distribution of minima and maxima to fit the distribution of minimal rainfall for Zimbabwe was employed. Maximum likelihood estimates of parameters were estimated. Figure 7.1 shows the graph of $-x_i$ annual rainfall for Zimbabwe.

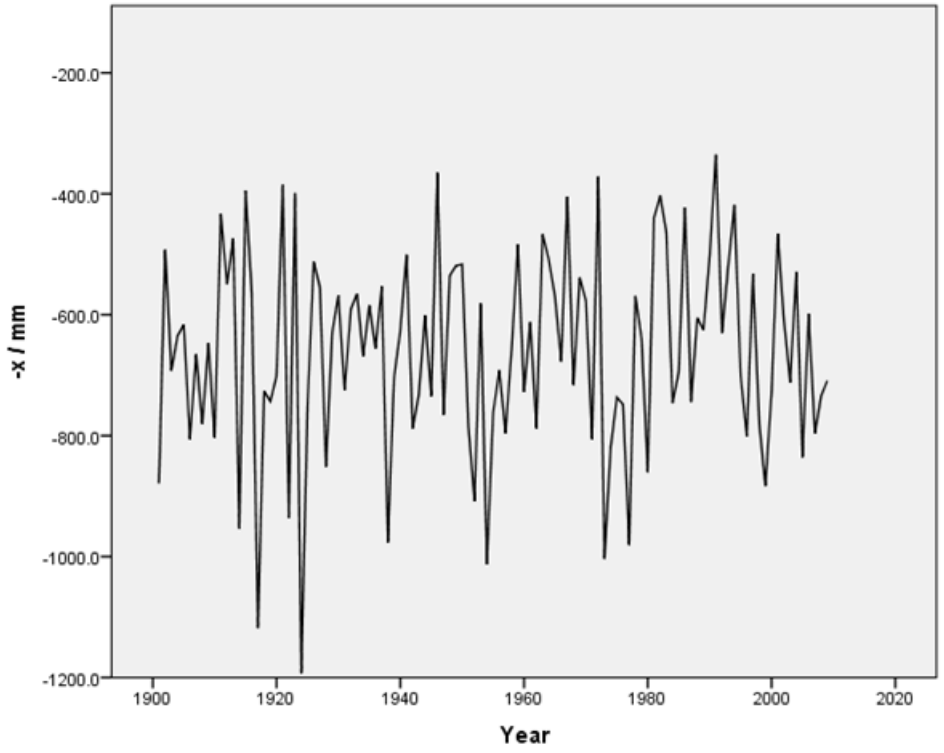

Figure 7.1: Time series plot of the $-x_i$ annual rainfall for Zimbabwe for the period 1901 to 2009.

Unit root test for stationarity

The augmented Dickey Fuller (ADF) test was used to check for stationarity of $-x_i$ annual rainfall for Zimbabwe. Table 7.1 shows the results of the ADF test.

Table 7.1: Unit root test to determine stationarity of minimum annual rainfall data for Zimbabwe for the period 1901 to 2009

Test	Test's critical values			Test statistic
	1%	5%	10%	
Augmented Dickey Fuller	-4.05	-3.45	-3.15	-4.26

The ADF test statistic $= -4.26$ is significant at 1%, 5% and 10% level of significance; we therefore reject the null hypothesis of no stationary at 1%, 5% and 10% levels of significance and conclude that the rainfall data are stationary. The ADF test also indicates that the data does not follow any trend. Therefore, we can determine the return levels of minima annual rainfall.

Descriptive statistics

Table 7.2 shows the descriptive statistics - specifically the coefficient of skewness and Jarque-Bera normality test - of the 109 years of annual rainfall data. The coefficient of skewness of minima annual rainfall ($-x_i$) is negative. This observation suggests that the rainfall data fits a distribution which is relatively long left tailed.

Table 7.2: Summary statistics of minimum annual rainfall data for Zimbabwe for the period 1901 to 2009

No. of obs.	Mean	Std. dev.	Min	Max	Skewness	Jarque-Bera statistic
109	-659.9300	169.2500	-1192.6000	-335.3000	-0.4501	$3.850(0.15)$

Fitting distributions to minimum mean annual rainfall

Normal distribution

Figure 7.2 shows the diagnostic plots illustrating the fit of minima mean annual rainfall data for the period 1901 to 2009 to the normal distribution.

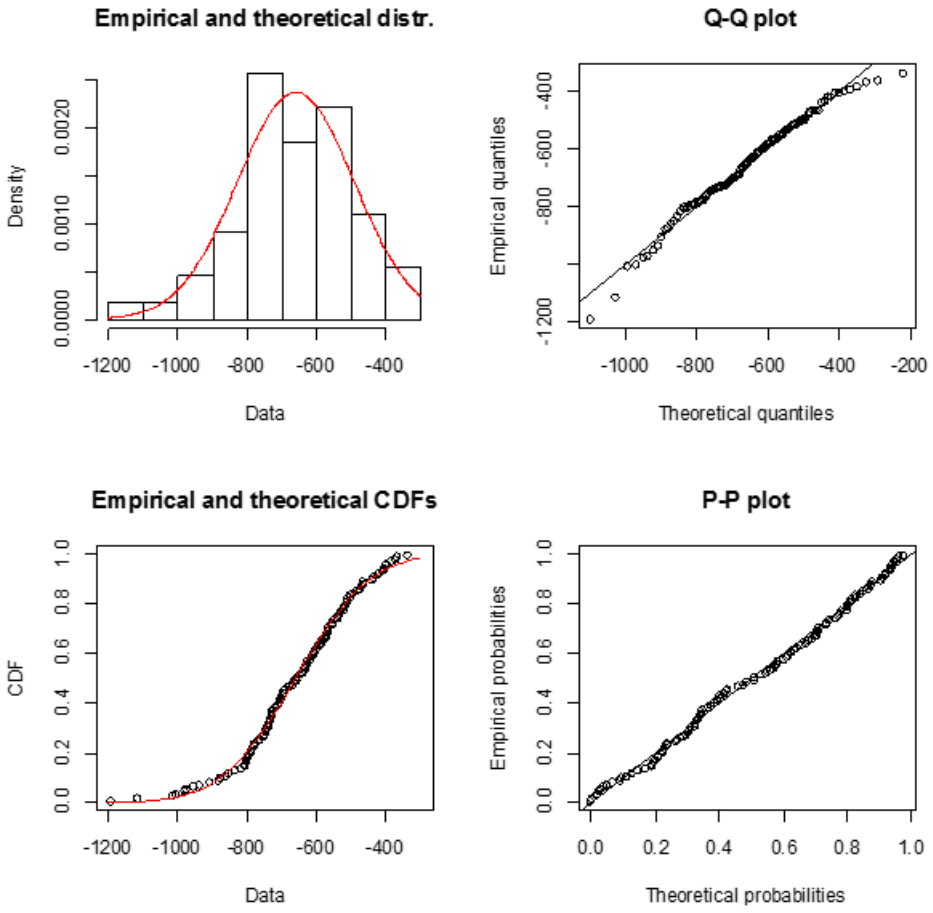

Figure 7.2: Diagnostic plots illustrating the fit of the minimum mean annual rainfall data for Zimbabwe to the normal distribution model, (a) Probability plot (top left panel), (b) Quantile plot (top right panel), (c) Return level plot (bottom left panel) and (d) Density plot (Bottom right panel)

The parameter estimates and their corresponding standard errors in brackets are:

$$\hat{\mu} = -659.9312(16.1362)$$

$$\hat{\sigma} = 168.4675(11.4101)$$

Figure 7.2 shows the minima mean annual rainfall data normal QQ-plot of minima annual rainfall for Zimbabwe. The normal QQ-plot of minima annual rainfall shows deviation from a normal distribution at both lower and upper tails of the

data. However, based on the p-value of the Jarque-Bera test, we fail to reject the null hypothesis of normality. The question is: If annual rainfall is normally distributed, then how do we account for extremely low rainfall (severe droughts) or extremely high rainfall (severe floods) events that have been recorded? The normal distribution approximates these events as negligible or close to zero. If the distribution of minima annual rainfall is heavy-tailed or skewed, the normal distribution may be misleading. Thus, the normal distribution is not a good fit for this rainfall data. The further one gets into the tails of the distribution, the rarer the event, but the event will be catastrophic if it happens. It is important to fit a distribution that is able to capture the probability of extreme minimum annual rainfall.

Generalised extreme value distribution

Figure 7.3 shows the diagnostic plots for the goodness of fit of the minima annual rainfall for Zimbabwe for the period 1901 to 2009.

Table 7.3 shows the maximum likelihood estimates of the GEVD model with their corresponding standard errors in brackets and negative log-likelihood (NLL) value.

Table 7.3: Maximum likelihood estimates (standard errors) and negative log-likelihood value of the GEVD parameters

$\hat{\xi}$	$\hat{\sigma}$	$\hat{\mu}$	NLL value
-0.4410	179.2709	-707.6685	710.3372
(0.0587)	(13.6584)	(18.7453)	

These results show that the data can be modelled using a Weibull class of distribution because $\hat{\xi} < 0$ (bounded tail). Combining estimates and standard errors, the 95% confidence intervals for ξ, σ and μ are $[-0.5561; -0.3259]$,$[154.8222; 203.7196]$ and $[-744.4090; -670.9278]$, respectively. ξ is significantly different from zero because zero is not contained in the interval.

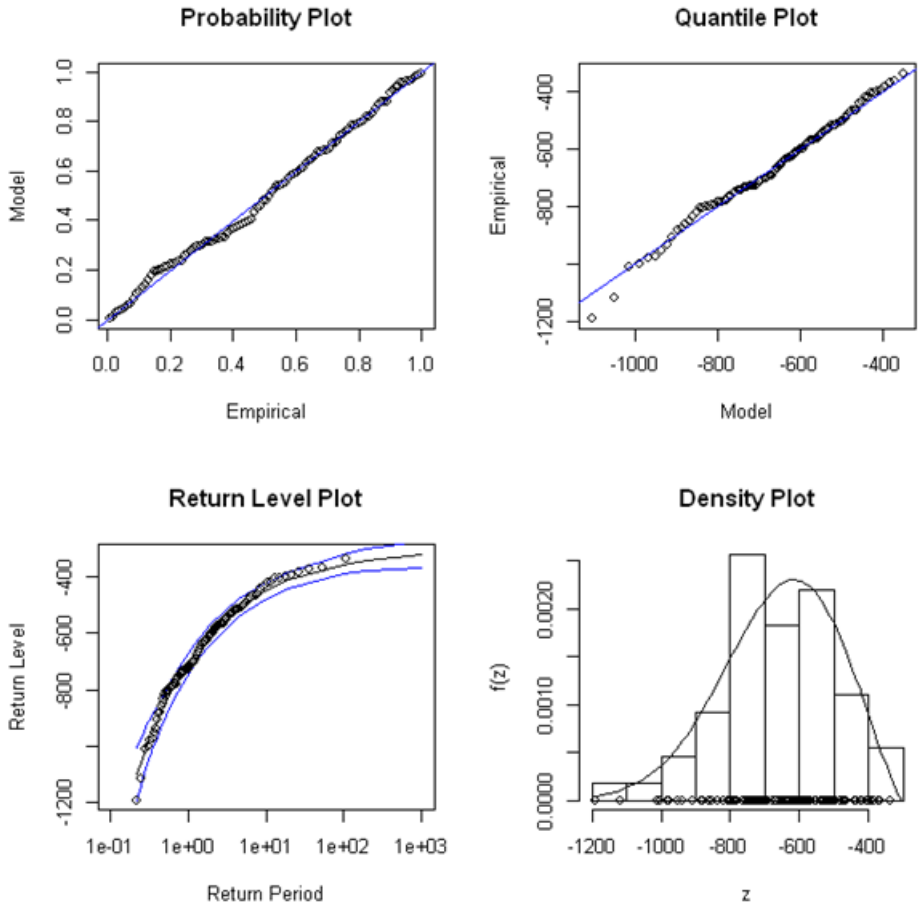

Figure 7.3: Diagnostic plots illustrating the fit of the minimum mean annual rainfall data for Zimbabwe for the period 1901-2009 to the GEVD model, (a) Probability plot (top left panel), (b) Quantile plot (top right panel), (c) Return level plot (bottom left panel) and (d) Density plot (Bottom right panel)

Model diagnostic

It is important to confirm that the data adequately fits the GEVD. Figure 7.3 shows the diagnostic plots of the GEVD. The quantiles of minima rainfall regressed against the quantiles of GEVD shows a straight line. This finding suggests that the data does not deviate from the assumption that they follow a GEVD. Table 7.4 shows the KS and AD statistics for checking adequacy of GEVD model.

The AD test statistic is less than its 5% critical value and the KS test statistic test leads

Table 7.4: KS and AD tests to determine whether minimum annual rainfall data for Zimbabwe for the period 1901-2009 follow a GEVD

Kolmogorov-Smirnov test		Anderson-Darling test	
statistic	critical value	statistic	critical value
0.058	0.130	0.241	2.50

to a decision of non-rejection of the null hypothesis. We conclude that the minimum annual rainfall for Zimbabwe follows the specified GEVD.

The maximum likelihood estimate for ξ is negative, corresponding to a bounded distribution, in which the 95% confidence interval does not contain zero. Greater accuracy of the confidence interval is achieved by the use of the profile likelihood. Figure 7.4 shows the profile likelihood of the generalised extreme value parameter ξ, from which a 95% confidence interval for ξ was obtained as approximately $[-0.55; -0.45]$, which is almost the same as the calculated 95% confidence interval.

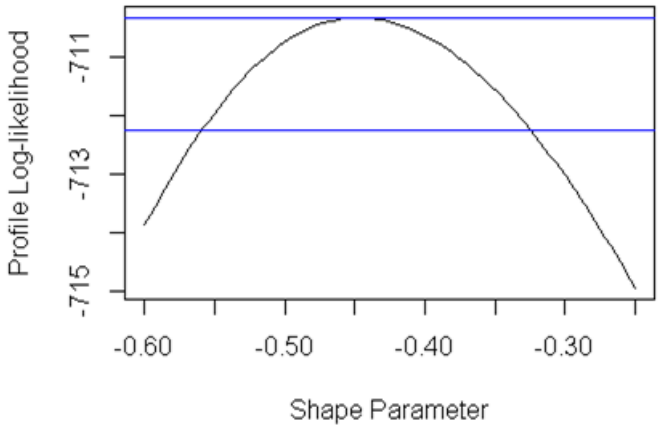

Figure 7.4: Profile likelihood for the GEVD parameter shape, for minimum annual rainfall for Zimbabwe for the period 1901-2009.

Bayesian analysis of minima annual rainfall data

The Markov Chain Monte Carlo method was applied to the annual minimum rainfall data. The GEVD scale parameter was re-parameterized as $\phi = \log(\sigma)$ to retain the positivity of this parameter. The prior density was chosen to be

$$\pi(\mu, \phi, \xi) = \pi_\mu(\mu)\pi_\phi(\phi)\pi_\xi(\xi),$$

where the marginal priors, $\pi_\mu(.)\pi_\phi(.)\pi_\xi(.)$ are

$$\mu \sim N(0, 400000),$$

$$\phi \sim N(0, 400000),$$

$$\xi \sim N(0, 10000),$$

for the three parameters of the GEVD, where, for example, denotes a Gaussian distribution with mean 0 and variance 400 000. These are independent normal priors with large variances. The variances were chosen to be large enough to make the distributions almost flat, corresponding to prior ignorance. In this chapter, 30 000 iterations of the algorithm were carried out. Figure 7.5 shows the MCMC trace plots.

To check that the chains had converged to the correct place, different starting points were used. All the chains converged. The estimated posterior densities for the GEVD parameters for Zimbabwe are given in Figure 7.6.

The posterior means and standard deviations for the GEVD parameters are given in Table 7.5.
Using non-informative priors, which are almost flat and add very little information to the likelihood, the posterior means were close to the maximum likelihood estimates of the GEVD parameters given in Table 7.3. The frequentist properties are preserved by using non-informative priors in the Bayesian statistics approach.

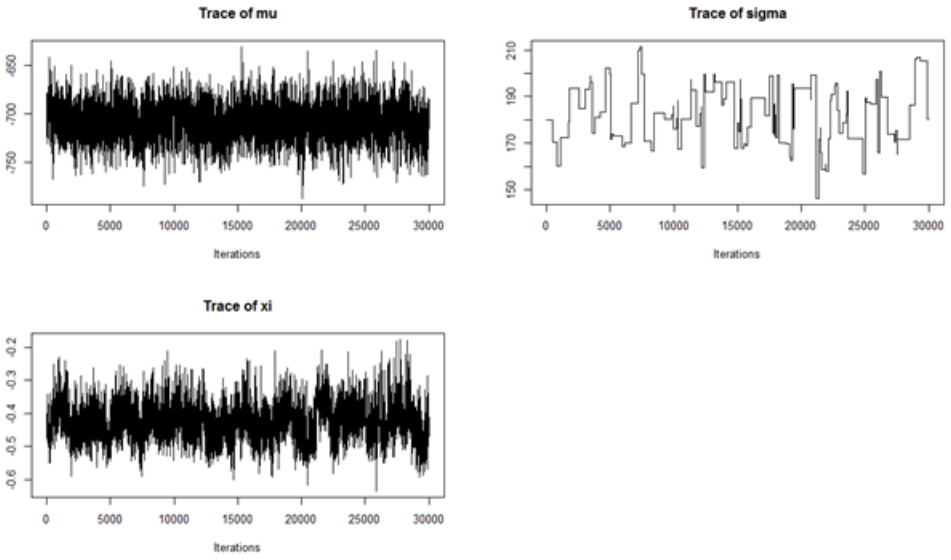

Figure 7.5: Trace plots of the GEVD parameters using non-informative priors for minimum annual rainfall for Zimbabwe for the period 1901-2009.

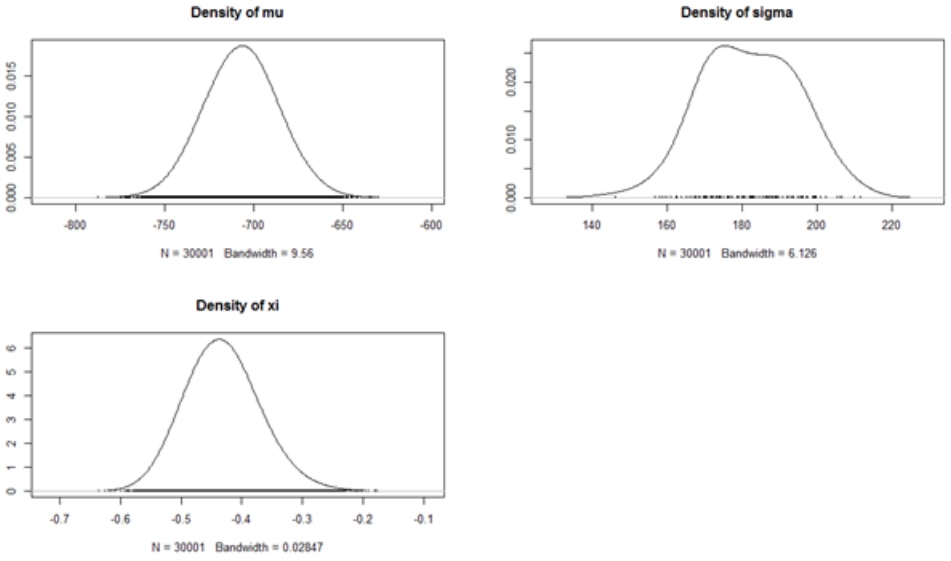

Figure 7.6: Trace plots of the GEVD parameters using non-informative priors for minimum annual rainfall for Zimbabwe for the period 1901-2009.

Return level estimate

The return levels or periods are estimated using the GEVD Model 1 based on the maximum likelihood (ML) estimates and Bayseian MCMC estimates. Rainfall less

Table 7.5: Posterior means (standard errors) of the GEVD Model 1 parameters

$\hat{\xi}$	$\hat{\sigma}$	$\hat{\mu}$
-0.4318	182.2600	-705.6514
(0.1501)	(38.0311)	(55.6321)

than 473 mm per annum is categorised by the Department of Meteorological Services in Zimbabwe as a meteorological drought. Table 7.6 shows the return level estimates at selected return intervals T using the GEVD. Mean annual rainfall is expected to be below the drought threshold value of 473 mm in every return period of about $T = 8$ years.

Table 7.6: Return level estimate (mm) at selected return intervals (T) determined using the GEVD

Method	$T = 5$	$T = 10$	$T = 15$	$T = 20$	$T = 25$	$T = 30$	$T = 50$	$T = 100$
ML	510.8254	451.6059	425.8784	410.5143	399.9734	392.1458	373.4254	354.6190
Bayesian	504.4229	443.2929	416.6044	400.6187	389.6276	381.4519	361.8436	341.4674

We observed that the estimated return levels based on the Bayesian MCMC parameters were significantly smaller than their corresponding estimates based on maximum likelihood estimation method. We accept the Bayesian approach based return level estimates as they are more realistic. The minimum mean annual rainfall for Zimbabwe was 335.3 mm, recorded in the 1991/1992 rainfall season. This is the worst drought in the recorded history of the country. The return level estimate of 335.2 mm is associated with a mean return period of about 100 years. That is, we expect a severe drought of similar magnitude or worse in every 100 years' time.

There is substantial evidence in the literature about weather/climate change, due to different natural drivers and human activity. It is worthwhile to investigate the influence of time trend and climatic drivers on extreme minimum rainfall. Thus

we also modelled a non-stationary GEVD model, allowing the location parameter to vary with (i) time (to be referred to as Model 2), (ii) $\mathrm{SOI}_{\mathrm{MAY}}$ (to be referred to as Model 3), (iii) $\mathrm{SOI}_{\mathrm{AUGUST}}$ (to be referred to as Model 4), (iv) $\mathrm{SDSLP}_{\mathrm{APRIL}}$ (to be referred to as Model 5) and (v) $\mathrm{SDSLP}_{\mathrm{AUGUST}}$ (to be referred to as Model 6). Table 7.6 shows the parameter estimates with standard errors in brackets and negative log-likelihood (NLL) values for the fitted non-stationary GEVD models.

Table 7.7: The maximum likelihood parameter estimates (standard errors) and negative log-likelihood values of non-stationary GEVD Models

Model	$\hat{\theta}_0$	$\hat{\theta}_1$	$\hat{\sigma}$	$\hat{\xi}$	NLL value
2	-725.0418	0.3684	180.0839	-4555	709.9620
	(29.7976)	(0.4254)	(14.0893)	(0.0649)	
3	-703.3994	-4.2782	176.4838	-0.5028	705.1606
	(18.5678)	(1.3612)	(14.4330)	(0.0684)	
4	-710.0861	-6.3409	161.1798	-0.3952	700.9601
	(16.9167)	(1.4513)	(12.0135)	(0.0584)	
5	-693.3414	39.4236	174.5347	-0.4398	707.1581
	(18.4444)	(15.1030)	(13.3703)	0.0599	
6	-679.7833	52.6916	173.5585	-0.4746	704.7572
	(18.9762)	(14.4412)	(13.7558)	0.0683	

The residual probability (left panel) and residual quantile plots (right panel) of Model 2 presented in Figure 7.7 (Section 7.6) shows no significant departure from the straight line. This seems to suggest that the extreme minimum annual rainfall for Zimbabwe varies with time. We use the D statistic to check whether allowing the location parameter to trend with time is worthwhile. The model pair (M_1, M_2) had a D statistic of $2[-709.9620 + 710.3372] = 0.7504 < \chi^2_{1,0.05} = 3.841$, we conclude that there is no evidence of significant trends with respect to time. The residual probability and residual quantile plots for Model 3 to Model 6 shown in Figure 7.8 to Figure 7.11 (Section 7.6) show no significant departure from the straight line. This suggest

that the natural weather/climatic change drivers seem to influence extreme minimum annual rainfall for Zimbabwe. We checked whether there is significant improvement to the stationary model (Model 1) if we allow the location parameter to trend with weather/climatic change drivers using the D statistic. The model pair (M_1, M_3) had a D statistic of $2[-705.1606 + 710.3372] = 10.3532 > \chi^2_{1,0.05} = 3.841$, we conclude that there is a significant influence of the SOI value for May on extreme minimum annual rainfall for Zimbabwe. The model pair (M_1, M_4) had a D statistic of $2[-700.9601 + 710.3372] = 18.7542 > \chi^2_{1,0.05} = 3.841$, we conclude that there is a significant influence of the SOI value for August on extreme minimum annual rainfall for Zimbabwe. The model pair (M_1, M_5) had a D statistic of $2[-707.1581+710.3372] = 6.3582 > \chi^2_{1,0.05} = 3.841$, we conclude that there is a significant influence of the SDSLP value for April on extreme minimum annual rainfall for Zimbabwe. The model pair (M_1, M_6) had a D statistic of $2[-704.7572 + 710.3372] = 11.1600 > \chi^2_{1,0.05} = 3.841$, we conclude that there is a significant influence of the SDSLP value for August on extreme minimum annual rainfall for Zimbabwe.

It is important to select the best performing non-stationary GEVD model for modelling extreme minimum annual rainfall for Zimbabwe. The goodness-of-fit tests namely; RRMSE, RMAE and PPCC values were used to select the most robust model. We use the i^{th} Landwehr plotting position, $F_i = \frac{1-0.35}{n}$ in place of p in (6.22) and

$$\hat{\mu} = \hat{\theta}_0 + \hat{\theta}_1 \times \text{value of the climatic change driver}$$

to estimate $Q(F_i)$ for the models. Table 7.8 shows the goodness-of-fit values of the non-stationary models.

The best model at least twice out of the three goodness-of-fit tests is selected as the best performing model. Results indicate that Model 5 is the best performing GEVD model since it has the least RRMSE and RMAE values. Thus, this study proposes

Table 7.8: Goodness-of-fit test results for non-stationary GEVD models (location parameter influenced by climate change drivers)

Model	RRMSE	RMAE	PPCC
3	0.0512	0.0325	0.9916
4	0.0560	0.0446	**0.9945**
5	**0.0319**	**0.0238**	0.9940
6	0.0396	0.0304	0.9934

that the best performing GEVD model for mean annual rainfall for Zimbabwe is:

$$F(y) = \exp\left\{-\left[1 - 0.4398\left(\frac{y - (-693.3414 + 39.4236\text{SDSLP}_{\text{APRIL}})}{174.5347}\right)\right]^{\frac{1}{0.4398}}\right\} \tag{7.14}$$

where y is the minima mean annual rainfall i.e. $y = -x$ where x is the observed mean annual rainfall for Zimbabwe. According to Cheng et al. (2014) estimating of return levels using non-stationary GEVD models is challenging since the behaviour of physically-based covariates is not constant. In the presence of $\text{SDSLP}_{\text{APRIL}}$ values of the next 100 or more years, the return level estimates of maxima mean annual rainfall for Zimbabwe can be calculated.

7.4 Concluding remarks

We modelled extreme minimum annual rainfall in Zimbabwe using the GEVD. Exploring the duality of maxima and minima, annual rainfall data from 1901 to 2009 were fitted to the stationary GEVD. The maximum likelihood estimation and Bayesian methods were used to obtain the estimates of the parameters. The stationary GEVD parameter estimates using the Bayesian approach were close to the maximum likelihood estimates with smaller standard deviations. Using non-informative priors, the frequentist properties of the model were preserved. This confirmed that using classical statistics is akin to using Bayesian statistics with non-informative priors. Expert opinion, when available in future, can be used to improve the model further.

Model diagnostics, which included the quantile plot, the K-S and AD tests, showed that the minimum annual rainfall follows a Weibull class of distribution. Return level estimates, which are the return levels expected to be exceeded in a certain period, were calculated for Zimbabwe. The 1992 record drought is likely to return in a mean return period of $T = 100$ years. The Department of Meteorological Services in Zimbabwe categorises a year with mean annual rainfall below 473 mm as a meteorological drought year. The mean annual rainfall is expected to be less than the drought threshold value of 473 mm in a mean return period of $T = 8$ years.

The ADF test showed that the minima annual rainfall data were stationary and had no trend. The non-stationary GEVD which allows the location parameter to trend with time was found not to perform better than the stationary GEVD. We also explored the possibility of the location parameter varying with climate change drivers. The main finding of this chapter was that the non-stationary GEVD model that allows the location parameter to vary with $\mathrm{SDSLP}_{\mathrm{APRIL}}$ out performs other competing models. This proposed model should be used only as a early warning tool for drought in Zimbabwe. Care should be taken, however, in using the proposed model in extrapolating into the future using historical trends, since the behaviour of $\mathrm{SDSLP}_{\mathrm{APRIL}}$ into the future has not been investigated in this study. Furthermore, we have incorporated one natural driver as a covariate in our proposed model, but many different natural and anthropogenic climate change drivers can be added as covariates to improve the model. An area of further research can be forecasting of $\mathrm{SDSLP}_{\mathrm{APRIL}}$.

7.5 Appendix

Diagnostic plots of non-stationary GEVD models of minimum mean annual rainfall for Zimbabwe

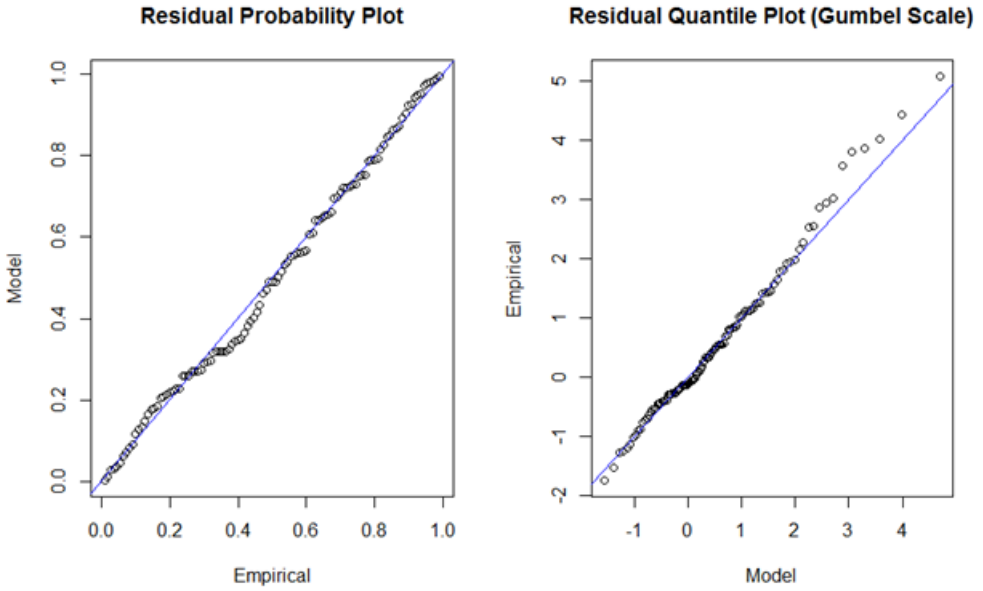

Figure 7.7: Diagnostic plot for GEVD Model 2

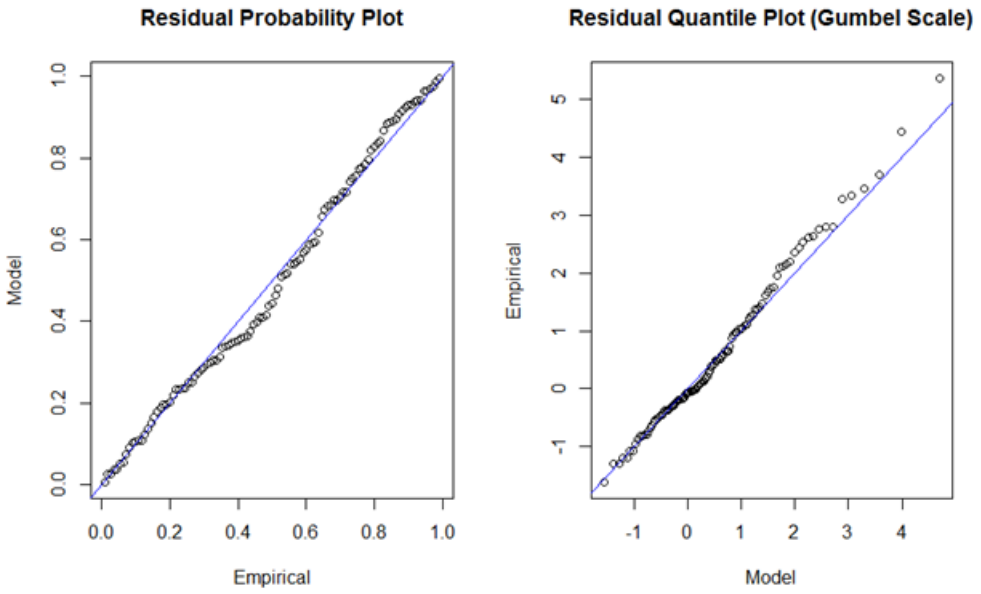

Figure 7.8: Diagnostic plot for GEVD Model 3

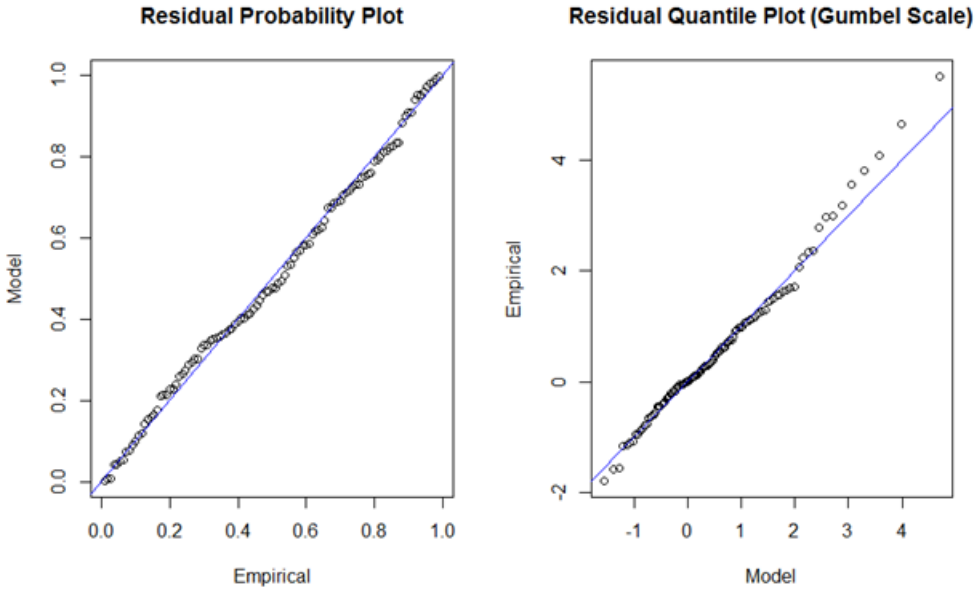

Figure 7.9: Diagnostic plot for GEVD Model 4

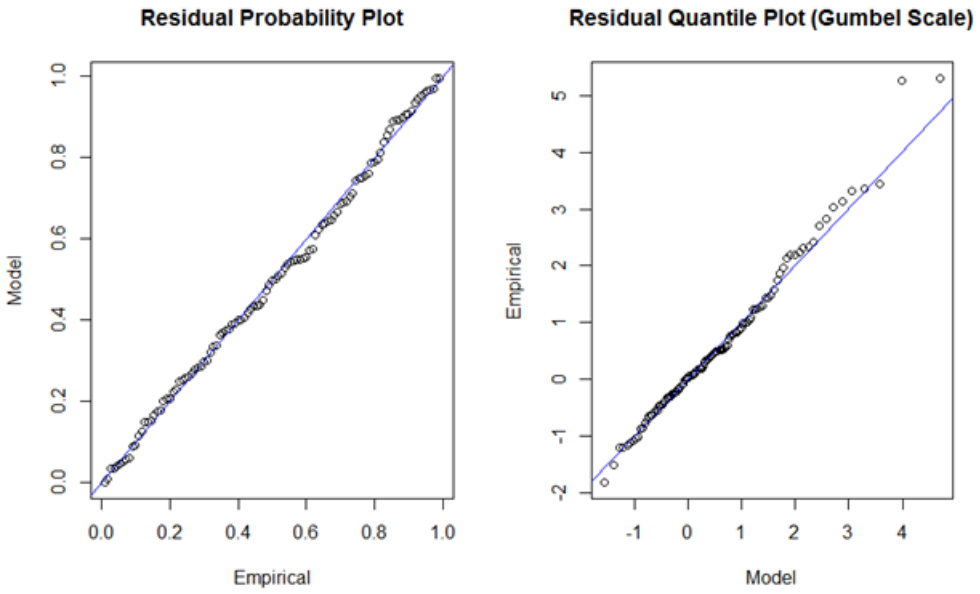

Figure 7.10: Diagnostic plot for GEVD Model 5

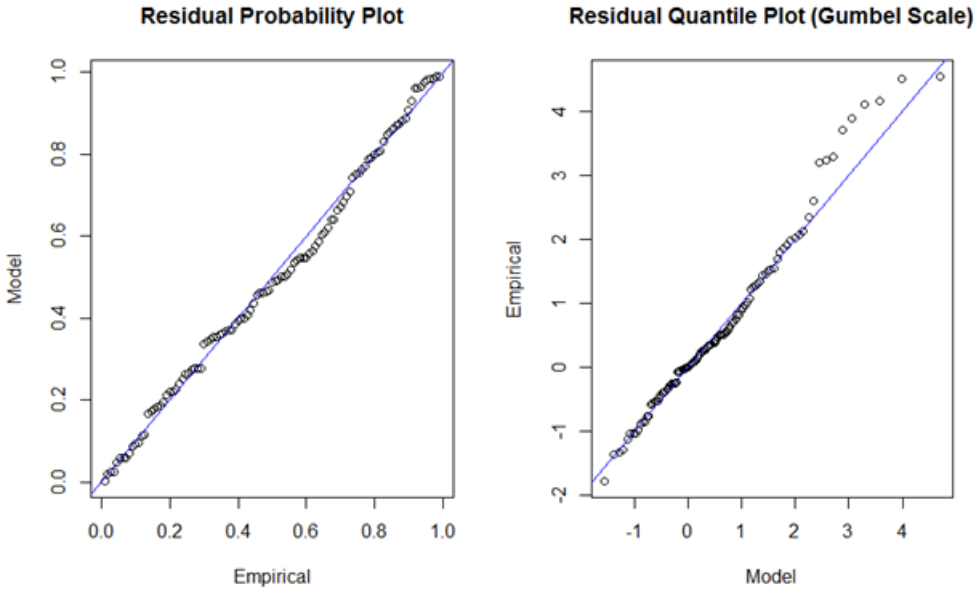

Figure 7.11: Diagnostic plot for GEVD Model 6

Chapter 8

Modelling mean annual rainfall extremes using a generalised Pareto distribution model

8.1 Introduction

Frequency analysis of extreme values of meteorological quantities have been widely used for problems related to agriculture, engineering design and risk management of buildings, roads and bridges. The information on extreme rainfall events is of obvious importance to society, especially to populations living in drought or flood risk areas (Sugahara et al., 2009). In recent years, Zimbabwe has been regularly affected by extreme rainfall that has an enormous influence on social development and threatens the country's efforts to revive its failing economy. Zimbabwe's economy is hugely dependent on rain-fed agriculture (Jury, 1996). Consequently, estimations of extreme rainfall events in Zimbabwe play a significant role in efficient risk appraisal and for the reduction of losses of the economy. Knowledge of the statistics pertaining to the occurrence of floods or droughts allows decision makers to design efficient mitigating measures. It helps in planning and constructing of water management and sewage systems. It also helps in supporting decisions regarding

how much land to use and the type of crops to grow in a particular season. Finally, knowledge of return levels of floods or droughts help inform and better prepare the citizens for cases of extreme rainfall (Brakenridge et al. 2016). To contribute to the achievement of these goals, this study focuses on modelling mean annual rainfall data for Zimbabwe using a different approach to analysing extreme values in rainfall.

Extreme value theory provides a theoretical framework to model the distribution of extreme events (Katz et al. 2002). Two approaches are commonly used for frequency analysis of extreme values: Block maxima approach and peaks-over-threshold (POT) approach. The block maxima approach consists of modelling a sequence of maximum (or minimum) values taken from a block of equal length such as the highest (or lowest) rainfall amount over the entire year. The probability distribution of extreme (maximum or minimum) values selected by this method converges asymptotically to the generalised extreme value distribution (GEVD). The block maxima method was discussed in detail in Chapter 6. The block maxima method has its shortcomings which we will discuss later in this chapter. The peaks-over threshold which utilises all data points above a specified high threshold converges asymptotically to a generalised Pareto distribution (GPD).

Applications of EVT were first published by Fisher and Tippett in 1928. Recent contributions to the statistical modelling of extremes are by Li et al. (2005), Bordi et al. (2007), McAleer et al. (2013), Ender and Ma (2014) and Li et al. (2015). The GPD model has been applied to model both stationary and non stationary rainfall amounts. Recently there has been a growing interest in investigating long-term trends in extreme rainfall attributed to weather/climate variability. Demaria et al. (2017) applied a non-stationary GEVD and GPD to observed sub-daily precipitation intensities from contrasting hydroclimatic environments in the USA. Sugahara et al. (2009) investigated the presence of long-term trends in daily rainfall of Sao Paulo,

Brazil. While Tan and Gan (2016) investigated the non stationarity of heavy precipitation over Canada using the GEVD, GPD and the Poisson distribution.

The main objective of this chapter is to perform frequency analysis of annual rainfall for Zimbabwe using GPD modelling particularly investigating whether the intensity of annual rainfall has changed over the years. Thus, frequency analysis of extreme rainfall in this study is carried out in a non-stationary context. There are at least two reasons for non-stationary modelling of extreme values of annual rainfall in Zimbabwe. The first is that over the years Zimbabwe has received a declining pattern of rainfall each year (Manatsa et al. 2008). This is attributed to global climate change which is a potential driver of extreme hydro-meteorological events (Trenberth, 1999; Felici et al., 2007; Parey et al., 2007). The second reason is related to the results obtained from some studies evidencing statistically significant positive trends in the extreme rainfall time series in other parts of the world. There is no work known to us on trend analysis of extreme rainfall in Zimbabwe.

The rest of the chapter is organised as follows: In Section 8.2 we discuss common pitfalls of the block maxima approach in modelling extreme values. We also present the research methodology in the same chapter. Section 8.3 presents the empirical results and discussion of the findings. Finally, Section 8.4 gives the concluding remarks.

8.2 Research methodology

In this section we discuss the background theory on stationary and non-stationary generalised Pareto distribution models. The data used in this chapter were described in Chapter 3. The parameters of the fitted models are estimated by the method of maximum likelihood. We first look at the pitfalls of the GEVD discussed in Chapter 6 and justify the application of GPD in modelling extreme rainfall for Zimbabwe.

8.2.1 Pitfalls of the GEVD

The block maxima approach, which is the basis for the GEVD is considered wasteful of data if more data on the extremes are available (Zhang et al., 2014). For example, if daily rainfall data is available and we use a month as the block size, then the maxima rainfall for the month is selected for modelling under this approach. Two methods have been developed to overcome this problem namely; the r largest order statistics method and the peaks-over threshold method.

8.2.2 r largest order statistics model

In Section 6.2 we supposed that $X_1, ..., X_n$ is a sequence of independent and identically distributed random variables and aimed to model the extreme behaviour of the X_i. For the r largest order statistics, we extend the limiting distribution by defining

$$M_n^k = k^{th} \text{ largest of } \{X_1, ..., X_n\}, \tag{8.1}$$

and identifying the limiting behaviour of this variable, for fixed k, as $n \to \infty$. The extremal types theorem (Theorem 8.1) is generalised by Theorem 6.4.

Theorem 8.1 *If there exist sequences of constants* $\{a_n > 0\}$ *and* $\{b_n\}$ *such that*

$$P_r \left\{ \frac{(M_n - b_n)}{a_n} \leq x \right\} \to F(x) \quad \text{as} \quad n \to \infty \tag{8.2}$$

for some nondegenerate distribution function, F*, so that* F *is the GEVD function given by (6.11), then, for fixed* k*,*

$$P_r \left\{ \frac{(M_n^k - b_n)}{a_n} \leq x \right\} \to F_k(x) \tag{8.3}$$

on $\{x : 1 + \xi(x - \mu)/\sigma > 0\}$*, where*

$$F_k(x) = \exp\{-\tau(x)\} \sum_{s=0}^{k-1} \frac{\tau(x)^s}{s} \tag{8.4}$$

with

$$\tau(x) = \left[1 + \xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}} \tag{8.5}$$

The theorem implies that, if the k^{th} largest statistic in a block is normalized in exactly the same way as the maximum, then the parameters correspond to the parameters of limiting GEVD of the block maxima. Model (8.4) is, however, difficult to use as a model. The problem is having each value of the largest r order statistic within each of several blocks, for some value of r (Smith, 1986). Therefore, in this thesis the r largest order statistic method is not used to model extreme rainfall for Zimbabwe.

8.2.3 Peaks-over threshold models

If the GEVD is considered wasteful, the r largest order statistics method is also regarded as wasteful of data as well. For example; if one block of data has a lot of extreme observations relative to other blocks, many of these may not be used as only the r largest are modelled. Modelling all observations exceeding a specified high threshold value, usually known as the peaks over threshold (POT) approach can be regarded as not wasteful of extreme data. The model of all observations exceeding a specified high threshold was first proposed by Pickands (1975) and a detailed discussion of the model was given by Davison and Smith (1990).

Suppose $X_1, ..., X_n$ is a sequence of independent and identically distributed random variables, having a common distribution G. Then, the X_i exceeding some high threshold u (usually known as threshold exceedances) can be regarded as extreme events. Let X denote an arbitrary term in the X_i sequence, then a description of the stochastic behavior of extreme events is given by the conditional probability:

$$P_r\{X > u + y \mid X > u\} = \frac{1 - G(u+y)}{1 - G(u)}, \quad y > 0. \tag{8.6}$$

The distribution of threshold exceedances would therefore be known if G was known.

In practice G is unknown, so the distribution is approximated.

8.2.4 Generalised Pareto distribution (GPD)

Theorem 8.2 *Let $X_1, ..., X_n$ be a sequence of independent and identically distributed random variables with common distribution function G, and let*

$$M_n = \max\{X_1, ..., X_n\}.$$

Let X denote an arbitrary term in the X_i sequence and suppose that G satisfies Theorem 6.1, so that for large n and a nondegenerate distribution function, F,

$$Pr\{M_n \leq x\} \approx F(x),$$

where

$$F(x) = \exp\left\{-\left[1 + \xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\} \tag{8.7}$$

for some μ, $\sigma > 0$ and ξ. Then for a specific threshold u, the distribution function of $X - u$, conditional on $X > u$, is approximately

$$F_u(y) = 1 - \left(1 + \frac{\xi y}{\widetilde{\sigma}}\right)^{-\frac{1}{\xi}} \tag{8.8}$$

defined on $\{y : y > 0 \text{ and } (1 + \xi y/\widetilde{\sigma}) > 0\}$, where

$$\widetilde{\sigma} = \sigma + \xi(u - \mu). \tag{8.9}$$

For large u, if $\xi = 0$, then

$$\lim_{u\to\infty} Pr(X - u \leq y \mid X > u) = F_u(y) = \left\{1 - \exp\left(\frac{y}{\widetilde{\sigma}}\right)^{-\frac{1}{\xi}}\right\}, y > 0 \tag{8.10}$$

which corresponds to an exponential distribution with parameter $\frac{1}{\widetilde{\sigma}}$. The generalised Pareto distribution is the name given to the distributions defined in (8.7) and (8.8)

(Scarrott and MacDonald, 2012; Magadia, 2010;). From Theorem 6.5, we can deduce that the shape parameter ξ of the GPD is in fact equal to the corresponding shape parameter of the GEVD while the scale parameter $\widetilde{\sigma}$ is equal to $\sigma+\xi(u-\mu)$. Thus, the GPD has two parameters namely; the scale parameter $\widetilde{\sigma}$ and the shape parameter ξ. The distribution function is given by:

$$F_u(y) = \begin{cases} 1 - \left(1 + \frac{\xi y}{\widetilde{\sigma}}\right)^{-\frac{1}{\xi}}, & \text{if} \quad \xi \neq 0, \\ 1 - \exp\left(\frac{y}{\widetilde{\sigma}}\right)^{-\frac{1}{\xi}}, & \text{if} \quad \xi = 0, \end{cases} \tag{8.11}$$

where $y > 0$ when $\xi \geq 0$, $0 \leq y \leq -\widetilde{\sigma}/\xi$ when $\xi < 0$, and the scale parameter $\widetilde{\sigma} > 0$ (Tsay, 2013).

The quantile function of the GPD is given by:

$$Q_N = \begin{cases} u + \frac{\widetilde{\sigma}}{\xi}\left[(\frac{n_y}{N_u}p)^{-\xi} - 1\right], & \xi \neq 0, \\ u - \widetilde{\sigma}\log(\frac{n_y}{N_u}(1-p)), & \xi = 0 \end{cases} \tag{8.12}$$

Where n_y is the number of observations per year, N is the number of exceedances above threshold u in a given sample. p, the probability of an individual observation exceeding the threshold u and

$$p = \frac{k}{n}$$

where $\frac{k}{n}$ is the sample proportion of points exceeding u. To obtain estimates of the quantiles (return levels), the estimate of $\hat{\zeta}_u$ and the maximum likelihood estimates of σ and ξ can be substituted into (8.12). Standard errors or confidence intervals of Q_N can be derived by the delta method (Rao, 1973). The mean and the variance of the GPD are $\frac{\widetilde{\sigma}}{1-\xi}$ and $\frac{\widetilde{\sigma}^2}{(1-\xi)^2(1-2\xi)}$ respectively.

8.2.5 Threshold selection

Before modelling data using the GPD, it is important to choose an appropriate threshold u. The choice of threshold is a matter of balancing bias and variance. Coles et al. (2001) pointed out that if the threshold is too low, it is more likely to violate

the asymptotic property of the model and cause bias; if the threshold is too high, it will generate few exceedances for estimation and results in high variances. A basic strategy is to select a threshold as low as possible such that the limiting approximation of the model can provide a reasonable result (Giles et al., 2016). The classical fixed threshold selection approach uses graphical diagnostics, essentially assessing aspects of the model fit, to make an a *priori* threshold choice. Some of the graphical diagnostics are Pareto quantile plot, mean excess plot and the parameter stability plot (Berning, 2010; Gomes and Gullou., 2015; de Haan and Lin, 2001;).

Pareto Quantile Plots

Beirlant et al. (1996) considered choosing the tail fraction to provide an optimal linear fit to the Pareto quantile plot. The Pareto quantile plot is a graphical method for inspecting the parameters of a Pareto distribution (Beirlant et al., 1996). If the Pareto distribution holds, there is a linear relationship between the logarithms of the observed values and the quantiles of the standard exponential distribution, since the logarithm of a Pareto distributed random variable follows an exponential distribution. The observations on the y-axis where the plot starts to follow a straight line is taken as the optimal threshold. The tail of the data follows a Pareto distribution if the logarithms of the observed values form almost a straight line. The left most point of the straight line (tail of the data) is the estimate of the optimal threshold u. The main drawback of this threshold selection procedure is after selection of the appropriate threshold there is no formal assessment of the certainty associated with the threshold choice.

Mean excess plot

Beirlant et al. (2004) defines the mean excess function with the finite expectation $E(Y) < \infty$ as:

$$e(u) = E(Y - u \mid Y > u),$$

i.e., the mean of exceedances over a threshold u. If the underlying distribution of $Y - u \mid Y > u$ follows a GPD, then the corresponding mean excess function is

$$e(u) = \frac{\widetilde{\sigma} + \xi u}{1 - \xi}, \tag{8.13}$$

provided that $\widetilde{\sigma} + \xi u > 0$ and $\xi < 1$. From (8.13), we can clearly see that the mean excess function must be linear in u. More precisely, $Y > u$ follows a GPD if, and only if, the mean excess function is linear in u (Coles et al., 2001). This provides us a way of selecting an appropriate threshold. Given the data set, we define the empirical mean excess function as:

$$\hat{e}_n(u) = \frac{\sum_{i=1}^{n}(Y_i - u)I_{\{Y_i > u\}}}{\sum_{i=1}^{n} I_{\{Y_i > u\}}}, \tag{8.14}$$

where n is the sample size. In this case the indicator function $I_{\{Y_i > u\}}$ is defined such that:

$$I_{\{Y_i > u\}} = \begin{cases} 1 & \text{if} \quad Y_i > u, \\ 0 & \text{Otherwise} \end{cases} \tag{8.15}$$

The empirical excess plot is a graphical representation of the locus of $(u, \hat{e}_n(u))$ and we can examine this plot to choose the threshold u such that $\hat{e}_n(u)$ is the lowest level where all the higher threshold based sample mean excesses are approximately linear for $Y > u$. Coles et al. (2001) acknowledged that the interpretation of the mean excess plots can be challenging.

The parameter stability plots

We exploit another useful property of the extremal index estimator to derive a further tool for threshold selection. The threshold stability of the extremal index estimator refers to its invariance to change in threshold above a suitably high threshold. That is, once a threshold is high enough, raising the threshold further should not dramatically change the estimated value of ξ. The parameter stability plot involves plotting the parameter estimates from the GPD against u, for a range values

of u. The parameter estimates should be stable (i.e. constant) above the threshold at which the GPD model becomes valid. Scarrot and MacDonald (2012) acknowledged that parameter stability plots do not provide firm conclusions, since inconsistencies are observed between the estimated shape parameter at the chosen threshold and higher thresholds.

Extremal mixture models

The drawback with fixed threshold approaches is that once the threshold has been chosen it is treated as fixed, so the associated subjectivity or uncertainty is ignored in subsequent inferences. (Scarrot and MacDonald, 2012). Extreme mixture models, in the last decade have been proposed which encapsulate the usual threshold model in combination with a component intended to capture all of the non-extreme distribution (bulk distribution) i.e. the procedure considers both the distribution of excesses (exceedances) and non-exceedances. The two distributions share information about the location of the threshold and the threshold in mixture models is defined as the parameter to be estimated. The mixture model accounts for uncertainty associated in threshold choice in subsequent inferences (Scarrot and MacDonald, 2012). The major drawbacks of the mixture models are that the asymptotic properties of their impromptu heuristics are not yet well understood and their behaviour at the threshold point is still unclear (Bernadara et al., 2014).

In this thesis, the commonly used Pareto quantile plot, mean excess plot and the parameter stability plot will be applied concurrently to select an appropriate threshold since recently developed mixture models still have challenges in their computational abilities. The subsection that follows reviews the declustering techniques associated with the POT approach and GPD distribution.

8.2.6 Declustering

The theory of extreme value distribution establishes a limiting distribution of series of random variables which are independent and identically distributed. Extreme events usually take place in clusters as a result of dependency in data (Beirlant et al., 2006; Bernadara et al., 2014). Time series data are known to be auto-correlated thus a naive selection of exceedances above a given threshold may lead to events that are no longer considered independent, but dependent. In order to reduce the dependencies of time series data, a technique called declustering is used (Coles et al., 2001; Ferro and Segers, 2003; Ribatet, 2006; Yilmaz and Perera, 2013). The R statistical package "clust" tries to identify peaks over a threshold while meeting the independence criteria. According to Ribatet (2006), clusters are identified as follows:

(i) The first exceedance initiates the first cluster;

(ii) The first observation under the threshold u "ends" the cluster unless *tim.cond* (time condition for independence) does not hold;

(iii) The next exceedance which hold *tim.cond* initiates a new cluster;

(iv) The process is iterated as needed.

The declustering approach has been criticized in the literature for being sensitive to the run length, r, that is chosen arbitrarily in cluster determination and discards of all data to leave out cluster maxima only (Coles et al., 2001).

Ferro and Segers (2003) proposed a declustering technique which involves the estimation of the extremal index, ξ and the method proposes an automatic selection of the run-length auxiliary parameter, r, used to identify independent clusters. The extremal index is given by:

$$\xi_u = \frac{2\left[\sum_{i=1}^{N-1}(T_i - 1)\right]^2}{(N-1)\sum_{i=1}^{N-1}(T_i - 1)(T_i - 1)} \tag{8.16}$$

where T_i denote the exceedance times and N is the number of threshold exceedances. In this case, u is a sufficiently high threshold. According to Coles et al. (2001), the extremal index measures the amount of clustering and $0 < \xi_u < 1$, and $\frac{1}{\xi_u}$ is the limiting mean cluster size. The exceedance times consist of two groups: one corresponding to inter-cluster times and the other corresponding to intra-cluster times. Under Ferro and Segers (2003) model, the extremal index ξ_u arises as the proportion of interdependence times which are also inter-cluster times. Assuming that we have observed N threshold exceedances, then the $\xi_u N$ largest interexceedance times are assumed to be approximately independent inter-cluster times. The value of ξ_u is used to identify the critical cut off interexceedance time that distinguishes inter from intra-cluster times. Once clusters have been identified, the cluster characteristics can be examined. Inference on these cluster maxima can be carried out and such inference is much more straight forward than the inference on the original dependent sequence.

In this thesis, we use the declustering approach as proposed by Ferro and Segers (2003). To check for goodness of fit of the Ferro and Segers (2003) model for estimated ξ_u, we use a QQ-plot of interexceedance times against standard exponential quantiles.

8.2.7 Estimation procedure of parameters for the Generalised Pareto Distribution

After determining an appropriate threshold, the parameters of a GPD can be estimated by method of maximum likelihood. Suppose we have a a sufficiently high threshold u and assume we have $y_i, ..., y_m$ exceedances over a sufficiently hight threshold from the original random variable $x_1, ..., x_n$ where $m < n$ i.e. m observations with $Y_i - u \geq 0$, the subsample $\{y_1 - u, ..., y_m - u\}$ has an underlying GPD distribution, where $Y_i - u \geq 0$ for $\xi \geq 0$, $0 \leq Y_i - u \leq -\frac{\widetilde{\sigma}}{\xi}$ for $\xi < 0$, then the logarithm

of the probability density function of y_i can be derived as:

$$\ln f_u(y_i - u) = \begin{cases} -\ln(\widetilde{\sigma}) - \frac{1+\xi}{\xi}\ln\left[1 + \xi\left(\frac{y_i - u}{\widetilde{\sigma}}\right)\right] & \text{for} \quad \xi \neq 0, \\ -\ln(\widetilde{\sigma}) - \frac{1}{\widetilde{\sigma}}(y_i - u) & \text{for} \quad \xi = 0. \end{cases} \tag{8.17}$$

Hence, the log-likelihood function $l(\xi, \widetilde{\sigma} \mid y_i - u)$ for the GPD is the logarithm of the joint density of the m observations, i.e.

$$l(\xi, \widetilde{\sigma} \mid y_i - u) = \begin{cases} -m\ln(\widetilde{\sigma}) - \frac{1+\xi}{\xi}\sum_{i=1}^{m}\ln\left[1 + \xi\left(\frac{y_i - u}{\widetilde{\sigma}}\right)\right] & \text{for} \quad \xi \neq 0, \\ -m\ln(\widetilde{\sigma}) - \frac{1}{\widetilde{\sigma}}\sum_{i=1}^{m}(y_i - u) & \text{for} \quad \xi = 0. \end{cases} \tag{8.18}$$

By differentiating (8.18) with respect to ξ we obtain:

$$\frac{dl}{d\xi} = \begin{cases} \frac{1}{\xi^2}\sum_{i=1}^{m}\ln\left[1 + \xi\left(\frac{y_i - u}{\widetilde{\sigma}}\right)\right] - \frac{1+\xi}{\widetilde{\sigma}\xi}\sum_{i=1}^{m}\frac{y_i - u}{1+\xi\left(\frac{y_i - u}{\widetilde{\sigma}}\right)}, & \text{for} \quad \xi \neq 0 \\ 0, & \text{for} \quad \xi = 0 \end{cases} \tag{8.19}$$

By differentiating (8.18) with respect to $\widetilde{\sigma}$ we obtain:

$$\frac{dl}{d\widetilde{\sigma}} = \begin{cases} -\frac{m}{\widetilde{\sigma}} + \frac{1+\xi}{\widetilde{\sigma}\xi}\sum_{i=1}^{m}\frac{\xi(y_i - u)}{\widetilde{\sigma} + \xi(y_i - u)}, & \text{for} \quad \xi \neq 0 \\ -\frac{m}{\widetilde{\sigma}} + \frac{1}{\widetilde{\sigma}^2}\sum_{i=1}^{m}(y_i - u), & \text{for} \quad \xi = 0 \end{cases} \tag{8.20}$$

The equations (8.19) and (8.20) have no analytical solutions. To obtain the estimates of ξ and $\widetilde{\sigma}$ we set the (8.19) and (8.20) to zero and solve using numerical techniques.

According to de Zea Bermudez and Turkman (2003) maximum likelihood estimators for GPD only exist for $\xi \leq 1$. In fact, if $\xi > 1$, the log-likelihood tends to infinity as $\widetilde{\sigma}/\xi$ approaches $y_{n:n}$. McNeil and Frey (2000) comment that consistency and asymptotic properties of the maximum likelihood estimators such as normality and efficiency hold uniquely if $\xi < 0.5$. According to Hosking and Wallis (1987), estimating the parameters of the GPD, algorithms used to compute maximum likelihood estimates may experience convergence problems.

8.2.8 Time-heterogenous GPD model

If we consider the stationary GPD model in (8.11) for $\xi \neq 0$, that is

$$F_u(y) = 1 - \left(1 + \frac{\xi y}{\widetilde{\sigma}}\right)^{-\frac{1}{\xi}} \tag{8.21}$$

and let (8.21) be M_0. We propose another GPD model, M_1 with scale parameter which trend with time i.e. $\ln\widetilde{\sigma}(t) = \widetilde{\sigma}_0 + \widetilde{\sigma}_1 t$ and $\xi(t) = \xi = \text{constant}$. Thus, the general form for non-stationary GPD model, M_1, is given by

$$F_u(\sigma(t), \xi, t) = 1 - \left(1 + \frac{\xi y}{\exp(\widetilde{\sigma}_0 + \widetilde{\sigma}_1 t)}\right)^{-\frac{1}{\xi}}, \quad \text{for } \xi \neq 0, \tag{8.22}$$

where $0 \leq y \leq x - u$.

8.2.9 Model diagnostics and goodness-of-fit

Before estimating return levels (8.12), we checked for model adequacy and goodness-of-fit. PP plots, QQ plots and return level plots are useful for assessing the adequacy of a fitted generalised Pareto model. Assuming a threshold u, threshold excesses $y_1 \leq ... \leq y_m$ and an estimated model $\hat{F}_u$, the probability plot consists of pairs

$$\{(i/(m+1), \hat{F}_u(y_i)); \;\; i = 1, ..., m\} \tag{8.23}$$

where

$$\hat{F}_u(y) = 1 - \left(1 + \frac{\hat{\xi} y}{\hat{\sigma}}\right)^{-\frac{1}{\hat{\xi}}}, \text{provided } \hat{\xi} \neq 0. \tag{8.24}$$

If the generalised Pareto model is adequate for modelling excesses of u, both the probability and quantile plots should consist of points that are approximately linear.

8.3 Empirical Results

In this section, the results of modelling mean annual rainfall for Zimbabwe using the GPD are presented. Results for the time-homogeneous GPD model (Model 1) and time-dependent GPD model (Model 2) are also presented. To fit Model 1 and Model 2, we need to check whether the mean annual rainfall data is independent and identically distributed (i.i.d.). In Chapter 3, the source and the properties of the rainfall data set was described and shown to be i.i.d., thus, we can fit the data to the generalised Pareto distribution.

8.3.1 Fitting time-homogeneous generalised Pareto distribution

To fit a GPD model, we check whether the tail of the mean annual rainfall data follows a Pareto distribution. Figure 8.1 shows the Pareto quantile plot for mean annual rainfall for Zimbabwe.

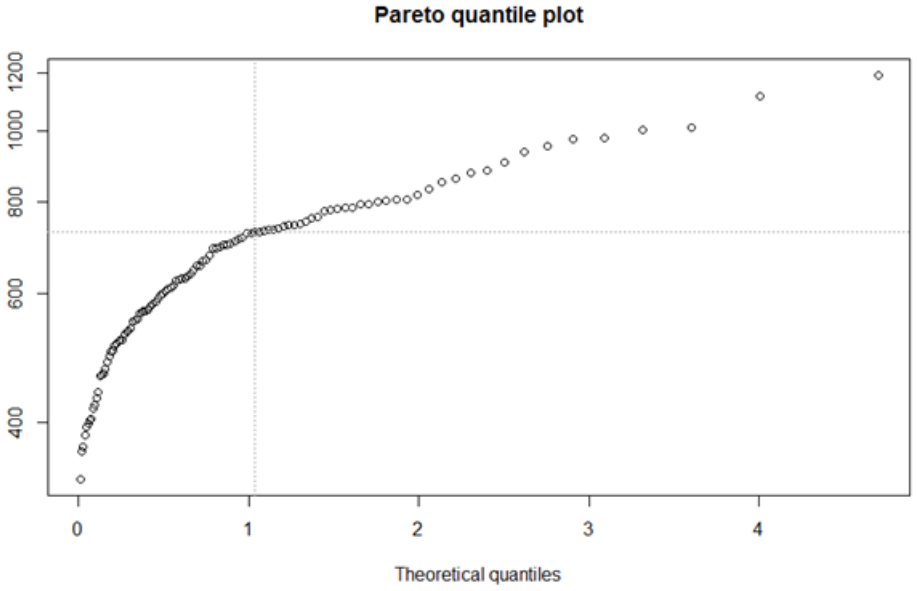

Figure 8.1: Pareto quantile plot for mean annual rainfall for Zimbabwe for the period 1901-2009.

From Figure 8.1, the tail of the data is almost a line confirming that the data may

follow a generalised Pareto distribution. To determine, the high threshold, u, the mean excess (mean residual life) and parameter stability plots were used to come up with a reasonable high threshold u. Figure 8.2 shows the mean excess plot of mean annual rainfall data for the period 1901-2009.

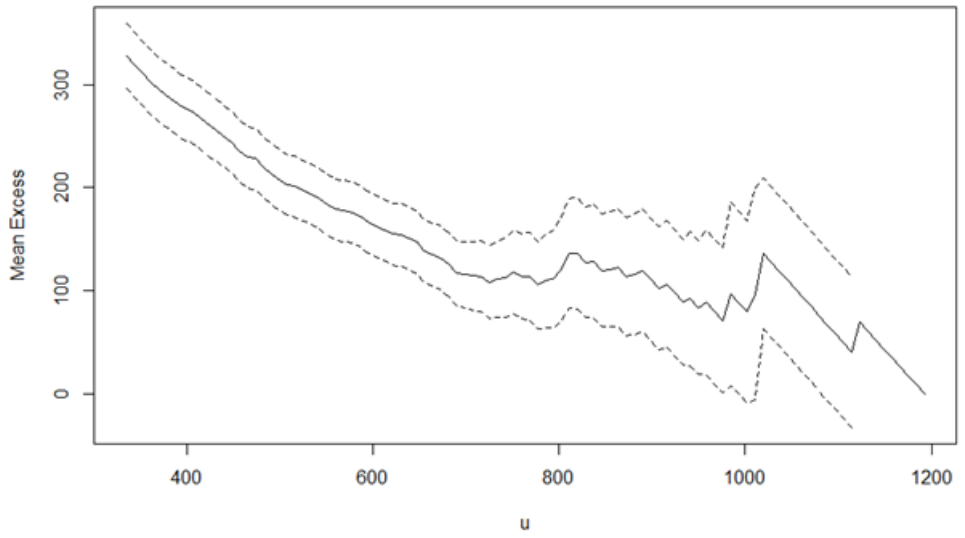

Figure 8.2: Mean excess plot for mean annual rainfall for Zimbabwe for the period 1901-2009.

The suitable threshold must lie where there is a positive gradient change in the mean excess. From Figure 8.2, the optimal threshold seems to lie around 725 mm. The selection of a suitable threshold is empirically done. To validate the threshold value, we used the parameter stability plot, shown in Figure 8.3.

Figure 8.3 shows that the estimated parameters are stable when $u \geq 725$. The Pareto quantile plot is then used to come up with the optimal threshold u of 727.9 mm. The threshold of 727.9 mm was chosen in order to meet the requirements of the bias-variance threshold trade-off, such as, it is high enough for asymptotic theorem to be considered accurate and low enough to have sufficient data to estimate the GPD

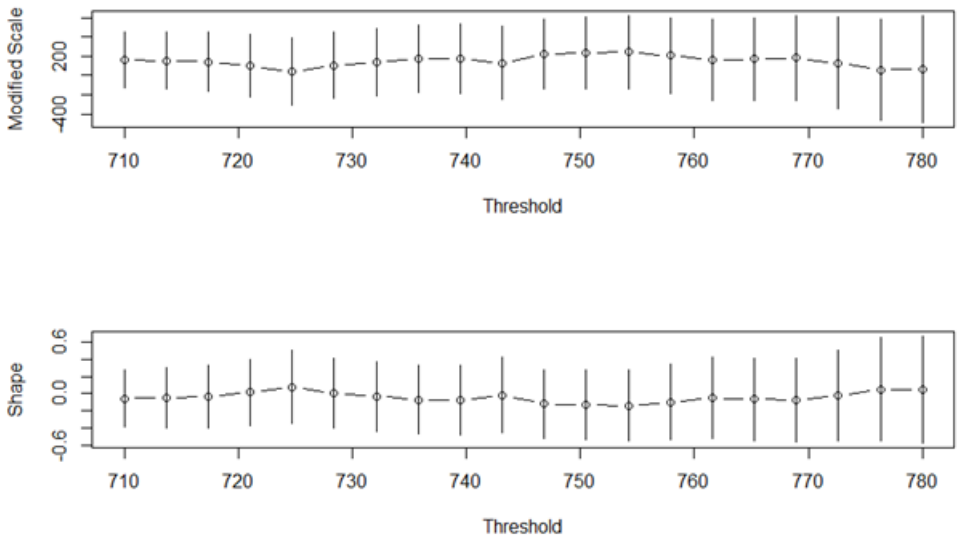

Figure 8.3: Parameter stability plot for mean annual rainfall for Zimbabwe for the period 1901-2009.

parameters. There are 38 observations above the threshold. Since the exceedances above the optimal threshold can not be assumed to be independent and identically distributed, declustering was performed and the results are shown in Figure 8.4.

We fit Model 1 (time homogenous GPD) and Model 2 (time-heterogenous GPD) to the declustered exceedances. Table 8.1 reports the maximum likelihood (ML) estimates with standard errors in the parentheses of the Model 1 and Model 2, namely; shape (ξ) and scale (σ) fitted to the mean annual rainfall data.

The shape parameter for Model 1 and Model 2 were significantly different from zero (p-values< 0.05). Figure 8.5 and Figure 8.6 shows the diagnostic plots of Model 1 and Model 2 respectively.

From Figure 8.5, the probability plot (top left panel) and the quantile plot (top right panel) suggest that the exceedances seems to follow the time-homogenous GPD

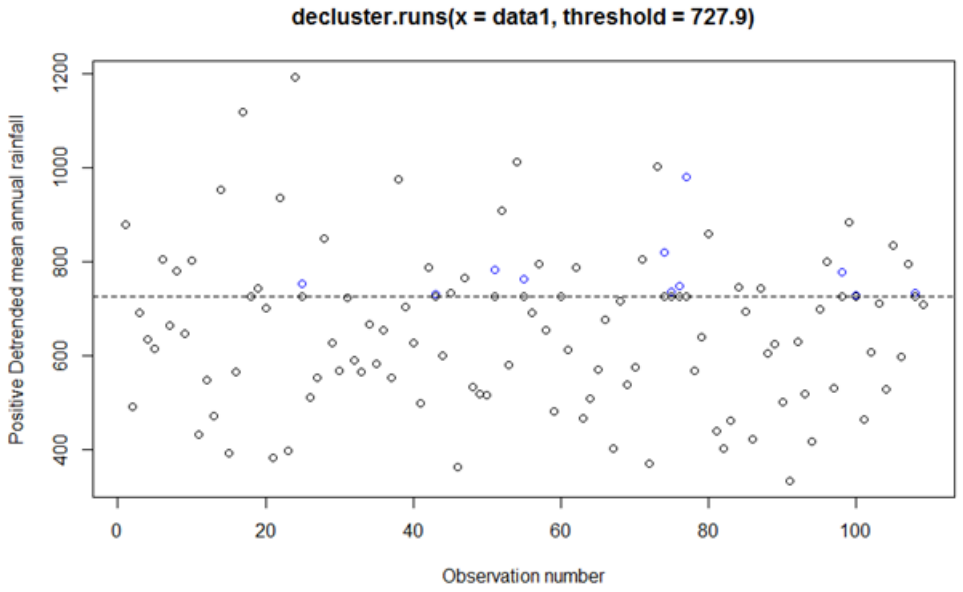

Figure 8.4: Plot of declustered exceedances for mean annual rainfall for Zimbabwe for the period 1901-2009.

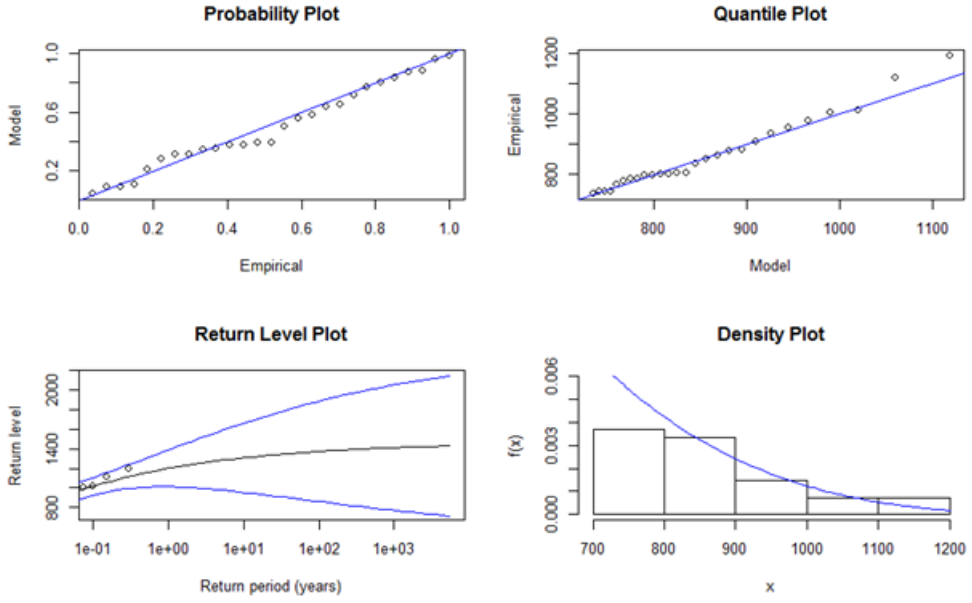

Figure 8.5: Diagnostic plots illustrating the fit of the mean annual rainfall data for Zimbabwe for the period 1901-2009 to the GPD Model 1, (a) Probability plot (top left panel), (b) Quantile plot (top right panel), (c) Return level plot (bottom left panel) and (d) Density plot (Bottom right panel)

Table 8.1: Maximum likelihood parameter estimates and negative log-likelihood of the time-homogenous and non-stationary GPD models for mean annual rainfall data for Zimbabwe

Model	$\hat{\sigma}_0$	$\hat{\sigma}_1$	$\hat{\xi}$	NLL value
M_1	166.4076	0.0000	-0.2242	159.0276
	(45.0578)		(0.1953)	
M_2	5.8406	-0.0101	-0.5145	157.2157
	(0.2947)	(0.0015)	(0.2361)	

model. The return level plot (bottom left panel) confirms that the time-homogenous GPD model is adequate to estimate return levels of the exceedances. Finally, the density plot seems consistent with the histogram of the data. Thus, all the diagnostic plots suggest that the exceedances follows the time-homogenous GPD model.

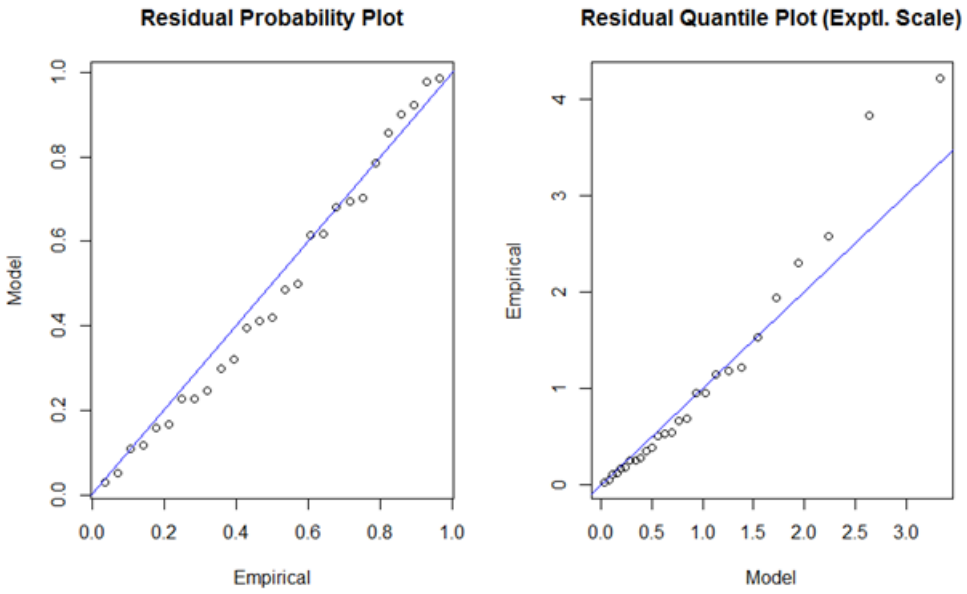

Figure 8.6: Diagnostic plots illustrating the fit of the mean annual rainfall data for Zimbabwe for the period 1901-2009 to the GPD Model 2, (a) Residual probability plot (left panel), (b) Residual quantile plot (right panel).

From Figure 8.6, the residual probability plot (left panel) shows no significant de-

parture from the straight line suggesting that the exceedances above the threshold follows a time-dependent GPD model. However, the residual quantile plot (left panel) seem to suggest that time-dependent GPD model may be adequate to estimate return levels at smaller return periods. At higher return periods the model is inadequate. The residual quantile plot departs from the straight line at the tail of the data.

In order to check whether allowing the scale parameter to trend with time is worthwhile, we use the D statistic discussed in Chapter 6. The model pair (M_1, M_2) had a D statistic value of $2(-157.2157 + 159.0276) = 3.6238$ and a critical value of $\chi^2_{1,0.05} = 3.841$. Since D statistic$= 3.6238 < \chi^2_{1,0.05} = 3.841$, this indicates that the time-dependent GPD model is not worthwhile over the time-homogenous GPD model. Thus the proposed model based on the findings of this study is the time-homogenous GPD model (Model 1). The time-homogenous GPD model for mean annual rainfall for Zimbabwe is given in (8.25)

$$F_{u=727.9}(\sigma, \xi) = 1 - \left(1 + \frac{-0.2242 y_i}{166.4076}\right)^{\frac{1}{0.2242}} \tag{8.25}$$

where $y_i = x_i - 727.9$ are the exceedances over the optimal threshold of 727.9 mm. We estimate return levels at selected time (T) intervals. Table 8.2 shows the return levels of mean annual rainfall for Zimbabwe at selected time (T) intervals.

Table 8.2: Return level estimate (mm) at selected return intervals (T) determined using the GPD Model 1

$T = 5$	$T = 10$	$T = 15$	$T = 20$	$T = 25$	$T = 30$	$T = 50$	$T = 100$
1061.5974	1120.3975	1150.7876	1170.7344	1185.3442	1196.7505	1226.3333	1531.2136

The highest mean annual rainfall of 1192.6 mm was recorded during 1924/25 rainfall season. Using the proposed GPD model, a high mean annual rainfall of this magnitude is likely to return once in every 30 year period. The return levels estimated

using the stationary GPD model (see Table 8.2) are slightly higher than the return level estimates using GEVD as reported in Chapter 6 (see Table 6.3). The reason for the slightly higher return levels using the GPD model is still unknown. One reason might be the high optimal threshold used, since the return level estimates should always greater than the threshold.

8.4 Concluding remarks

We extended the works of Li et al. (2004), Sugahara (2009) and Demaria et al. (2017) who fitted daily rainfall data to stationary and time-dependent GPD models using maximum likelihood estimation. In this chapter, we have shown that the methods can be applied to averaged annual rainfall data. We are the first to model mean annual rainfall data for a drought-prone country such as Zimbabwe using the GPD model. We modelled mean annual rainfall data for Zimbabwe using time-homogenous GPD, with a high threshold value of $u = 727.9$ mm. The study has shown that exceedances above the 727.9 mm threshold do not vary with time. Using the proposed model, the estimates of return levels at selected time intervals are slightly higher than those obtained using the stationary GEVD. According to Ferreira and de Haan (2015), the POT method is most efficient if the number of exceedances is much larger than the number of blocks. In Chapter 6, we used yearly blocks for GEVD, i.e. 109 blocks and the number of exceedance is 38 which is less than the number of blocks. However, this study has revealed that the shape parameter, ξ is less than zero and results obtained from the data and findings in this study suggest that the proposed model is reliable. Therefore, in the case where the number of observations is equal to the number of blocks, the POT method seems to produce reliable results.

Chapter 9

Generalised extreme value regressions with binary dependent variable: An application to predicting meteorological drought probabilities

9.1 Introduction

Meteorological drought although rare, impacts severely on the environment and economy of a country. Many significant events across several disciplines such as hydrology, meteorology, medicine, epidemiology and finance are rare. In hydrology and meteorology, extreme value theory is applied to model temperature, floods, minima rainfall and wind speed. The methodology for modelling occurrence of rare events are well established. The two-parameter exponential distribution derived from extreme value theory is generally used to model extreme data (see Lu, 2004). The Generalised Pareto and Generalised Extreme Value Distributions are also commonly used to model rare and extreme events in finance and hydrology (see Em-

brechts et al., 1997, Coles et al., 2001, Koutsoyiannis, 2004; Nadarajah and Choi, 2007 and Bhunya et al., 2007;). Meteorological drought is the occurrence of below normal rainfall. A rainfall season or a calendar year can be declared as a drought period if the average seasonal/annual rainfall received in that particular rainfall season or year is below the average normal rainfall. Thus, a rainfall season/year can be categorised as experiencing a drought or no drought (binary).

Droughts affect a large number of people worldwide and causes tremendous economic losses, environmental damage, social hardships, yet droughts are the least understood of all weather phenomena (Obasi, 1994). In Zimbabwe, droughts occur frequently and are often severe and impacting negatively on the country's economic performance. At least 50% of the Gross Domestic Product (GDP) in Zimbabwe is derived from rain-fed agriculture production (Jury, 1996). Thus, apart from immediate food shortages, droughts have a significant impact on the overall performance of the Zimbabwe agriculturally based economy. Over the past three decades, Zimbabwe has been ravaged by erratic rainfall patterns which has sometimes led to droughts. Since independence in 1980, Zimbabwe recorded the lowest average rainfall in the 1991/1992 rainfall season and the highest average rainfall in the 1999/2000 rainfall season (Zimbabwe Central Statistical Report, 2014). The 1991/1992 rainfall season, was influenced by an El Niño year whilst for the 1999/2000 season was affected by a La Nina year. Cyclone Eline accounted for most of the abnormal rainfall experience in the 1999/2000 rainfall season.

From year 2001 to 2003, Zimbabwe had rainfall in the first half of the rainfall season only and a dry spell in the second half, resulting in severe drought in some parts of the country. From year 2004 to 2008, Zimbabwe received an average rainfall in the northern parts of the country whilst other parts received very little rainfall or no rainfall at all. In the 2009/2010 rainfall season, 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. Generally, the rainfall patterns are becoming more and more erratic making it difficult to plan. For the 2015/2016 rainfall season, the Southern African Regional Climate Outlook Forum (SARCOF) predicted a 90% chance that an El Niño will develop during the period October-December 2015 for countries in southern Africa. The occurrence of an El Niño is likely to result in poor rains in Zimbabwe. Given this background on the rainfall and droughts patterns in Zimbabwe, research on predicting drought for the country, in any given year, becomes of paramount importance. Timely predictions will assist farmers with planning for each rainfall season, for instance; using drought resistant crops or varieties given the expected drought probability. It is imperative that a simple and accurate tool be proposed to predict the occurrence of drought at a reasonable time lag. This study argues, since droughts are rare and extreme, using the GEVD regression model can add value in the prediction of drought probabilities.

In this study we modelled binary rare events data, i.e. drought as the dependent variable. Several models of binary dependent variables have been suggested in literature by considering different link functions: logit, probit, log-log and contemporary log-log models. The logistic regression model is the most commonly used model. However, the logistic regression has been shown to underestimate the probability of rare events and the logit link function is a symmetric function, thus the response curve approaches zero as the same rate approaches one (King and Zeng, 2001). Calabrese and Osmetti (2013) proposed the Generalised Extreme Value Distribution (GEVD) regression model for modelling rare binary responses. The proposed model, however, has not been previously used to model extreme rainfall patterns. In this study, our aim are to (i) propose a model to predict meteorological drought probabilities for Zimbabwe that can be used as a drought early warning system, (ii) compare the relative performance of the logistic regression model and the GEVD regression model in predicting drought probabilities for Zimbabwe.

This chapter is organized as follows. In Section 9.2 we present the research methodology. The description of the data set for the chapter is given in Section 9.3. In Section 9.4, the proposed models are outlined. In Section 9.5, the results of the fitted models and their implications are discussed. Finally, the concluding remarks and appendix are provided in Section 9.6 and Section 9.7 respectively.

9.2 Research methodology

In this section the logistic regression model and its drawbacks in modelling rare events data are discussed. The GEVD regression model is also discussed.

9.2.1 The logistic regression model

Suppose $y_1, ..., y_n$ be realizations of Bernoulli independent random variables $Y_1, ..., Y_n$ that can take the value one and zero with probabilities π_t and $1 - \pi_t$ for $t = 1, 2, ..., n$ respectively. A generalised linear model (GLM) considers a link function and covariate vector $\mathbf{x}_t$ such that

$$g(\pi) = \boldsymbol{\beta}'\mathbf{x}_t \tag{9.1}$$

The probability π_t is obtained by applying the inverse function of $g(.)$ i.e.

$$\pi_t = g^{-1}(\boldsymbol{\beta}'\mathbf{x}_t). \tag{9.2}$$

In the logistic model, the probability π_t is a logistic cumulative distribution function

$$\pi(\mathbf{x}_t) = \frac{\exp(\boldsymbol{\beta}'\mathbf{x}_t)}{1 + \exp(\boldsymbol{\beta}'\mathbf{x}_t)} \tag{9.3}$$

with

$$\boldsymbol{\beta}' = [\beta_0, \beta_1, ..., \beta_k] \qquad \mathbf{x}' = [1, x_1, ..., x_n].$$

Thus the logit link function is

$$\text{logit}(\pi(\mathbf{x}_t)) = \ln\left(\frac{\pi(\mathbf{x}_t)}{1-\pi(\mathbf{x}_t)}\right) = \boldsymbol{\beta}'\mathbf{x}_t. \tag{9.4}$$

If we set

$$y_t = \ln\left(\frac{\pi(\mathbf{x}_t)}{1-\pi(\mathbf{x}_t)}\right) \tag{9.5}$$

The transformed logistic regression model is

$$y_t = \boldsymbol{\beta}'\mathbf{x}_t. \tag{9.6}$$

Because logistic regression predicts probabilities, rather than classes, we estimate the parameter vector $\boldsymbol{\beta}'$ using the method of maximum likelihood. The log-likelihood function of the logistic regression model is

$$l(\boldsymbol{\beta}) = \sum_{t=1}^{n} y_t(x_{tk}\beta_k) - n_t \log\left(1 - \exp\left\{\sum_{k=0}^{K} x_{tk}\beta_k\right\}\right). \tag{9.7}$$

The estimates of the parameters are obtained by setting the first derivative of the log-likelihood function with respect to each β to zero and solving the systems of equations. The resultant system of equations have no closed form solutions, we however approximately solve then numerically using the Newton's method for numerical optimization.

9.2.2 The Generalised extreme value distribution (GEVD) regression model

The logistic regression model underestimates the probability of a rare event (Calabrese and Osmetti, 2011). In meteorology and hydrology maxima of temperatures, precipitation and river discharges have been recorded for many decades (Hosking and Wallis, 1987). The Extreme Value Theorem (EVT) provides a robust theoretical frame work to model the distribution of rare events. The Generalised Extreme Value Distribution (GEVD) was recommended for hydrological frequency analysis

(Bunya et al., 2007). There are three parameters that describe the GEVD and these are: location (μ), scale (σ) and shape (ξ). The GEVD is a family of continuous probability distributions developed within extreme value theorem. The GEVD unites the Gumbel, Fréchet and Weibull class of distributions into a single family to allow for a continuous range of possible shapes. Based on the extreme value theorem the GEVD is the limiting distribution of properly normalized maxima of a sequence of independent and identically distributed random variables (Beirlant et al., 2004). Thus, the GEVD is used to model the maxima of long (finite) sequence random variables. The detailed description of the GEVD was given in Chapter 6 pages 80-84.
The quantiles of the GEVD are given in terms of parameters and the cumulative probability p by

$$Q(p) = \mu + \frac{\sigma}{\xi}\left[(-\ln(1-p))^{-\xi} - 1\right], \xi \neq 0. \tag{9.8}$$

In this study we used the quantile function of the GEVD as the link function of the GLM as proposed by Calabrese and Osmetti (2013). The GLM with the quantile function as the link function is referred to as a GEVD regression model. The link function of the GEVD regression model is given by

$$\frac{[-\ln\pi(\mathbf{x}_t)]^{-\xi} - 1}{\xi} = \boldsymbol{\beta}'\mathbf{x}_t \tag{9.9}$$

For the interpretation of the parameters β and ξ, we suppose that the value of the j-th regressor (with $j = 1, 2, ..., k$) is increased by one unit and all the other independent variables remain unchanged. Let $\mathbf{x}^*$ be the new covariate values, whereas $\mathbf{x}$ denotes the original covariate values. From (9.11), we deduce that $\beta_j = g(\pi(\mathbf{x}^*)) - g(\pi(\mathbf{x}))$ with $j = 1, 2, ..., k$. This means that if the parameter β_j (with $j = 1, 2, ..., k$) is positive and all the other parameters are fixed, by increasing the j-th regressor the estimate $\pi(\mathbf{x})$ decreases. Otherwise, if β_j is negative, by increasing the j-th regressor the estimate $\pi(\mathbf{x})$ of the GEVD model also increases.

Again, we analyse the parameter β_0: for all fixed values of ξ and for a null independent variable, β_0 has a positive monotonic relationship with the estimate of $\pi(\mathbf{x})$. Finally, we analyse the influence of the ξ parameter on $\pi(\mathbf{x})$. We find that for $\beta_0 = 0$ and by considering null values for all the covariates, from GEVD model we obtain an estimate $\pi(\mathbf{x})$ that is almost equal to e^{-1} for all the values of ξ. This means that $\pi(\mathbf{x})$ variations depend on the covariate variations and not on ξ variations. The maximum likelihood method is used to estimate the parameters of the GEVD regression model. Let $\mathbf{Y} = (Y_1, Y_2, ..., Y_n)$ be a simple random sample of size n, the log-likelihood function of the GEVD regression model is

$$l(\beta, \xi) = \sum_{t=1}^{n} \left\{ -y_t \left[1 + \xi(\boldsymbol{\beta}'\mathbf{x}_t)\right]^{-\frac{1}{\xi}} + (1 - y_t) \ln \left[1 - \exp\left\{ - \left[1 + \xi(\boldsymbol{\beta}'\mathbf{x}_t)\right]^{-\frac{1}{\xi}} \right\}\right] \right\} \tag{9.10}$$

The first derivative of the log-likelihood function with respect to each parameter are

$$\frac{dl(\beta, \xi)}{d\beta_j} = -\sum_{t=1}^{n} x_{tj} \frac{\ln[\pi(\mathbf{x}_t)]}{1 + \xi\boldsymbol{\beta}'\mathbf{x}_t} \frac{y_t - \pi(\mathbf{x}_t)}{1 - \pi(\mathbf{x}_t)}, \quad j = 0, 1, ..., k, \tag{9.11}$$

$$\frac{dl(\beta, \xi)}{d\xi} = -\sum_{i=1}^{n} \left[\frac{1}{\xi^2} \ln(1 + \xi\boldsymbol{\beta}'\mathbf{x}_t) - \frac{\boldsymbol{\beta}'\mathbf{x}_t}{\xi(1 + \xi\boldsymbol{\beta}'\mathbf{x}_t)} \right] \frac{y_t - \pi(\mathbf{x}_t)}{1 - \pi(\mathbf{x}_t)} \ln[\pi(\mathbf{x}_t)]. \tag{9.12}$$

which are set to zero and we solve the system of equations using a constrained optimization method with $\{\mathbf{x}_t : 1 + \xi\boldsymbol{\beta}'\mathbf{x}_t > 0\}$ (Calabrese and Osmetti, 2013).

9.3 The data

Average annual rainfall data for Zimbabwe over the period 1901-2015 (115 years) were obtained from Department of Meteorological Services in Zimbabwe. Meteorological drought, which is the scarcity of rainfall, is caused by various natural phenomena such as El Niña, global warming and sea level pressures. The department of Meteorological Services in Zimbabwe uses a threshold of 75% of normal

average annual rainfall to declare a year as a drought year. The average annual precipitation claims for Zimbabwe range from 663 mm (Shoko, 2014) to 867 mm (http://www.climatetemp.info/zimbabwe). This gives a range average of 760 mm. We use 570 mm (75% of 760 mm) as the drought threshold in this study. Mean annual rainfall of less than 570 mm is categorized as drought indexed 1, otherwise mean annual rainfall is categorized as no drought indexed 0. Thus, the drought index can be treated as binary data

$$Y_t = \begin{cases} 1, & \text{If mean annual rainfall } < 570\text{mm} \\ 0, & \text{Otherwise} \end{cases} \tag{9.13}$$

We aim to propose a simple early warning drought predictive model using weather/climate change drivers such as the Southern Oscillation Index (SOI) and the standardised Darwin sea level pressure (SDSLP) anomalies. The Southern Oscillation is the most notable large scale climatic variation that occurs from year to year (Panu and Sharma, 2002). This is measured using the Southern Oscillation Index (SOI) which is the difference in sea level air pressure between Tahiti (in Mid Pacific) and Darwin (in Australia) (Mason and Jury, 1997). The SOI values range from about -35 to +35. A strong and consistent negative SOI pattern is related to El Niño, the abnormal warming of ocean surface temperatures in the eastern tropical Pacific that recurs every few years. The El Niño phenomena is usually (but not always) associated with below normal rainfall. On the other hand, a strong and consistent positive SOI pattern is related to La Niña, the cooling of surface ocean waters in the eastern tropical Pacific ocean. La Niña usually precedes above normal rainfall. The monthly SOI and the SDSLP values are obtained from National Weather Service Climate Prediction Centre (http://www.longpaddock.qld.gov.au.). The drought index for each year is matched to average monthly SOI values and SDSLP anomalies observed in the same year and the previous year. Thus the data set is of the form:

$$S = \{\mathbf{x}_{t1}, \mathbf{x}_{t2}, Y_t\}, \quad \text{for } t = 1, 2, ..., 115,$$

where $\mathbf{x}_{t1}$ and $\mathbf{x}_{t2}$ are dimensional vectors containing average monthly SOI and SD-SLP values respectively that have been matched to mean annual rainfall recorded in year t and Y_t is the drought index for year t.

9.4 The models

In this chapter, we estimate drought probabilities using natural weather/climatic drivers discussed in Chapter 3. We apply the logistic and GEVD regression model with

(i) $\mathrm{SOI}_{\mathrm{MAY}}$ (to be referred to as Model 1),

(ii) $\mathrm{SOI}_{\mathrm{AUGUST}}$ (to be referred to as Model 2),

(iii) $\mathrm{SDSLP}_{\mathrm{APRIL}}$ (to be referred to as Model 3) and

(iv) $\mathrm{SDSLP}_{\mathrm{AUGUST}}$ (to be referred to as Model 4)

as covariates to obtain drought probabilities for Zimbabwe. The Deviance (D) statistic is used to check for goodness-of-fit and selection of the best performing model. The D statistic is defined as:

$$D = -2\sum_{t=1}^{n}\left[y_t \ln\left(\frac{\hat{\pi}_i}{y_t}\right) + (1 - y_t \ln\left(\frac{1-\hat{\pi}_t}{1-y_t}\right)\right] \tag{9.14}$$

where $\hat{\pi}_t = \hat{\pi}_{x_t}$. Smaller values of the D-statistic indicate that the data fits the model better. The rule of thumb for goodness-of-fit of the model is that the D statistic must be less than the degrees of freedom, which are equal to $n - p$, where n is the number of observations and p is the number of parameters estimated. A model with with D statistic which is 1.5 times the degrees of freedom is acceptable (Hosmer and Lemeshow, 2000). The model with the smallest D statistic is selected as the best logistic model. For continuous predictor variable, Hosmer and Lemeshow (2000) criticised the use of the D statistic as a model selection criterion urging that the statistics are still valid but their values become unreliable. Hosmer and Lemeshow (2000)

proposed grouping based on the values of the estimated probabilities. The Hosmer and Lemeshow, $\hat{C}$ statistics is given by

$$\hat{C} = \sum_{t=1}^{n} \frac{(y_t - \hat{\pi}_t)^2}{\hat{\pi}_t(1 - \hat{\pi}_t)} \tag{9.15}$$

where $\hat{\pi}_t$ is the estimated probabilities. The statistic follows a chi-square distribution with $n - p - 1$ degrees of freedom and p is the number of parameters. p-value of the $\hat{C}$ statistic greater than 0.05 indicates good fit. The advantage of the $\hat{C}$ is it provides a single, easily interpretable value that can be used to assess fit. The main disadvantage of the $\hat{C}$ statistic is that in the process of grouping we may miss an important deviation from fit due to a small number of individual data points. Therefore, before finally accepting that the model fits the data well, the usual residual analysis must be checked i.e. residuals are uncorrelated and have constant variance.

9.5 Empirical results

In this section, the results of modelling drought probabilities using the logistic and GEVD regression models with SOI and SDSLP values (discussed in Chapter 3) as covariates were presented. In order to avoid over-fitting of the proposed models, resulting in over-optimistic estimates of the predictive accuracy, we divide our data into in-sample data set (1901-1980) and out-of-sample data set (1981-2015).

9.5.1 Estimation results using the logistic regression model

Table 9.1 reports the parameter estimates with standard errors in brackets, deviance statistic and p-value of $\hat{C}$ statistic obtained by applying the logistic regression model using different climatic variables to the in-sample data set.

From Table 9.1, all the models are significant at 5% level of significance (p-values$<$ 0.05). Using the D statistic, Model 3 has the lowest D statistic and is selected as the best performing model. The selected model seems to fit the data quite well since

Table 9.1: Parameter estimates, deviance statistic and p-value of $\hat{C}$ statistic for logistic regression models

Model	$\hat{\beta}_0$ $\hat{\beta}_1$	Deviance statistic	p-value of $\hat{C}$ statistic
1	$-0.7923(0.2422)$ $-0.0159^{*}(0.0242)$	98.9460	0.5775
2	$-0.8134(0.2489)$	95.9650	0.1545
3	$-0.5710(0.2624)$ $0.6546(0.3256)$	95.0020	0.4335
4	$-0.6000(0.2880)$ $0.3571(0.3177)$	98.0600	0.2990

D statistic $= 95.0020 < 1.5 \text{ times } 78 \text{ degrees of freedom}$. This is confirmed by the p-value of $\hat{C}$ statistic $= 0.4335 > 0.05$. We checked for prediction accuracy for the best fitting logistic regression model. Figure 9.1 shows the time series plot of mean annual rainfall and drought years as predicted by the best fitting logistic model (grey columns).

From Figure 9.1, Model 1 seems to classify periods with mean annual rainfall below the drought threshold value of 570 mm as drought, with the notable exception of years; 1921 (385 mm), 1923 (399 mm) and 1946 (365 mm). To adequately predict drought, it is important to model the tail distribution of drought indices. We apply the GEVD regression to estimate drought probabilities using weather/climatic variables. Results indicate that the intercept is not significant at 5% level of significance. We therefore apply the GEVD regression model without intercept to check the influence of climatic variables in predicting the drought probabilities. The results of the GEVD regression models without intercept are presented in Table 9.2.

The fitted GEVD regression models are significant with p-values of the parameter

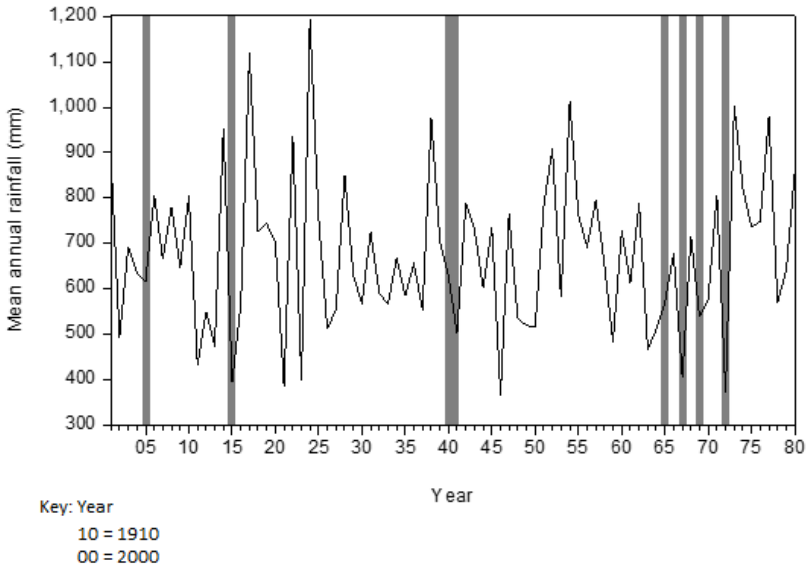

Figure 9.1: Plot of mean annual rainfall and predicted drought years from selected logistic regression model (in-sample data).

estimates < 0.05 except Model 1. The best performing model is Model 3 which has the lowest D statistic. Model 3, i.e. $g(\hat{\pi}) = 0.3455\text{SDSLP}_{\text{APRIL}}$ with the link function $g(\hat{\pi}) = \frac{[-\ln\pi(x_i)]^{\xi}-1}{\xi}$, seems to fit the data well since D statistic $= 95.5419 < 1.5 \text{ times } 79 \text{ degrees of freedom}$. This is confirmed by the p-value of $\hat{C}$ statistic $= 0.3941 > 0.05$. The residuals from Model 3 are uncorrelated and do not show significant heteroscedasticity (see ACF and PACF correlogram of residuals and squared residuals in Section 9.7). In order to access the predictive accuracy of the best GEVD regression model, we plot the time series of mean annual rainfall and check if the model would pick mean annual rainfall below the drought threshold value of 570 mm. Figure 9.2 shows the time series plot of mean annual rainfall and drought years as predicted by the best fitting GEVD regression model (grey columns).

The selected GEVD regression model seems to classify drought years better than the best fitting logistic model. From Figure 9.2 the selected GEVD regression model seems able to classify drought periods as drought except for years 1923 (399 mm) and 1946 (365 mm). This shows an improvement from the predictive accuracy of the selected logistic model. Thus, we formally compare the predictive accuracy of the

Table 9.2: Parameter estimates, deviance statistic and p-value of $\hat{C}$ statistic for GEVD regression models

Model	$\hat{\beta}_1$	Deviance statistics	p-value of $\hat{C}$ statistic
1	-0.0121	99.9430	0.5295
2	-0.0296	96.7919	0.2274
3	0.3455	95.5419	0.3941
4	0.2088	98.2806	0.2870

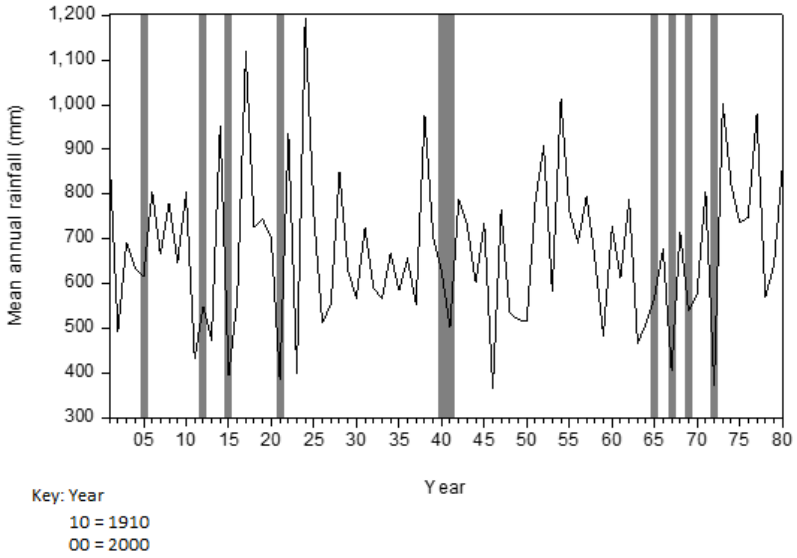

Figure 9.2: Plot of mean annual rainfall and predicted drought years from the selected GEVD regression model (in-sample data).

best performing GEVD regression here proposed with the best fitting logistic regression model. We assess the predictive accuracy of the two best fitting models using performance measures namely; the Mean Square Error (MSE) and Mean Absolute Error (MAE). The MSE and MAE are defined as

$$MSE = \frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2 \qquad MAE = \frac{1}{n}\sum_{t=1}^{n} \mid y_t - \hat{y}_t \mid \tag{9.16}$$

where y_t and $\hat{y}_t$ are the actual and the predicted drought indices respectively. The

model with a lower MSE and MAE better forecast the dependent variable.

The classification of a rainfall season as a drought (mean annual rainfall below the drought threshold) before the onset of the rainfall season is important because it is very costly to classify a rainfall season as a good rainfall season when it turns out to be a drought. Inaccurate classification of the rainfall season will lead farmers to choose the wrong type of crop for the season. For example, if the rainfall season is classified as no drought, farmers will choose crop varieties which are not drought resistant. This will lead to food shortages and non-performance of agriculture dependent industries. For this reason, we focus our attention on the tail of the response curve for values of the dependent variable equal to one that represent drought. We compare the two models by computing MSE and MAE only for drought years i.e.we consider only the positive errors $y_i - \hat{y}_t > 0$ and $n = 13$ is the number of years classified as drought. In order to avoid the problem of over-fitting the data, we calculate MSE and MAE using the out-of-sample data set (1981-2015). The in-sample and out-of-sample sizes are reported in Table 9.3.

Table 9.3: The in-sample and out-of-sample sizes

Classification	In-sample	Out-of-sample
Drought	25	13
No drought	55	22

Table 9.4 reports the MSE and MAE for the selected logistic and GEVD regression models.

From the results reported in Table 9.4, the proposed GEVD regression model performs better than the logistic regression model having the least MSE and MAE values. Therefore, the proposed model for predicting drought probabilities for Zimbabwe is $\hat{y}_i = 0.3455\text{SDSLP}_{\text{APRIL}}$, where $\hat{y}_i = \frac{[-\ln\pi(\mathbf{x})]^{-\hat{\xi}}-1}{\hat{\xi}}$. Since variations of $\pi(\mathbf{x})$ do not depend on variations of ξ, if we use the value of $\hat{\xi} = -0.4398$ for the best fit-

Table 9.4: The in-sample and out-of-sample sizes

Goodness-of-fit test	Models	
	Logistic regression	GEVD regression
MSE	0.3751	0.3608
MAE	0.5898	0.5839

ting minima annual rainfall for Zimbabwe (see Chapter 7), then any $\mathrm{SDSLP}_{\mathrm{APRIL}} \geq 0.9797$ will indicate the possibility of drought in the coming raining season.

9.6 Concluding remarks

In this study, we extended the work of Shoko and Shoko (2014) and Manatsa et al. (2008) by proposing a GLM regression model with the quantile of GEVD distribution as the link function for predicting drought probabilities for Zimbabwe. Manatsa et. al. (2008) found the MAMJ Darwin anomalies (average of March to June Darwin sea level pressure anomalies) as an earlier and more skilful predictor of Zimbabwean droughts. The main findings of this study is that the SDSLP anomaly for April alone is a skillful predictor of droughts in Zimbabwe. We have also proposed that any SDSLP anomaly greater than or equal to 0.9797 is an indicator of meteorological drought in Zimbabwe. This is a distinct departure from earlier related research cited in literature. The proposed model can be used as an early drought warning tool that reduces the risk of drought by farmers, government agencies, non-governmental organizations and citizens.

9.7 Appendix

Diagnostic plots for Model 3

Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
. | . |	. | . |	1	-0.061	-0.061	0.3091	0.578
.*| . |	.*| . |	2	-0.083	-0.087	0.8923	0.640
. | . |	. | . |	3	-0.023	-0.034	0.9376	0.816
. |*. |	. |*. |	4	0.122	0.112	2.2215	0.695
. | . |	. | . |	5	-0.057	-0.047	2.5046	0.776
.*| . |	.*| . |	6	-0.172	-0.165	5.1420	0.526
.*| . |	.*| . |	7	-0.141	-0.174	6.9256	0.437
.*| . |	.*| . |	8	-0.113	-0.194	8.0829	0.425
. | . |	. | . |	9	0.031	-0.024	8.1735	0.517
. | . |	. | . |	10	0.045	0.053	8.3602	0.594
. | . |	. |*. |	11	0.060	0.092	8.7037	0.649
.*| . |	.*| . |	12	-0.135	-0.142	10.457	0.576
. | . |	. | . |	13	0.063	-0.043	10.848	0.624
|* |	| |	14	0.149	0.044	13.056	0.522
. | . |	. | . |	15	0.052	0.028	13.333	0.577
.*| . |	. | . |	16	-0.080	-0.010	13.993	0.599
. | . |	. | . |	17	-0.061	-0.047	14.383	0.640
. | . |	.*| . |	18	-0.005	-0.073	14.385	0.704
. | . |	. | . |	19	0.010	0.029	14.396	0.760
.*| . |	.*| . |	20	-0.096	-0.095	15.408	0.753
. | . |	. | . |	21	-0.015	0.027	15.432	0.801
. |*. |	. |*. |	22	0.091	0.133	16.377	0.797
. | . |	. |*. |	23	0.074	0.100	16.999	0.809
.*| . |	.*| . |	24	-0.087	-0.137	17.891	0.808
. | . |	.*| . |	25	-0.020	-0.121	17.939	0.845
. |*. |	. | . |	26	0.097	0.040	19.094	0.832
. | . |	. | . |	27	0.055	0.046	19.472	0.852
. | . |	. | . |	28	-0.048	-0.003	19.758	0.873
.*| . |	.*| . |	29	-0.087	-0.073	20.725	0.869
. | . |	. | . |	30	-0.006	-0.061	20.730	0.896
.*| . |	.*| . |	31	-0.079	-0.122	21.571	0.896
. | . |	.*| . |	32	-0.059	-0.163	22.042	0.906
. | . |	. | . |	33	0.060	0.032	22.538	0.915
. | . |	. |*. |	34	0.035	0.098	22.711	0.930
. | . |	. |*. |	35	0.052	0.084	23.109	0.938
. | . |	. | . |	36	0.061	-0.055	23.658	0.943

Figure 9.3: ACF and PACF correlogram of residuals from the best fitting Model 3

Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
. | . |	. | . |	1	-0.030	-0.030	0.0754	0.784
. | . |	. | . |	2	0.048	0.047	0.2717	0.873
. *| . |	. *| . |	3	-0.098	-0.096	1.0954	0.778
. |* . |	. |* . |	4	0.114	0.108	2.2132	0.697
. *| . |	. *| . |	5	-0.109	-0.098	3.2582	0.660
. *| . |	. *| . |	6	-0.130	-0.155	4.7454	0.577
. *| . |	. *| . |	7	-0.099	-0.078	5.6339	0.583
. *| . |	** | . |	8	-0.183	-0.221	8.6834	0.370
. | . |	. | . |	9	0.022	0.010	8.7298	0.463
. | . |	. | . |	10	0.021	0.039	8.7706	0.554
. | . |	. *| . |	11	-0.049	-0.111	8.9980	0.622
. *| . |	. *| . |	12	-0.084	-0.090	9.6810	0.644
. |* . |	. |* . |	13	0.148	0.086	11.840	0.541
. |* . |	. |* . |	14	0.140	0.077	13.791	0.465
. | . |	. | . |	15	0.025	0.003	13.854	0.537
. | . |	. | . |	16	-0.031	-0.041	13.952	0.602
. | . |	. | . |	17	0.029	-0.006	14.041	0.664
. | . |	. *| . |	18	-0.045	-0.070	14.256	0.712
. | . |	. | . |	19	-0.013	-0.017	14.274	0.767
. *| . |	. *| . |	20	-0.095	-0.071	15.266	0.761
. | . |	. | . |	21	-0.035	0.006	15.401	0.802
. | . |	. |* . |	22	0.056	0.121	15.761	0.828
. | . |	. | . |	23	0.050	0.013	16.044	0.853
. | . |	. *| . |	24	-0.065	-0.093	16.533	0.868
. *| . |	. | . |	25	-0.070	-0.053	17.114	0.878
. | . |	. | . |	26	0.047	0.000	17.380	0.897
. |* . |	. |* . |	27	0.143	0.120	19.905	0.835
. | . |	. *| . |	28	-0.064	-0.083	20.429	0.848
. | . |	. *| . |	29	-0.057	-0.074	20.840	0.865
. | . |	. | . |	30	-0.061	-0.046	21.329	0.877
. *| . |	. *| . |	31	-0.088	-0.172	22.362	0.871
. | . |	. | . |	32	0.000	-0.015	22.362	0.897
. | . |	. | . |	33	-0.047	-0.014	22.665	0.912
. | . |	. | . |	34	-0.007	0.012	22.673	0.931
. | . |	. | . |	35	0.005	0.036	22.676	0.946
. |* . |	. | . |	36	0.139	-0.009	25.566	0.902

Figure 9.4: ACF and PACF correlogram of squared residuals from the best fitting Model 3

Chapter 10

Conclusion

10.1 Introduction

In this thesis a simple but skillful early warning tool for disaster risk reduction associated with extreme rainfall was presented. Frequency analysis of mean annual rainfall data was presented and discussed using various statistical methods. The main focus of the thesis was on the extensive application of methods based on the statistics of extremes of mean annual rainfall data series and the identification of an earlier and more skillful predictor of extreme annual rainfall for Zimbabwe.

Modelling mean annual rainfall for a drought-prone country such as Zimbabwe has potential use in disaster-risk reduction, meteorology, agriculture and in water harvesting planning and management. Zimbabwe's agro-based economy relies heavily on rain-fed agriculture, thus timely and reliable prediction of mean annual rainfall before the onset of the rainfall season can play a pivotal role in enhancing disaster preparedness and adaptation to the impacts of extreme rainfall. This thesis therefore, presented models that can be used to predict extreme mean annual rainfall at least 5 months before the onset of a rainfall season. The thesis was organised into various chapters each addressing research objectives stated in Chapter 1. The main findings of the various chapters of the thesis are presented in this chapter.

10.2 Thesis summary

Mean annual rainfall data for the period 1901-2015 obtained from the Department of Meteorological services in Zimbabwe was used as the main data set. The data set was divided into two; mean annual rainfall for the period 1901-2009 (in-sample data set for Chapters 4-7) and mean annual rainfall for the period 2010-2015 (out-of-sample data set used in Chapter 4). In Chapter 9 the data set was divided into two parts as follows: in-sample data set (1901-1980) and out-of-sample (1981-2015). The monthly Southern Oscillation Index and standardized Darwin Sea Level Pressure anomalies data set for the period 1901-2015 was obtained from NOAA. Correlation analysis was used to find correlations between mean annual rainfall and natural climatic change variables. At a longer period before the onset of the rainfall season in mid-November, SOI for May and SOI for August were found to be highly correlated with mean annual rainfall. SDSLP anomalies for April and August were also found to be highly correlated with mean annual rainfall for Zimbabwe. Correlations of SOI and SDSLP anomalies with mean annual rainfall at a lead time of more than one year were found to be statistically insignificant. This finding is not in agreement with the findings of Shoko and Shoko (2014), who found that SOI values for April and May of the previous year were correlated with the amount of rainfall. Therefore, SOI for May, SOI for August, SDSLP for April and SDSLP for August were selected as predictors of mean annual rainfall for Zimbabwe.

In order to use the selected rainfall predictors in a linear regression model, correlation analysis between the predictors was performed. SOI values for August was found to be highly correlated to the SDSLP annomalies for August. Principal component analysis was applied to extract a component of SDSLP anomaly for August which is orthogonal to the SOI value for August. Goodness-of-fit tests namely MAPE and RMSE were used to select the best regression model for predicting mean annual rainfall. Weighted regression model using the the SOI value for August and a component of SDSLP anomaly for August which is not explained by SOI as pre-

dictors was found to be the best fitting model. The proposed model has three main drawbacks. Firstly, predictions of mean annual rainfall can only be done at least two months before the onset of the rainfall season. This provides a short space of time for farmers to select the type of crop for the next rainfall season and other stakeholders in government and non-governmental organisations involved in disaster risk reduction. Secondly, the proposed model seems to slightly over-forecast the mean annual rainfall, therefore it was concluded that the predictive model still need to be improved. Finally, the linear regression model assumes that the dependent variable in this case mean annual rainfall is normal but in the literature rainfall amounts have been shown to be non linear. Therefore, in order to fully understand the mean annual rainfall patterns, the distribution analysis of mean annual rainfall was explored.

In the literature, the two-parameter gamma, two-parameter lognormal and the two-parameter log-logistics distributions are some of the distributions which usually fit to rainfall amounts well. The performance of the three candidature distributions in fitting to mean annual rainfall for Zimbabwe was assessed using RRMSE, RMAE and PPCC goodness-of-fit tests. The two-parameter gamma distribution was established at the best fitting parent distribution. This finding is consistent with the findings of Cho et al. (2004), Husak et al. (2007) and Stagge et al. (2015). This thesis noted that the tail of the parent distributions sometimes diverges in the extreme region, thus, the performance of extreme value distributions in fitting to the mean annual rainfall data were explored.

The characteristics of extreme mean annual rainfall for Zimbabwe was explored in three parts. Firstly, extreme maxima annual rainfall analysis was explored using the block maxima approach. A yearly block was used to fit the stationary GEVD model to maxima annual rainfall data using both the maximum likelihood method and Bayesian MCMC approach. In the absence of expert knowledge, non-informative prior was used. Using the Anderson-Darling test, the thesis established that the

stationary GEVD model adequately fitted the maxima mean annual rainfall for Zimbabwe. The maximum likelihood parameter estimates were found to be close to their Bayesian counter parts, but the latter produced parameter estimates with smaller standard deviations. Thus, the inclusion of the prior distribution improves the precision of tail quantile estimates. Furthermore, it was established that the return levels based on the Bayesian approach are substantially higher than their corresponding maximum likelihood based return levels. The possibility of statistics of extremes in changing climate was explored using the maximum likelihood estimation method. The GEVD model which allows the location parameter to vary with time and selected climate change drivers was fitted to maxima annual rainfall data. The thesis revealed that the extreme maxima rainfall for Zimbabwe does not vary with time. The GEVD model which incorporate SDSLP anomalies for April as a covariate of the location parameter was established to perform better than the stationary GEVD model for maxima annual rainfall for Zimbabwe. This improvement in fit establishes the indicator of meteorological volatility known as SDSLP anomalies for April as a single predictor of extreme maxima annual rainfall for Zimbabwe. This finding is very important for planning and policy-making by the government and the people of Zimbabwe.

Secondly, statistics of extreme minima annual rainfall was explored using the block maxima method. The duality of maxima and minima distributions was explored to fit GEVD to minima annual rainfall. The maximum likelihood estimation method and the Bayesian MCMC approach were used to estimate parameters of stationary the GEVD model. As noted earlier, the inclusion of the prior distribution although it was non-informative improved the estimation of return level estimates. The return level estimates, based on the Bayesian approach are significantly smaller than their corresponding return level estimates based on the maximum likelihood estimation method. Using lower expected return levels in drought risk reduction is more important, than preparing for a higher return level. Then the actual rainfall amount

received is less than expected. The possibility of statistics of extremes in changing climate was also explored using the maximum likelihood estimation method. The GEVD model which allows the location parameter to vary with time and meteorological indicators was fitted to minima annual rainfall data. The thesis revealed that extreme minima rainfall for Zimbabwe does not vary with time. The GEVD model which incorporate SDSLP anomalies for April as a covariate of the location parameter was established to perform better than the stationary GEVD model for minima annual rainfall for Zimbabwe. This improvement in fit establishes the indicator of meteorological volatility known as SDSLP anomalies for April as a single predictor of extreme minima annual rainfall for Zimbabwe. This is consistent with the findings of Manatsa et al. (2008) that the Darwin sea level pressure anomalies are a more superior predictor of Zimbabwean droughts than SOI values. Findings in this study improves on the work of Manatsa et al. (2008) and is very important for planning and policy-making by the government and the people of Zimbabwe.

Thirdly, extreme maxima annual rainfall was modelled using the POT approach of EVT. The thesis established that the time-homogenous GPD model fitted the data well. This thesis revealed that return levels, estimated using parameters of the time-homogenous GPD model were significantly higher than their corresponding estimates generated using the stationary GEVD model. The reasons for this significant difference is unknown.

Finally, this research attempted to propose an early warning drought monitoring tool for rare events. Mean annual rainfall data were categorised into drought, indexed 1 and no drought, indexed 0, thus, a binary dependent variable was generated. In order to propose a good and reliable drought monitoring tool for the country, the predictive accuracy of a logistic regression model and a GLM model, using the quantile function of GEVD as the link function were compared. The significant meteorological indicators were used as explanatory variables. This thesis estab-

lished that the GEVD regression model performed better than the logistic regression model in predicting drought probabilities. Furthermore, this thesis revealed that SDSLP anomalies for April is associated with drought probabilities. This is not surprising since this research had established SDSLP anomalies for April as a skillful predictor of mean annual rainfall for Zimbabwe. It is hoped that these findings will contribute towards decision making in Zimbabwe and help reduce the impacts of drought on people and wildlife.

10.3 Summary of the key findings

We provide a summary of key findings of this thesis, which are as follows:

1. The meteorological indicator known as the Standardized Darwin Sea Level Pressure anomalies for April is an earlier (6 months before the onset of rainfall season) and more skillful predictor of mean annual rainfall for Zimbabwe and any SDSLP anomaly of April greater than 0.9797 is an indicator of drought in the forthcoming rainfall season.

2. The gamma distribution is the most suitable parent distribution to model mean annual rainfall for Zimbabwe, however, it fails to describe the tail distribution of mean annual rainfall.

3. The non-stationary GEVD model with SDSLP anomalies of April as a covariate is the most suitable model for both maxima and minima annual rainfall for Zimbabwe.

4. Extreme annual rainfall for Zimbabwe does not change with time.

5. For suitably declustered exceedances, time-homogenous GPD models provide appropriate models for mean annual rainfall for Zimbabwe.

6. The GEVD regression model for binary rare events performs better than the logistic regression model in modelling drought probabilities for Zimbabwe.

The key objective of this thesis was to provide an early warning system for floods and droughts in Zimbabwe using extensive statistical methods. Overall, our findings appear adequate to answer the key objective. Our findings are important and contribute to decision making, enhancing knowledge about extreme annual rainfall and in developing strategies to reduce the impacts of disasters associated with extreme annual rainfall.

10.4 Limitations of the thesis

This thesis adds value to the growing body of theoretical and empirical research mainly on application of EVT methodology in meteorology. We have given a brief yet arduous treatment of EVT, providing the academic and researcher with a quick and handy overview of the main theories and models. One theoretical limitation of this thesis is some topics and details which others may have found important have been left out. However, we have striven to include the results that we consider the most important to the application of EVT in meteorology.

On the empirical front, one of the limitations of this thesis is that we only used aggregated national annual rainfall data to develop models that can be used for drought/flash floods early warning systems in Zimbabwe. It would have been interesting to develop models for drought/flash floods early warning systems for each agro-ecological region in Zimbabwe. Another limitation is that we only used SOI and SDSLP values as weather/climate predictors. Other weather/climate predictors such as wind speed, temperature etc would have been used to improve the models.

10.5 Ideas for further research

This thesis focussed on developing tools that can be used as an early warning system for drought/flash floods in Zimbabwe. However, there is still potential and need for further development and refinement, both theoretically and empirically. We there-

fore, propose possible ideas for further research.

- In this thesis aggregated national annual rainfall data is used. Further research should look at modelling mean annual rainfall for each of the natural regions in Zimbabwe. Zimbabwe is divided into 5 agro-ecological regions, known as natural regions, on the basis of the rainfall regime, soil quality and vegetation. Therefore, developing early warning systems for each natural region will assist in predicting the occurrence of drought/flash floods for a specific natural region. It will also assist in developing drought/flash floods coping mechanism for a specific region.

- One area which needs further research is investigating the applicability of spatial-zero-inflated models in modelling the daily summer rainfall count data with extra zeros in Zimbabwe.

- Modelling daily summer rainfall with weather/climate-change predictors using a Poisson-EVT generalised linear modelling approach.

References

ADIKU, S.G.K., DAYANANDA, P.W.A., ROSE, C.W. and DOWUONA, G.N.N., (1997). An analysis of the within-season rainfall characteristics and simulation of the daily rainfall in two savanna zones in Ghana. *Agricultural and Forest Meteorology*, 86(1-2), pp.51-62.

AHMAD, M.I., SINCLAIR, C.D. and WERRITTY, A., (1988). Log-logistic flood frequency analysis. *Journal of Hydrology*, 98(3-4), pp.205-224.

AKSOY, H., (2000). Use of gamma distribution in hydrological analysis. *Turkish Journal of Engineering and Environmental Sciences*, 24(6), pp.419-428.

ARSHAD, M., RASOOL, M.T. and AHMAD, M.I., (2003). Anderson Darling and modified Anderson Darling tests for generalised Pareto distribution. *Pakistan Journal of Applied Sciences*, 3(2), pp.85-88.

BALI, T.G., (2003). The generalised extreme value distribution. *Economic Letters*,79:423-427.

BALI, T.G., (2007). A generalised extreme value approach to financial risk measurement. *Journal of Money, Credit and Banking*, 39(7), pp.1613-1649.

BARTELS, R. (1982). The rank version of von Neumann's ratio test for randomness,

Journal of the American Statistical Association, 77(377), pp. 40-46.

BEIRLANT, J., TEUGELS, J.L. and VYNCKIER, P., (1996). *Practical analysis of extreme values*. Leuven University Press.

BEIRLANT, J., GOEGEBEUR, Y., SEGERS, J., de WALL, D. and FERRO, C. (2004). *Statistics of extremes: theory and applications*. John Wiley & Sons Ltd, West Sussex.

BEIRLANT, J., GOEGEBEUR, Y., SEGERS, J. and TEUGELS, J., (2006). *Statistics of extremes: theory and applications*. John Wiley & Sons Ltd, West Sussex.

BERGER, A., MELICE, J.L. and DEMUTH, C.L., (1982). Statistical distributions of daily and high atmospheric SO2-concentrations. *Atmospheric Environment (1967)*, 16(12), pp.2863-2877.

BERNARDARA, P., MAZAS, F., KERGADALLAN, X. and HAMM, L., (2014). A two-step framework for over-threshold modelling of environmental extremes. *Natural Hazards and Earth System Sciences*, 14(3), p.635.

BERNING, T.L., (2010). *Improved estimation procedures for a positive extreme value index* (Doctoral dissertation, Stellenbosch: University of Stellenbosch).

BESAG, J.,(2001). Markov chain Monte Carlo for statistical inference. *Center for Statistics and the Social Sciences*, 9, pp.24-25.

BHUNYA, P.K., JAIN, S.K., OJHA, C.S. and AGARWAL, A., (2007). Simple parameter estimation technique for three-parameter generalised extreme value distribution. *Journal of Hydrologic Engineering*, 12(6), pp.682-689.

BORDI, I., FRAEDRICH, K., PETITTA, M. and SUTERA, A., (2007). Extreme value analysis of wet and dry periods in Sicily. *Theoretical and Applied Climatology*, 87(1), pp.61-71.

BOX, G.E.P. and PIERCE, D.A., (1970). Distribution of the autocorrelations in autoregressive moving average time series models. *Journal of American Statistical Association*, 65, pp. 1509-1526.

BUCKLE, C., (1996). *Weather and climate in Africa*. Longman.

BULU, A. and AKSOY, H., (1998), June. Low flow and drought studies in Turkey. *In Low Flows Expert Meeting* pp. 133-141.

BURNSIDE, C. and DOLLAR, D. (2004). *Aid, Policies and Growth*, World Bank Research, Working Papers on International Economics, Trade and Capital Flows, No. 1777.

BRAKENRIDGE, G.R., SYVITSKI, J.P., NIEBUHR, E., OVEREEM, I., HIGGINS, S.A., KETTNER, A. and PRADES, L., (2016). *Design with nature: Causation and avoidance of catastrophic flooding*, Myanmar. Earth-Science Reviews.

BROOCK, W.A., SCHEINKMAN, J.A., DECHERT, W.D. and LEBARON, B., (1996). A test for independence based on the correlation dimension. *Econometric reviews*, 15(3), pp.197-235.

CALABRESE, R. and OSMETTI, S.A., (2011). Generalised extreme value regression for binary rare events data: an application to credit defaults. *Bulletin of the International Statistical Institute LXII, 58th Session of the International Statistical Institute*, pp.5631-5634.

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

CHENG, L., AGHAKOUCHAK, A., GILLELAND, E. and KATZ, R.W., (2014). Non-stationary extreme value analysis in a changing climate. *Climatic Change*, 127(2), pp.353-369.

CHIKOBVU, D. and CHIFURIRA, R., (2015). Modelling of extreme minimum rainfall using generalised extreme value distribution for Zimbabwe. *South African Journal of Science*, 111(9-10), pp.01-08.

CHIKOBVU, D. and SIGAUKE, C., 2013. Modelling influence of temperature on daily peak electricity demand in South Africa. *Journal of Energy in Southern Africa*, 24(4), pp.63-70.

CHIKODZI, D., ZINHIVA, H., SIMBA, F.M. and MURWENDO, T., (2013). Reclassification of agroecological zones in Zimbabwe-the rationale, methods and expected benefits: the case of Masvingo Province. *Journal of Sustainable Development in Africa*, 15(1), pp.104-116.

CHINGOMBE, W., PEDZISAI, E., MANATSA, D., MUKWADA, G. and TARU, P., (2015). A participatory approach in GIS data collection for flood risk management, Muzarabani district, Zimbabwe. *Arabian Journal of Geosciences*, 8(2), pp.1029-1040.

CHINHAMU, K. HUANG, C.K., HUANG, C.S. and HAMMUJUDDY, J., (2015). Empirical analyses of extreme value models for the South African Mining Index. *South African Journal of Economics*, 83(1), pp.41-55.

CHO, H.K., BOWMAN, K.P. and NORTH, G.R., (2004). A comparison of gamma and lognormal distributions for characterizing satellite rain rates from the tropical rainfall measuring mission. *Journal of Applied Meteorology*, 43(11), pp.1586-1597.

COLES, S., BAWA, J., TRENNER, L. and DORAZIO, P., (2001). *An introduction to statistical modeling of extreme values (Vol. 208)*. London: Springer.

COLES, S. and POWELL, E.A., (1996). Bayesian methods in extreme value modelling: a review and new developments. *International Statistics Review/Revue Internatinale de Statistique*, pp. 119-136.

COLES, S. and TAWN, J., (2005). Bayesian modelling of extreme surges on the UK east coast. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 363(1831), pp.1387-1406.

COLLIER, P. and GUNNING, J.W., (1999). Why has Africa grown slowly? *The Journal of Economic Perspectives*, 13(3), pp.3-22.

CONOVER, W.J., (1999). *Practical Nonparametric Statistics. Third Edition*, New York: John Wiley & Sons, Inc.

CRISCI, A., GOZZINI, B., MENEGUZZO, F., PAGLIARA, S. and MARACCHI, G., (2002). Extreme rainfall in a changing climate: regional analysis and hydrological implications in Tuscany. *Hydrological Processes*, 16(6), pp.1261-1274.

DALRYMPLE, T., (1960). *Flood frequency analyses*. Water Supply: Geological Survey, Reston, Virginia, USA , 1543-A.

DAN'AZUMI, S., SHAMSUDIN, S. and ARIS, A., (2010). Modelling the distribution

of rainfall intensity using hourly data. *Stochastic Environmental Research and Risk Assessment*, pp. 1-15.

DAVISON, A.C. and SMITH, R.L., (1990). Models for exceedances over high thresholds. *Journal of the Royal Statistical Society. Series B (Methodological)*, pp.393-442.

de HAAN, L. and LIN, T., (2001). On convergence toward an extreme value distribution in C [0, 1]. *Annals of probability*, pp.467-483.

DEKA, S., BORAH, M. and KAKATY, S.C., (2009). Distributions of annual maximum rainfall series of north-east India. *European Water*, 27(28), pp.3-14.

DEMARIA, E.M., GOODRICH, D. and KEEFER, T., (2017). Frequency Analysis of Extreme Sub-Daily Precipitation under Stationary and Non-Stationary Conditions across Two Contrasting Hydroclimatic Environments. *Hydrology and Earth System Sciences*, pp. 1-28.

de ZEA BERMUDEZ, P. and TURKMAN, M.A.A., (2003). Bayesian approach to parameter estimation of the generalised Pareto distribution. *Test*, 12(1), pp.259-277.

DIEBOLT, J., GUILLOU, A. and RACHED, I., (2007). Approximation of the distribution of excesses through a generalised probability-weighted moments method. *Journal of Statistical Planning and Inference*, 137(3), pp.841-857.

DODMAN, D. and MITLIN, D., (2015). The national and local politics of climate change adaptation in Zimbabwe. *Climate and Development*, 7(3), pp.223-234.

DUTTA, S., ESSADDAM, N., KUMAR, V. and SAADI, S., (2015). How does electronic trading affect efficiency of stock market and conditional volatility? Evidence

from Toronto Stock Exchange. *Research in International Business and Finance*, 39(B), pp. 867-877.

EMBRECHTS, P., RESNICK, S.I. and SAMORODNITSKY, G., (1999). Extreme value theory as a risk management tool. *North American Actuarial Journal*, 3(2), pp.30-41.

ENDER, M. and MA, T., (2014). Extreme value modeling of precipitation in case studies for China. *International Journal of Scientific and Innovative Mathematical Research (IJSIMR)*, 2(1), pp.23-36.

ENGLE, R.F., (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica: Journal of the Econometric Society*, pp.987-1007.

FAO. (2001). *Global Watch Report*. United Nations, Rome.

FARRELL, P.J. and ROGERS-STEWART, K., (2006). Comprehensive study of tests for normality and symmetry: extending the Spiegelhalter test. *Journal of Statistical Computation and Simulation*, 76(9), pp.803-816.

FELICI, M., LUCARINI, V., SPERANZA, A. and VITOLO, R., (2007). Extreme value statistics of the total energy in an intermediate-complexity model of the midlatitude atmospheric jet. Part II: trend detection and assessment. *Journal of the atmospheric sciences*, 64(7), pp.2159-2175.

FENG, S., NADARAJAH, S. and HU, Q.,(2007). Modeling annual extreme precipitation in China using the generalised extreme value distribution. *Journal of the Meteorological Society of Japan*. Series. II, 85(5), pp.599-613.

FERREIRA, A. and de HAAN, L., (2015). On the block maxima method in extreme value theory: PWM estimators. *The Annals of statistics*, 43(1), pp.276-298.

FERRO, C.A. and SEGERS, J., (2003). Inference for clusters of extreme values. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(2), pp.545-556.

FISHER, R.A. and TIPPETT, L.H.C., (1928). *Limiting forms of the frequency distribution of the largest or smallest member of a sample*. In Mathematical Proceedings of the Cambridge Philosophical Society,(Vol. 24, No. 2, pp. 180-190). Cambridge University Press.

FITZGERALD, D.L., (2005). Analysis of extreme rainfall using the log logistic distribution. *Stochastic Environmental Research and Risk Assessment*, 19(4), pp.249-257.

GAMERMAN, D., (1997). Sampling from the posterior distribution in generalised linear mixed models. *Statistics and Computing*, 7(1), pp.57-68.

GAOINI, E., DEY, D. and RUGGERI, F., (2009). *Bayesian modeling of flash floods using generalised extreme value distribution with prior elicitation*. University of Connecticut, Department of Statistics.

GELFAND, A.E. and SMITH, A.F., (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American statistical association*, 85(410), pp.398-409.

GEMAN, S. and GEMAN, D., (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*, (6), pp.721-741.

GILES, D.E., FENG, H. and GODWIN, R.T., (2016). Bias-corrected maximum likeli-

hood estimation of the parameters of the generalised Pareto distribution. *Communications in Statistics-Theory and Methods*, 45(8), pp.2465-2483.

GILKS, W.R., RICHARDSON, S. and SPIEGELHALTER, D.J., (1996). Introducing markov chain monte carlo. *Markov chain Monte Carlo in practice*, 1, p.19.

GILLI, M. and KLLEZI, E., (2006). An application of extreme value theory for measuring financial risk. *Computational Economics*, 27(2), pp.207-228.

GNEDENKO, B., (1943). Sur la distribution limite du terme maximum d'une serie aleatoire. *Annals of mathematics*, pp.423-453.

GOMES, M.I. and GUILLOU, A., (2015). Extreme value theory and statistics of univariate extremes: a review. *International Statistical Review*, 83(2), pp.263-292.

GOVERNMENT OF ZIMBABWE. (2001). *The agricultural sector of Zimbabwe*. Statistical Bulletin, Harare.

GRASSBERGER, P. and PROCACCIA, I., (1983). Measuring the strangeness of strange attractors. *Physical Review* ,9D, pp. 189-208.

GWIMBI, P., (2009). Linking rural community livelihoods to resilience building in flood risk reduction in Zimbabwe. *Jámbá: Journal of Disaster Risk Studies*, 2(1), pp.71-79.

HASAN, H., AHMAD RADI, N.F. and KASSIM, S., (2012), Modeling of extreme temperature using generalised extreme value (GEV) distribution: a case study of Penang. *In World Congress on Engineering*, 1(2012), pp. 181-186.

HARRIS, J.M., (2003). Sustainability and sustainable development. *International Society for Ecological Economics*, 1(1), pp.1-12.

HASTINGS, W. K. , (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1), 97-109.

HEO, J.H., KHO, Y.W., SHIN, H., KIM, S. and KIM, T., (2008). Regression equations of probability plot correlation coefficient test statistics from several probability distributions. *Journal of Hydrology*, 355(1-4), pp. 1-15.

HILBORN, R. and MANGEL, M., (1997). *The ecological detective: confronting models with data (Vol. 28)*. Princeton University Press.

HOELL, A., FUNK, C., ZINKE, J., and HARRISON, L., (2017). Modulation of the Southern Africa precipitation response to the El Niño Southern Oscillation by the subtropical Indian Ocean Dipole. *Climate Dynamics*, 48, pp.2529-2540.

HOSKING, J.R., (1984). Testing whether the shape parameter is zero in the generalised extreme-value distribution. *Biometrika*, 71(2), pp.367-374.

HOSKING, J.R., (1985). Algorithm as 215: Maximum-likelihood estimation of the parameters of the generalised extreme-value distribution. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 34(3), pp.301-310.

HOSKING, J.R. and WALLIS, J.R., (1987). Parameter and quantile estimation for the generalised Pareto distribution. *Technometrics*, 29(3), pp.339-349.

HOSMER, D.W. and LEMESHOW, S., (2000). *Interpretation of the fitted logistic regression model. Applied Logistic Regression*, 2^{rd} edition, pp.47-90.

HUSAK, G.J., MICHAELSEN, J. and FUNK, C., (2007). Use of the gamma distribution to represent monthly rainfall in Africa for drought monitoring applications. *International Journal of Climatology*, 27(7), pp.935-944.

ISMAIL, S.A., (1987). Long-range seasonal rainfall forecast for Zimbabwe and its relation with El-Nio/Southern Oscillation (ENSO). *Theoretical and applied climatology*, 38(2), pp.93-102.

JAYNE, T.S., GOVEREH, J., CHILONDA, P., MASON, N., CHAPOTO, A. and HAANTUBA, H., (2007). *Trends in agricultural and rural development indicators in Zambia*. Lusaka: Food Security Research Project.

JENKINSON, A. F., (1955). The frequency distribution of the annual maximum (or minimum) values of meteorological elements. *Quarterly Journal of the Royal Meteorological Society*, 81(348), pp. 158-171.

JURY, M.R., (1996). Regional teleconnection patterns associated with summer rainfall over South Africa, Namibia and Zimbabwe. *International Journal of Climatology*, 16(2), pp.135-153.

KATZ, R.W., PARLANGE, M.B. and NAVEAU, P., (2002). Statistics of extremes in hydrology. *Advances in water resources*, 25(8), pp.1287-1304.

KAUNDA, C.S., KIMAMBO, C.Z. and NIELSEN, T.K., (2012). *Hydropower in the context of sustainable energy supply: a review of technologies and challenges*. ISRN Renewable Energy.

KING, G. and ZENG, L., (2001). Logistic regression in rare events data. *Political anal-*

ysis, 9(2), pp.137-163.

KOLMOGOROV, A.N., (1933). Sulla determinazione empirica delle leggi di probabilita. Giorn. Ist. *Ital. Attuari*, 4, pp.1-11.

KORATTIKARA, A., CHEN, Y. and WELLING, M., (2014). Austerity in MCMC land: Cutting the Metropolis-Hastings budget. *In Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pp. 181-189.

KOTZ, S. and NADARAJAH, S., (2000). *Extreme value distributions*: theory and applications. World Scientific.

KOUTSOYIANNIS, D., (2004). Statistics of extremes and estimation of extreme rainfall: II. Empirical investigation of long rainfall records/Statistiques de valeurs extrmes et estimation de prcipitations extrmes: II. Recherche empirique sur de longues sries de prcipitations.*Hydrological Sciences Journal*, 49(4).

KOUTSOYIANNIS, D. AND BALOUTSOS, G., (2000). Analysis of a long record of annual maximum rainfall in Athens, Greece, and design rainfall inferences. *Natural Hazards*, 22(1), pp.29-48.

KUAN, C.M., (2008). *Lecture on time series diagnostic tests*. Institute of Economics Academia.

LEADBETTER, M.R., LINDGREN, G. and ROOTZN, H., (2012). *Extremes and related properties of random sequences and processes*. Springer Science & Business Media.

LÉLÉ, S.M., (1991). Sustainable development: a critical review. *World development*, 19(6), pp.607-621.

LI, Z., LI, Z., ZHAO, W. and WANG, Y., (2015). Probability modeling of precipitation extremes over two river basins in northwest of China. *Advances in Meteorology*,

LI, Z., BRISSETTE, F. and CHEN, J., (2013). Finding the most appropriate precipitation probability distribution for stochastic weather generation and hydrological modelling in Nordic watersheds. *Hydrological Processes*, 27(25), pp.3718-3729.

LI, Y., CAI, W. and CAMPBELL, E.P., (2005). Statistical modeling of extreme rainfall in southwest Western Australia. *Journal of climate*, 18(6), pp.852-863.

LU, H.C., (2004). Estimating the emission source reduction of PM 10 in central Taiwan. *Chemosphere*, 54(7), pp.805-814.

LUDWIG, D., (1996). Uncertainty and the assessment of extinction probabilities. *Ecological Applications*, 6(4), pp.1067-1076.

LJUNG, G.M. and BOX, G.E.P., (1978). On measure of lack of fit in time series model. *Biometrika*, 65, pp. 297-303.

MACLEOD, A.J., (1989). A remark on algorithm AS 215: Maximum-likelihood estimation of the parameters of the generalised extreme-value distribution. *Applied Statistics*, 38(1), pp.198-199.

MADAMOMBE, E.K., (2004). *Zimbabwe: Flood management practices-Selected flood prone areas Zambezi Basin*. Unpublished Paper, Wmo/Gwp Associated Programme on Flood Management.

MADDALA, G.S. and KIM, I.M., (1998). *Unit roots, cointegration, and structural change*

(No. 4). Cambridge University Press.

MADSEN, H., RASMUSSEN, P.F. and ROSBJERG, D., (1997). Comparison of annual maximum series and partial duration series methods for modeling extreme hydrologic events: 1. At-site modeling. *Water Resources Research*, 33(4), pp.747-757.

MAGADIA, J., (2010). Value-at-risk modeling via the peaks-over-threshold approach. *Annual Bangko Sentral ng Pilipinas-University of the Philippines Professional Chair Lectures*.

MAKARAU, A. (1995). Intra-seasonal oscillatory models of the Southern Africa summer circulation, Unpublished PhD Thesis, University of Cape Town.

MAMOMBE, V., KIM, W. and CHOI, Y.S., (2017). Rainfall variability over Zimbabwe and its relation to large-scale atmosphere-ocean process. *International Journal of Climatology*, 37, pp. 963-971.

MANATSA, D., CHINGOMBE, W., MATSIKWA, H. and MATARIRA, C.H., (2008). The superior influence of Darwin Sea level pressure anomalies over ENSO as a simple drought predictor for Southern Africa. *Theoretical and Applied Climatology*, 92(1), pp.1-14.

MANATSA, D., MUSHORE, T. and LENOUO, A. , (2017). Improved predictability of drought over southern Africa using the standardized precipitation evapotranspiration index and ENSO. *Theoretical and Applied Climatology*, 127, pp.259-274.

MANGOYANA, S. and MEDA, T., (2001). *Food Security and Sustainability in Zimbabwe*. University of Zimbabwe Publications, Harare.

MANNING, W.G. and MULLAHY, J., (2001). Estimating log models: to transform or not to transform?. *Journal of health economics*, 20(4), pp.461-494.

MARTINS, E.S. and STEDINGER, J.R., (2000). Generalised maximum likelihood generalised extreme value quantile estimators for hydrologic data. *Water Resources Research*, 36(3), pp.737-744.

MASON, S.J., and JURY, M.R., (1997). Climatic variability and change over Southern Africa: A reflection on underlying process. *Progress in Physical Geography*, 21(1), pp. 23-50.

MATARIRA, C.H., (1990). Drought over Zimbabwe in a regional and global context. *International Journal of Climatology*, 10(6), pp.609-625.

MATARIRA, C.H. and UNGANAI, L.S., (1994).*A rainfall prediction scheme for southern Africa based on the southern oscillation*. Technical report, SADC-REWU, Harare.

MAYOORAN, T. and LAHEETHARAN, A., (2014). The Statistical Distribution of Annual Maximum Rainfall in Colombo District. *Sri Lankan Journal of Applied Statistics*, 15(2).

MAZVIMAVI, D., (2008). Investigating possible changes of extreme annual rainfall in Zimbabwe. em Hydrology & Earth System Sciences Discussions, 5(4), pp.1765-1784.

MAZVIMAVI, D., (2010). Investigating changes over time of annual rainfall in Zimbabwe. *Hydrology and Earth System Sciences*, 14(12), pp.2671-2679.

McALEER, M., JIMÉNEZ-MARTÍN, J.Á. and PÉREZ-AMARAL, T., (2013). Has the Basel Accord improved risk management during the global financial crisis? *The*

North American Journal of Economics and Finance, 26, pp.250-265.

McKEE, T.B., DOESKEN, N.J., and KLEIST, J., (1993). The relationship of drought frequency and duration to time scales. *In Proceedings of the 8th Conference on Applied Climatology*, 17(22), 179-183.

McNEIL, A.J. and FREY, R., (2000). Estimation of tail-related risk measures for heteroscedastic financial time series: an extreme value approach. *Journal of empirical finance*, 7(3), pp.271-300.

METROPOLIS, N., ROSENBLUTH, A.W., ROSENBLUTH, M.N., TELLER, A.H., and TELLER, E., (1953). Equation of state calculations by fast computing machine. *Journal of Chemical Physics*, 21, 1087-1091.

MINKA, T.P., (2002). *Estimating a Gamma distribution*. Microsoft Research, Cambridge, UK, Technical. Report.

MUCHURU, S., LANDMAN, W,A., De WITT, D. and LÖTTER, D., (2014). Seasonal rainfall predicability over the Lake Kariba catchment area. *Water SA*, 40, pp. 461-470.

MUSHORE, T.D., (2013). Climatic changes, erratic rains and the necessity of constructing water infrastructure: Post 2000 land reform in Zimbabwe. *International Journal of Scientific & Technology Research*, 2(8), pp. 304-310.

NADARAJAH, S. and CHOI, D., (2007). Maximum daily rainfall in South Korea. *Journal of Earth System Science*, 116(4), pp.311-320.

NATIONAL ENVIRONMENTAL RESEARCH COUNCIL (NERC). (1975). *Flood studies report*. Wallingford: Institute of Hydrology.

NGUYEN, C.C., GAUME, E. and PAYRASTRE, O., (2014). Regional flood frequency analyses involving extraordinary flood events at ungauged sites: further developments and validations. *Journal of Hydrology*, 508, pp.385-396.

NIYIMBANIRA, F., (2013). An Econometric Evidence of the Interactions between Inflation and Economic Growth in South Africa. *Mediterranean Journal of Social Sciences*, 4(13), p.219.

OBASI, G.O.P., (1994). WMO's role in the international decade for natural disaster reduction. *Bulletin of the American Meteorological Society*, 75(9), pp.1655-1661.

PANU, U.S. and SHARMA, T.C., (2002). Challenges in drought research: some perspectives and future directions. *Hydrological Sciences Journal*, 47(S1), pp.S19-S30.

PAREY, S., MALEK, F., LAURENT, C. and DACUNHA-CASTELLE, D., (2007). Trends and climate evolution: Statistical approach for very high temperatures in France. *Climatic Change*, 81(3), pp.331-352.

PERMAN, R., MA, Y., McGILVRAY, J. and COMMON, M., (2003). *Natural Resource and Environmental Economics* 3^{rd} edition, New York: Pearson Addison Wesley.

PICKANDS III, J., (1975). Statistical inference using extreme order statistics. *The Annals of Statistics*, pp.119-131.

POCERNICH, M.J. (2002). *Application of Extreme Value Theory and Threshold models to hydrological event*, (Master's thesis, University of Colorado).

PRESCOTT, P. and WALDEN, A.T., (1983). Maximum likelihood estimation of the

parameters of the three-parameter generalised extreme-value distribution from censored samples. *Journal of Statistical Computation and Simulation*, 16(3-4), pp.241-250.

RAKHECHA, P.R. and SOMAN, M.K., (1994). Trends in the annual extreme rainfall events of 1 to 3 days duration over India. *Theoretical and Applied Climatology*, 48(4), pp.227-237.

RAO, C.R. (1973). *Linear Statistical Inference and its Applications*, 2^{rd} edition. Wiley, New York.

RAZALI, N.M. and WAH, Y.B., (2011). Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. *Journal of statistical modeling and analytics*, 2(1), pp.21-33.

RENARD, B., LANG, M. and BOIS, P., (2006). Statistical analysis of extreme events in a non-stationary context via a Bayesian framework: case study with peak-over-threshold data. *Stochastic environmental research and risk assessment*, 21(2), pp.97-112.

RESNICK, S.I., (2013). *Extreme values, regular variation and point processes*. Springer.

RICHARD, Y., TRZASKA, S., ROUCOU, P. and ROUAULT, M., (2000). Modification of the southern African rainfall variability/ENSO relationship since the late 1960s. *Climate Dynamics*, 16(12), pp.883-895.

RIBATET, M., (2006). *A user's guide to the POT package (version 1.4). month. ROCHA, A.M.C., (1992). The influence of global sea surface temperatures on southern African summer climate* (Doctoral dissertation, University of Melbourne).

ROCHA, A.M.C., (1992). *The influence of global sea surface temperatures on Southern*

African summer climate, PhD Thesis, University of Melbourne.

ROGHANI, R., SOLTANI, S. and BASHARI, H., (2016). Influence of Southern Oscillation on autumn rainfall in Iran (1951-2011). *Theoretical and Applied Climatology*, 124(2), pp.411-423.

ROPELEWSKI, C.F. and HALPERT, M.S., (1987). Global and regional scale precipitation patterns associated with the El Nio/Southern Oscillation. *Monthly weather review*, 115(8), pp.1606-1626.

ROTH, M., BUISHAND, T.A., JONGBLOED, G., TANK, A.K. and VAN ZANTEN, J.H., (2014). Projections of precipitation extremes based on a regional, non-stationary peaks-over-threshold approach: A case study for the Netherlands and north-western Germany. *Weather and Climate Extremes*, 4, pp.1-10.

RURINDA, J., MAPFUMO, P., van WIJK, M.T., MTAMBANENGWE, F., RUFINO, M.C., CHIKOWO, R. and GILLER, K.E., (2014). Sources of vulnerability to a variable and changing climate among smallholder households in Zimbabwe: A participatory analysis. *Climate Risk Management*, 3, pp. 65-78.

RURINDA, J., MAPFUMO, P., van WIJK, M.T., MTAMBANENGWE, F., RUFINO, M.C., CHIKOWO, R. and GILLER, K.E., (2013) Managing soil fertility to adapt to rainfall variability in smallholder cropping systems in Zimbabwe. *Field Crops Research*, 154, pp. 211-225.

SACHS, J.D. and WARNER, A.M., (1997). Sources of slow growth in African economies. *Journal of African economies*, 6(3), pp.335-376.

SAKULSKI, D., JORDAAN, A., TIN, L. and GREYLING, C., (2014). *Fitting theoretical*

distributions to Rainy Days for Eastern Cape Drought Risk assessment. In Proceedings of Daily Meteo. org/2014 Conference (p. 48).

SATISH, P., GIRI, R.K. and SATISH, C., (2014). Environmental hazard-landslides and avalances (Kasmir region). *International Journal of Physics and Mathematical Sciences*, 4(3), pp. 87-99.

SCARROTT, C. and MACDONALD, A., (2012). A review of extreme value threshold estimation and uncertainty quantification. *REVSTAT-Statistical Journal*, 10(1), pp.33-60.

SCHMIDLI, J., GOODESS, C.M. FREI, C., HAYLOCK, M.R., HUNDECHA, Y., RIBALAYGUA, J. and SCHMITH, T., (2007). Statistical and dynamic downscaling of precipitation: An evaluation and comparison of scenarios for the European Alps. *Journal of Geophysical Research Atmospheres*, 112(D4).

SHOKO, K. and SHOKO, N., (2014). Drought and El nino phases in Zimbabwe. *The Dyke*, 17(8.3), pp. 1-18.

SMITH, S.V., (1985a). Studies of the effects of cold fronts during the rainy season in Zimbabwe. *Weather*, 40(7), pp.198-203.

SMITH, E., (2005). *Bayesian modelling of extreme rainfall data* (Doctoral dissertation, University of Newcastle upon Tyne).

SMITH, R.L., (1985b). Maximum likelihood estimation in a class of nonregular cases. *Biometrika*, 72(1), pp.67-90.

SMITH, R.L., 1986. Extreme value theory based on the r largest annual events. *Jour-*

nal of Hydrology, 86(1-2), pp.27-43.

SMITH, R.L. and GOODMAN, D.J., (2000). Bayesian risk analysis. *Extremes and Integrated Risk Management*, pp.235-251.

SMITH, R.L. and NAYLOR, J.C., (1987). A comparison of maximum likelihood and Bayesian estimators for the three-parameter Weibull distribution. *Applied Statistics*, pp.358-369.

STAGGE, J.H., TALLAKSEN, L.M., GUDMUNDSSON, L., VAN LOON, A.F. and STAHL, K., (2015). Candidate distributions for climatological drought indices (SPI and SPEI). *International Journal of Climatology*, 35(13), pp.4027-4040.

STEPHENS, M.A., (1974). EDF statistics for goodness of fit and some comparisons. *Journal of the American statistical Association*, 69(347), pp.730-737.

SUGAHARA, S., DA ROCHA, R.P. and SILVEIRA, R., (2009). Non stationary frequency analysis of extreme daily rainfall in Sao Paulo, Brazil. International Journal of Climatology, 29(9), pp.1339-1349.

SUHAILA, J., CHING-YEE, K., FADHILAH, Y. and HUI-MEAN, F., (2011). Introducing the mixed distribution in fitting rainfall data. *Open Journal of Modern Hydrology*, 1(2), pp. 11.

SUHAILA, J. and JEMAIN, A.A., (2007). Fitting daily rainfall amount in Peninsular Malaysia using several types of exponential distributions. *Journal of applied sciences research*, 3(10), pp.1027-1036.

SURMAN, P.G., BODERO, J. and SIMPSON, R.W., (1987). The prediction of the num-

bers of violations of standards and the frequency of air pollution episodes using extreme value theory. *Atmospheric Environment (1967)*, 21(8), pp.1843-1848.

TAN, X. and GAN, T.Y., (2017). Non-stationary analysis of the frequency and intensity of heavy precipitation over Canada and their relations to large-scale climate patterns. *Climate Dynamics*, 48(9-10), pp.2983-3001.

TRENBERTH, K.E., (1999). *Conceptual framework for changes of extremes of the hydrological cycle with climate change*. In Weather and Climate Extremes (pp. 327-339). Springer Netherlands.

TRÚCK, S. and LIANG, K., (2012). Modelling and forecasting volatility in the gold market. *International Journal of Banking and Finance*, 9(1), pp. 48-80.

TORRANCE, J.D., (1990). The Southern Oscillation and the rainy season in Zimbabwe. *Zimbabwe Science News*, 24, pp.4-6.

TOSH, C. and DASGUPTA, S., (2014). Lower bounds for the gibbs sampler over mixtures of Gaussians. *In Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pp. 1467-1475.

TSAY, R.S., (2013). *Multivariate Time Series Analysis: with R and financial applications*. John Wiley & Sons.

UNECA. (2000). *HIV/AIDS and education in Eastern and Southern Africa: Leadership challenge and the way forward*. Addis Ababa, UNECA.

UNGANAI, L.S., (1996). Historic and future climatic change in Zimbabwe. *Climate research*, pp.137-145.

von MISES, R. (1954). La distribution de la plus grande de n valeurs. *American Mathematical Society, Providence, RI, Selected Papers*, 2, pp. 271-294.

WASHINGTON, R. and PRESTON, A., (2006). Extreme wet years over Southern Africa: Role of Indian Ocean sea surface temperatures. Journal of Geophysical Research: *Atmospheres*, 111(D15).

WAYLEN, P.E.T.E.R. and HENWORTH, S., (1996). A note on the timing of precipitation variability in Zimbabwe as related to the Southern Oscillation. *International journal of climatology*, 16(10), pp.1137-1148.

WEBSTER, P.J., (1981). Mechanisms determining the atmospheric response to sea surface temperature anomalies. *Journal of the Atmospheric Sciences*, 38(3), pp.554-571.

WILLMOTT, C.J. and MATSUURA, K.,(2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate research*, 30(1), pp.79-82.

YILMAZ, A.G. AND PERERA, B.J.C., (2013). Extreme rainfall non-stationarity investigation and intensity-frequency-duration relationship. *Journal of Hydrologic Engineering*, 19(6), pp.1160-1172.

ZHANG, T., GEORGIOPOULOS, M. and ANAGNOSTOPOULOS, G.C., (2014). *Online model racing based on extreme performance*. In Proceedings of the 2014 Annual Conference on Genetic and Evolutionary Computation (pp. 1351-1358). ACM.

ZIMBABWE CENTRAL STATISTICAL OFFICE HANDBOOK (2014). *October-Environmental Statistics Report*.

ZIMBABWE CENTRAL STATISTICAL OFFICE HANDBOOK (2010). *October-Environmental Statistics Report*.

ZIMBABWE CENTRAL STATISTICAL OFFICE HANDBOOK (1994). *October-Environmental Statistics Report*.

ZIN, W.Z.W., JEMAIN, A.A. and IBRAHIM, K., (2009). The best fitting distribution of annual maximum rainfall in Peninsular Malaysia based on methods of L-moment and LQ-moment. *Theoretical and applied climatology*, 96(3-4), pp.337-344.