STATISTICS OF EXTREMES WITH APPLICATIONS TO EXTREME FLOOD HEIGHTS IN THE LOWER LIMPOPO RIVER BASIN OF MOZAMBIQUE

by

DANIEL MAPOSA

THESIS

Submitted in fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

STATISTICS

in the

FACULTY OF SCIENCE AND AGRICULTURE
(School of Mathematical and Computer Sciences)

at the

UNIVERSITY OF LIMPOPO

PROMOTER: Prof. JJ Cochran (University of Alabama, USA)

CO-PROMOTER: Prof. M Lesaoana

2016

Declaration

I, **Daniel Maposa**, declare that the thesis hereby submitted to the University of Limpopo, for the degree of Doctor of Philosophy in Statistics has not been submitted by me for a degree at this or any university; that it is my work in design and in execution, and that all material contained herein has been duly acknowledged.

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

Maposa, D. (DR)

Abstract

Statistics of extremes has seen much growth both in theory and application since its early theoretical developments almost a century ago in the 1920s and its first major applications to real-life problems pioneered by Emil Gumbel in the early 1940s. Although the theory and applications of extreme value theory (EVT) have been extensively advanced and utilised in most developed countries, in terms of applications little has been done in many developing countries in Africa despite the abundance of areas of applications and raw data in some of these countries. In hydrology, the choice of flood frequency probability distributions for a particular site or region remains the subject of ongoing research. The work contained in this thesis is a contribution towards this area and it addresses this problem in one of the developing and economically challenged countries in Africa, Mozambique, in the lower Limpopo River basin (LLRB). The LLRB is a basin characterised by extreme natural hazards, alternating between extreme floods and severe droughts.

This thesis is based on an extensive application of EVT to extreme flood heights data in the LLRB of Mozambique at three sites: Chokwe, Combomune and Sicacate hydrometric stations. Two fundamental approaches of EVT, block maxima and peaks-over-threshold (POT), are used in this thesis. Recent theoretical results by Ferreira and de Haan (2015) have shown that despite its inefficiency due to data lost as a result of blocking, the block maxima approach is more efficient in a number of situations than the POT approach, and the two approaches are quite comparable for large sample sizes. A number of

candidate distributions are investigated for their goodness-of-fit to the annual daily maximum flood heights in a block maxima realisation at each site. The findings reveal that the GEV distribution is the most appropriate distribution to apply in the LLRB and the distribution can be recommended as the likelihood function for regional and spatial extremes flood frequency analysis in the basin. The thesis addresses the issue of cumulative effects on daily flood heights through a comparative analysis of six annual maxima moving sums. The findings demonstrate that the six annual maxima time series models are not significantly different based on the characteristics considered in this thesis.

In an attempt to reduce uncertainties in the estimates, a Bayesian Markov chain Monte Carlo (MCMC) approach with a conjugate prior and a GEV likelihood function is used to model the tails of the extreme flood heights in the basin. The findings reveal that the addition of prior information in Bayesian MCMC substantially reduces uncertainties in the estimates and improves precision in the predicted extreme floods. The r largest order statistics models developed in this thesis are generally promising and the standard errors of the estimates of the parameters are substantially reduced. In order to account for climate change impact, nonstationary models are considered with the long-term trend and seasonal oscillation index (SOI) (a meteorological variable indicator) as covariates of the parameters of the GEV distribution and the generalised Pareto distribution (GPD).

Among the major contributions of this thesis is a proposed procedure for the determination of the 8 days window period used in extracting independent r largest order values within the same year for the r largest order statistics approach. A summary of the key findings and contributions of this thesis are given in Chapter 9. Moreover, contributions by the study topic in each chapter are given at the end of each chapter.

Dedication To

My mother Mutsengwa and late father July

My brother Enock Maposa

My daughters:
Nyasha Maposa and Anesu Maposa

Acknowledgments

To begin with, I would like to express my gratitude to my academic promoter Professor James J. Cochran (University of Alabama) and co-promoter Professor 'Maseka Lesaoana (University of Limpopo) for their patience, inspiration and guidance throughout my PhD studies and co-authoring papers based on this thesis. I am very grateful to my academic 'brother' Dr. Caston Sigauke, for journeying with me through the world of statistics of extremes and Bayesian statistics. I would also like to thank various reviewers who reviewed several papers from this thesis which were meant for publication in peer-reviewed journals for their valuable comments. Indeed the review comments broadened my thinking horizon in the world of statistics of extremes and have greatly improved the quality of this thesis.

I am very grateful to the Mozambique National Directorate of Water (DNA), for providing data used for this thesis. Special thanks go to Mr Isac Filimone of DNA who went all his way to provide me with all the necessary data used in this thesis. I am also thankful to the United Nations Office for the Coordination of Humanitarian Affairs-Southern Africa (UN OCHA) for providing me with weekly updates reports of floods in Southern Africa, particularly for the Limpopo River basin in Mozambique. I am also greatly indebted to the DST-NRF Centre of Excellence in Mathematical and Statistical Sciences (CoE-MaSS) of South Africa who provided funds for the postgraduate studies for the year 2014. I would also like to express my gratitude to the World Bank and Statistics South Africa for funding my trip to attend the World Statis-

tics Congress (WSC2015) organised by the International Statistical Institute (ISI). I am also greatly indebted to Monash South Africa for funding my trips to attend the Extreme Value Analysis (EVA2013) Conference in Shanghai, China, 2013 and the International Disaster and Risk Conference (IDRC2014), in Davos, Switzerland, 2014. University of Limpopo, I thank you for your continuous support towards my PhD studies.

I would like to thank my UL fellow PhD students and colleagues TB Darikwa, Alex Boateng and the Department of Statistics and Operations Research as a whole for their support and encouragement. I would also like to thank my colleagues at Monash South Africa, Prof. Ilse Niemann-Struweg, Prof. HB Klopper, Prof. Eita (now at UJ), Dr. Kwame (now at UJ), my brothers-in-chief Mazanai Musara and Brian Sibanda, not forgetting my sisters-in-chief Mpho Majola, Jennifer and Lindo. Ilse and HB thank you so much for supporting me to attend international conferences and believing in me throughout the years I spent with you at Monash University. To the Almighty Lord, I am very grateful for the gift of life, good health and wisdom. To my wonderful and exquisite wife, Nzokuhle, and lovely children Nyasha and Anesu, I thank you for your patience and support and I can declare that I am finally available at home again. To Mavy, I am grateful to you for your understanding and maturity during the testing times of my study duration. Last, but not least, I am greatly indebted to my elder brother Enock Julai Maposa for always believing in me and always encouraging me to continue with education to the PhD level, I dedicate this thesis to him.

Contents

List of Figures

List of Tables

List of Special Symbols and Special Abbreviations

$G_\xi(x)$	extreme value distribution function for flood height maxima
$G(\mu(t), \sigma(t), \xi(t))$	nonstationary generalised extreme value distribution
$H_\xi(x)$	generalized Pareto distribution function
$H(\sigma(t), \xi(t))$	nonstationary generalised Pareto distribution
x_p	quantile function
D	deviance statistic
$P(\theta)$	prior distribution
$P(x\|\theta)$	likelihood function
$P(\theta\|x)$	posterior distribution
$P(x_{n+1}\|x)$	posterior predictive density of a future observation x_{n+1}
u	a sufficiently high threshold
log	natural logarithm
ln	natural logarithm
$\mathcal{D}(G_\xi(x))$	domain of attraction of $G_\xi(x)$
ξ	extreme value index (EVI)
F_u	empirical exceedance distribution function of u
x_F	right end point of distribution function F
$a.s.$	almost surely
iid	independent and identically distributed
pdf	probability density function

Distributions

Generalised extreme value family of max-stable distributions

Gumbel distribution

Distribution function

$$G(x) = \exp\left\{-\exp\left[-\frac{x-\mu}{\sigma}\right]\right\}, \xi = 0$$

Fréchet distribution

Distribution function

$$G_\xi(x) = \exp\left\{-\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\}, \xi > 0$$

Weibull distribution

Distribution function

$$G_\xi(x) = \exp\left\{-\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\}, \xi < 0$$

Generalised Extreme Value Distribution, $\xi \neq 0$

Notation	$X \sim \text{GEV}(\mu, \sigma, \xi)$
Distribution function	$G_\xi(x) = \exp\left\{-\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\}$
Density function	$g_\xi(x) = \frac{1}{\sigma}\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}-1}\exp\left\{-\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\}$
Range	$\left\{x\|1+\xi\left(\frac{x-\mu}{\sigma}\right)>0\right\}$
Parameters	$\xi \neq 0, -\infty < \mu < \infty, \sigma > 0$
EVI	ξ

Generalised Extreme Value Distribution, $\xi = 0$

Notation	$X \sim \text{GEV}(\mu, \sigma, 0)$
Distribution function	$G_\xi(x) = \exp\left\{-\exp\left(-\frac{x-\mu}{\sigma}\right)\right\}$
Density function	$g_\xi(x) = \frac{1}{\sigma}\exp\left(-\frac{x-\mu}{\sigma}\right)\exp\left\{-\exp\left(-\frac{x-\mu}{\sigma}\right)\right\}$
Range	$-\infty < x < \infty$
Parameters	$-\infty < \mu < \infty, \sigma > 0$
EVI	0

Generalised Pareto Distribution

Notation $Y \sim \text{GPD}(\sigma, \xi)$

Distribution function

$$H_\xi(y) = P\left(X \leq u + y | X \geq u\right) = \begin{cases} 1 - \left(1 + \left(\xi \frac{y}{\sigma}\right)\right)_+^{-\frac{1}{\xi}}, & \text{for } \xi \neq 0, \;\; 0 \leq y \leq x_F - u, \\ 1 - \exp\left(-\frac{y}{\sigma}\right), & \text{for } \xi = 0, \;\; 0 \leq y \leq x_F - u, \;\; \sigma > 0. \end{cases}$$

Gamma Distribution with two parameters

$F(x) = \frac{\Gamma_{x/\beta}(\alpha)}{\Gamma(\alpha)}$

Gamma Distribution with three parameters

$F(x) = \frac{\Gamma_{(x-\gamma)/\beta}(\alpha)}{\Gamma(\alpha)}$

Generalised Gamma Distribution with three parameters

$F(x) = \frac{\Gamma_{(x/\beta)^\kappa}(\alpha)}{\Gamma(\alpha)}$

Log-Pearson Type 3 Distribution

$F(x) = \frac{\Gamma_{(\ln(x)-\gamma)/\beta}(\alpha)}{\Gamma(\alpha)}$

Lognormal Distribution with two parameters

$$F(x) = \Phi\left(\frac{\ln(x)-\mu}{\sigma}\right)$$

Lognormal Distribution with three parameters

$$F(x) = \Phi\left(\frac{\ln(x-\gamma)-\mu}{\sigma}\right)$$

List of Abbreviations and Acronyms

ACF	Autocorrelation Function
A-D	Anderson-Darling
ARA Sul	Regional Water Administration for the South in Mozambique
AM1	Annual Daily Maximum
AM2	Annual 2-Day Maximum
AM5	Annual 5-Day Maximum
AM7	Annual 7-Day Maximum
AM10	Annual 10-Day Maximum
AM30	Annual 30-Day Maximum
AMDS	Annual Maximum Daily Series
BBC	British Broadcasting Corporation
CAPA	Catchment parameter method
CDF	Cumulative Distribution Function
CII	Chartered Insurance Institute
CNN	Cable News Network
DFID	British Department for International Development
DNA	Mozambique's National Directorate of Water
E31	Code for Pafuri Hydrometric Station
E33	Code for Combomune Hydrometric Station
E35	Code for Chokwe Hydrometric Station

E36	Code for Sicacate Hydrometric Station
ENHANS	Extreme Natural Hazards and Societal Implications
ENSO	El Niño Southern Oascillation
EVT	Extreme Value Theory
FFA	Flood Frequency Analysis
FFEWS	Flood Forecasting and Early Warning System
Ga2	Two-parameter Gamma Distribution
Ga3	Three-parameter Gamma Distribution
GEV	Generalised Extreme Value
GIZ	Germany's Deutsche Gesellschaft f*ü*r Internationale Zusammenarbeit
GLO	Generalised Logistic Distribution
GLTP	Great Limpopo Transfrontier Park
GN	Generalised Normal Distribution
GoF	Goodness-of-Fit
GPD	Generalised Pareto Distribution
GeoSSFM	Geo-spatial Streamflow Forecasting Modelling
GML	Generalised Maximum Likelihood
HFA	Hyogo Framework for Action
IDRC	International Disaster and Risk Conference
IFRC	International Federation of Red Cross
IFD	Intensity-Frequency-Duration
INGC	Mozambique's National Institute for Disaster Management
IPCC	Intergovernmental Panel on Climate Change
K-S	Kolmogorov-Smirnov
LGa3	Three-parameter Log-Gamma Distribution
LIMCOM	Limpopo Watercourse Commission
LLRB	Lower Limpopo River Basin
LN2	Two-parameter Log-Normal Distribution
LN3	Three-parameter Log-Normal Distribution

LP3	Log-Pearson Type 3
LRB	Limpopo River Basin
LSE	Least Squares Estimators
MAD	Mean Absolute Difference
MADI	Mean Absolute Deviation Index
MAP	Mean Annual Precipitation
MCMC	Markov Chain Monte Carlo
MDGs	Millennium Development Goals
MDI	Maximal data information
ML	Maximum likelhood
MLE	Maximum Likelihood Estimator
MOM	Method of Moments
NWSRFS	National Weather Services' River Forecasting System
NWSRFS	National Weather Services' River Forecasting System
OCHA ROSA	(UN) Organisation for the Coordination of Humanitarian Affairs - Regional Office for Southern Africa
P3	Pearson Type 3
POT	Peaks-Over-Threshold
PPCC	Probability Plot Correlation Coefficient
PWM	Probability Weighted Moments
REFSSA	Regional Estimation of Extreme Flood Peaks by Selective Statistical Analys
RMSE	Root Mean Square Error
RRMSE	Relative Root Mean Square Error
SADC	Southern African Development Community
RMF	Regional maximum flood
SDGs	Sustainable Development Goals
SOI	Seasonal Oscillation Index
UK	United Kingdom

UN	United Nations
UNCHS	United Nations Centre for Human settlements
UNDP	United Nations Development Programme
UNEP	United Nations Environmental Programme
UNESCO	United Nations Educational, Scientific and Cultural Organisation
US$	United States Dollars
USA	United States of America
USAID	United States Agency for International Development
WMO	World Meteorological Organisation

Research Outputs

The following sections give a list of research outputs from this thesis.

Peer Reviewed Journal Publications

1. Maposa, D., Cochran, JJ. and Lesaoana, M. (2016).
   Modelling extreme flood heights in the lower Limpopo River basin of Mozambique using a time-heterogeneous generalised Pareto distribution. To appear soon in: *Statistics and Its Interface*, **9**(4).

2. Maposa, D., Cochran, JJ. and Lesaoana, M. (2016).
   Modelling nonstationary annual maximum flood heights in the lower Limpopo River basin of Mozambique. To appear soon in: *Jàmbá: Journal of Disaster Risk Studies*, 8(1), Art#185. http://dx.doi.org/10.4102/jamba.v8i1.185.

3. Maposa, D., Cochran, JJ. and Lesaoana, M. (2016). Comparative analysis of annual maxima time series models along the lower Limpopo River basin of Mozambique. In: Awotona, A. (Ed.), (2016). Planning for Community-based Disaster Resilience Worldwide: Learning from Case Studies in Six Continents. To appear soon in: *Routledge, Taylor & Francis Group*, p.115-144.

4. Maposa, D., Cochran, JJ. and Lesaoana, M. (2015). Construction of flood frequency curves in the lower Limpopo River basin of Mozambique using Bayesian and Markov chain Monte Carlo methods. *Proceedings of*

the 60th International Statistical Institute (ISI) World Statistics Congress, 26-31 July 2015, Rio de Janeiro, Brazil, ISBN: 978-90-73592-35-3 *C* ISI 2015, www.isi-web.org.

5. Maposa, D., Cochran, JJ. and Lesaoana, M. (2015). Fighting flooding in Mozambique. *ORMS Today*, **42**(2), April 2015 Edition.

6. Maposa, D., Cochran, JJ., Lesaoana, M. and Sigauke, C. (2014). Estimating high quantiles of extreme flood heights in the lower Limpopo River basin of Mozambique using model based Bayesian approach. *Natural Hazards and Earth System Sciences, Discussion*, **2** (8): 5401-5425, doi:10.5194/nhessd-2-5401-2014.

7. Maposa, D., Cochran, JJ. and Lesaoana, M. (2014). Investigating the goodness-of-fit of ten candidate distributions and estimating high quantiles of extreme floods in the lower Limpopo River Basin, Mozambique. *Journal of Statistics and Management Systems*, **17**(3): 265-283, doi:10.1080/09720510.2014.927602.

Extended Abstracts in Conference Proceedings

1. Maposa, D., Cochran JJ. and Lesaoana, M. (2014). Estimating high quantiles of extreme flood heights in the lower Limpopo River basin of Mozambique using model based Bayesian approach. Proceedings of the International Disaster and Risk Conference (IDRC Davos 2014), 24-28 August 2014, Extended Abstracts, p.435-438, **Davos, Switzerland**. Available online at: The IDRC Davos 2014 Conference Proceedings `http://idrc.info/outcomes/conference-proceedings/`

2. Maposa, D., Cochran, JJ., Lesaoana, M. and Sigauke, C. (2014). Estimating high quantiles of extreme flood heights in the lower Limpopo River basin of Mozambique using model based Bayesian approach. Faculty of

Science and Agriculture Postgraduate Research Conference (FSA-PGD 2013), 2-3 October 2014, Extended Abstracts, Bolivia Lodge, **Polokwane, South Africa**.

3. Maposa, D., Cochran, JJ. and Lesaoana, M. (2013). Investigating the goodness-of-fit of ten candidate distributions and estimating high quantiles of extreme flood heights in the lower Limpopo River Basin, Mozambique. Faculty of Science and Agriculture Postgraduate Research Conference (FSA-PGD 2013), 3-4 October 2013, Extended Abstracts, Bolivia Lodge, **Polokwane, South Africa**.

4. Maposa,D., Cochran, JJ. and Lesaoana, M. (2012). Fitting the distributions and estimating the return period of extreme floods in Mozambique, Limpopo River Basin. Faculty of Science and Agriculture Postgraduate Research Conference (FSA-PGD 2013), 4-5 October 2012, Extended Abstracts, Bolivia Lodge, **Polokwane, South Africa**.

International Conferences

1. Maposa, D., Cochran, JJ. and Lesaoana, M. (2015). 60^{th} International Statistical Institute (ISI) World Statistics Congress (WSC 2015). Construction of flood frequency curves in the lower Limpopo River basin of Mozambique using Bayesian and Markov chain Monte Carlo methods, 26-31 July 2015, **Rio de Janeiro, Brazil**.

2. Maposa, D. Cochran, JJ. and Lesaoana, M. (2014). International Disaster and Risk Conference [IDRC2014], (2014). Estimating high quantiles of extreme flood heights in the lower Limpopo River basin of Mozambique using model based Bayesian approach, 24-28 August 2014. Congress Centre, **Davos, Switzerland**. Available online at:
   The IDRC Davos 2014 Conference Proceedings (http://idrc.info/

outcomes/conference-proceedings/

3. Maposa, D., Cochran, JJ. and Lesaoana, M. (2013). Extreme Value Analysis [EVA2013] Conference, (2013). Investigating the goodness-of-fit of ten candidate distributions and estimating high quantiles of extreme floods in the lower Limpopo River Basin, Mozambique, 8–12 July 2013. Fudan University, **Shanghai, China**. Available online at: http://people.math.gatech.edu/~peng/EVAFuDan/doc/EVA2013-ProgramAbstract.pdf , pp.60.

Other Conferences

1. Maposa, D., Cochran, JJ. and Lesaoana, M. (2015). Modelling nonstationary annual maximum flood heights in the lower Limpopo River basin of Mozambique. South African Statistical Association (SASA 2015), 29 November - 2 December 2015, University of Pretoria, South Africa.

2. Maposa, D., Cochran, JJ. and Lesaoana, M. (2015). Modelling extreme flood heights in the lower Limpopo River basin of Mozambique using a time-heterogeneous generalised Pareto distribution. Faculty of Science and Agriculture Postgraduate Research Conference (FSA-PGD 2015), 1-2 October 2015, Bolivia Lodge, Polokwane, South Africa.

3. Maposa, D., Cochran, JJ. and Lesaoana, M. (2014). Estimating high quantiles of extreme flood heights in the lower Limpopo River basin of Mozambique using model based Bayesian approach. Faculty of Science and Agriculture Postgraduate Research Conference (FSA-PGD 2013), 2-3 October 2014, Bolivia Lodge, Polokwane, South Africa.

4. Maposa, D., Cochran, JJ. and Lesaoana, M. (2013). Investigating the goodness-of-fit of ten candidate distributions and estimating high quantiles of extreme flood heights in the lower Limpopo River Basin, Mozam-

bique. South African Statistical Association (SASA 2013), 4-8 November 2013, The Ranch Hotel, Polokwane, South Africa.

5. Maposa, D., Cochran, JJ. and Lesaoana, M. (2013). Investigating the goodness-of-fit of ten candidate distributions and estimating high quantiles of extreme flood heights in the lower Limpopo River Basin, Mozambique. Monash University (South Africa Campus), 31 October 2013, Monash South Africa, Johannesburg, South Africa.

6. Maposa, D., Cochran, JJ and Lesaoana, M. (2013). Investigating the goodness-of-fit of ten candidate distributions and estimating high quantiles of extreme flood heights in the lower Limpopo River Basin, Mozambique. Faculty of Science and Agriculture Postgraduate Research Conference (FSA-PGD 2013), 3-4 October 2013, Bolivia Lodge, Polokwane, South Africa.

7. Maposa, D., Cochran, JJ. and Lesaoana, M. (2012). Fitting the distributions and estimating the return period of extreme floods in Mozambique, Limpopo River Basin. Faculty of Science and Agriculture Postgraduate Research Conference (FSA-PGD 2013), 4-5 October 2012, Bolivia Lodge, Polokwane, South Africa.

Chapter 1

Introduction and background

1.1 Introduction

The beginning of the 21^{st} century has been marked by an unusual number of natural disasters worldwide (CNN, 2015; Wikipedia, 2015; Mondlane et al., 2013; MunichRe, 2013; WMO, 2013; Kron et al., 2012; Maree, 2011; MunichRe, 2011; Cuamba and Maúre, 2008; BBC, 2005; UN, 2005; Wisner et al., 2004). Recent catastrophic events such as the devastating earthquakes in Nepal in 2015 (the most severe being the Gorkha earthquake) which also affected parts of China, India and Bangladesh; severe earthquakes and devastating tsunamis in Japan in 2011 and the Indonesian islands of Sumatra and Andaman in 2004; earthquakes in Haiti in 2010, China in 2008 and Pakistan in 2005; hurricane Katrina in USA in 2005; flooding in Mozambique in 2013 and 2000, South Africa in 2011 and Pakistan in 2010; and flooding and landslides in Brazil in 2011, remind us that a holistic approach is needed to understand such phenomena, predict such catastrophic events, and mitigate the impact of natural disasters.

A natural disaster may pose an intolerable threat to society. Unfortunately there is currently no way to describe the origin and complex dynamics of natural disasters by a constitutive set of fundamental equations. The most damaging and least understood in the realm of natural disasters are the so-called extreme events. In different contexts extreme events are called critical transitions, disasters, or catastrophes. Typical examples of extreme events are severe earthquakes, floods, landslides, droughts, tsunamis, heat waves, etc. Although extreme natural events are rare, they are important to understand because they have the potential to inflict great damages to populations, economies, and the environments such as loss of life, partial or total loss of infrastructure, and large environmental damage like those caused by catastrophic events mentioned in the previous paragraph. For instance, the recent 25 April 2015 Nepal earthquake whose epicentre was in Gorkha district near the city of Kathmandu and the 12 May 2015 follow-up Nepal earthquake whose epicentre was near the Chinese border between the capital of Kathmandu and Mount Everest posed an intolerable threat to society. These two recent earthquakes in Nepal killed a combined total of more than 8,000 people and injured more than 19,000 people in less than a month, and the 25 April 2015 earthquake is reported to be the worst natural disaster to strike Nepal since the Nepal-Bihar earthquake in 1934 (CNN, 2015; Wikipedia, 2015).

According to Mondlane et al. (2013) and MunichRe (2013) the severe earthquake in Haiti in 2010 was the most deadly disaster since 1900 and the Japanese 2011 giant earthquakes and devastating tsunamis were both the most expensive and most complex disasters ever to happen, resulting in large scale economic, social, asserts losses and loss of lives. These disasters are clear evidence that, despite modern technology at our disposal and advanced research, human beings continue to be exposed to natural disasters.

Despite the shortage of quality or sufficient amount of data to use in the analysis of trends in extreme events, data collected since 1950 indicate clear evidence of changes in some extreme events (MunichRe, 2013). According to Mondlane et al. (2013) extreme events will intensify in the future and continue to endanger food security and human lives. Along the same lines of argument, Dr. Walter J. Ammann, the Conference Chair of the International Disaster and Risk Conference (IDRC) held in Davos, Switzerland, on 24-28 August 2014, asserted that the scope, intensity, complexity of risks and natural disasters such as floods, earthquakes and forest fires are on the rise in these recent years (Stal et al., 2014; WMO, 2013). In Arya et al. (2014), the Director-General of UNESCO, Irina Bokova, stated that: "Every year, more than 200 million people are affected by natural hazards, and the risks are increasing – especially in developing countries, where a single major disaster can set back healthy economic growth for years. As a result, approximately one trillion dollars have been lost in the last decade alone." In line with these views, the 2014 IDRC held in Davos, advocated for collaborative efforts in disaster risk management and reduction (Stal et al., 2014).

Mondlane et al. (2013) argues that rainfall and water patterns are the main environmental concerns in Africa. On a worldwide scale floods and droughts account for approximately 90 percent (i.e. vast majority) of all people affected by natural disasters (Smakhtin, 2014). This is certainly true in the Limpopo River basin (LRB) in Southern Africa, particularly Mozambique, where several efforts in disaster risk management and reduction are underway in the LRB region to mitigate the devastating impact of persistent flooding. The lower Limpopo River basin (LLRB) of Mozambique is characterised by extreme natural hazards alternating between extreme floods and severe droughts. The primary focus of this research is extreme floods in Mozambique along the LLRB.

Extreme floods or extreme events in general are usually of interest because of their tremendous impact on humans and the economy (Cooley, 2005). It is for this reason that the discipline of statistics is used extensively to characterise the behaviour of extreme events (Blain and Meschiatti, 2014; de Haan and Ferreira, 2006; Cooley, 2005; Coles and Pericchi, 2003). The branch of statistics that is concerned with the distribution of data of unusually low or high values in the tail of a probability distribution is called extreme value theory (EVT), (Reiss and Thomas, 2007; Beirlant et al., 2004; Coles, 2001). In EVT one attempts to best estimate the probability of unusual events (Khuluse et al., 2009; Cooley, 2005). Disciplines to which EVT has been applied include, but are not limited to, medicine, hydrology, meteorology, finance, environmental studies, and economics. For example, in hydrology EVT has been found to be useful in fitting models to historical data to estimate the return periods of extreme floods (Blain and Meschiatti, 2014; Ahammed and Hewa, 2012; Atroosh and Mustafa, 2012; Khuluse et al., 2009; Nguyen, 2009; van den Brink and Können, 2009; Coles et al., 2003; Coles, 2001; Gumbel, 1958). In finance EVT has been used extensively in insurance and financial risk analysis.

1.2 Research problem

1.2.1 Background of the Limpopo River basin of Mozambique

Mozambique is located on the Indian Ocean coast of Southern Africa sharing borders with Tanzania in the north, Malawi, Zambia and Zimbabwe in the west, and South Africa and Swaziland in the south. It is one of the developing countries in Southern Africa. Agricultural and other economic activities such as fishing in the LLRB form the backbone of the Mozambican economy. Mozambique is a downstream country through which nine transboundary

rivers which have their origins in neighbouring countries (including the great Zambezi River) pass.

The Limpopo River rises at the confluence of Marico and Crocodile rivers at the limit of North West Province in South Africa and flows northeast along the border of South Africa with Botswana. The name Limpopo is a modified version of the original local Sepedi name *diphororo tsâ meetse* meaning 'gushing strong waterfalls' (Chilundo et al., 2008). The river separates South Africa from Botswana and Zimbabwe by creating an international border of nearly 900 km, and then flows eastwards through Mozambique to the Indian Ocean. The river enters Mozambique at Pafuri, a point where the three countries Mozambique, South Africa and Zimbabwe meet and flows for the next 561 km within Mozambique before emptying into the Indian Ocean at Zongoene approximately 60 km downstream of the town of Xai-Xai (WMO, 2012; Hakala and Pekonen, 2008). The LRB significantly narrows in the coastal area where the river course meanders for nearly 70 km through its lower valley, from Xai-Xai town to the ocean and forms a circular alluvial valley inland of about 15 km in diameter as it enters the ocean at Zongoene (Hakala and Pekonen, 2008).

Among the African rivers that drain into the Indian Ocean, the Limpopo River is the second largest in length to the Zambezi River which hosts Mozambique's largest hydropower station at Cahora Bassa Dam (WMO, 2012). The floodplains of the Limpopo River are fertile and heavily populated.

The LRB catchment area covers more than 412,500 km^2 and is drained by the Limpopo River and its tributaries including the Olifants (Elefantes in Portuguese or Elephants in English) and Changane rivers which drain the largest areas. The total length of the main Limpopo River is about 1,750 km and is located between latitudes 22S – 26S and longitudes 26E – 35E (WMO, 2012;

Hakala and Pekonen, 2008). Its catchment area distribution among Botswana, Mozambique, South Africa and Zimbabwe is $20\%, 20\%, 45\%$ and 15% respectively, with most of the catchment lying under semi-arid conditions (WMO 2012).

The LRB is home to about 14 million people in the riparian states of Botswana, Mozambique, South Africa and Zimbabwe (WMO, 2012; Chilundo et al., 2008). The urban centres throughout the SADC region such as Gaborone and Francistown of Botswana, Pretoria, parts of Johannesburg, and Polokwane all of South Africa, Beitbridge, Bulawayo and Gwanda of Zimbabwe, Chokwe and Xai-Xai of Mozambique are the major users of the basin's water resources, supplying industries, power stations and municipalities. In rural areas, the basin's water is primarily used for domestic purposes, irrigation, livestock and watering (WMO, 2012; Hakala and Pekonen, 2008).

The LRB comprises several tributaries and sub-basins. Its largest tributary, the Crocodile River, originates from the Witwatersrand of South Africa in Johannesburg and drains a catchment area of 29,000 km^2. The Olifants and its tributaries form the largest sub-basin of the LRB, covering a catchment area of 79,000 km^2, with about 84% of the catchment area located in South Africa and bring the greatest amount of water to the Limpopo River (Mohamed, 2014; Hakala and Pekonen, 2008). The other large sub-basin is the Changane River, which is characterised by a very low run-off and exists entirely within Mozambique covering a catchment area of 43,000 km^2 (Chilundo et al., 2008; Hakala and Pekonen, 2008). The largest dam in the LRB of Mozambique, Massingir Dam, is built on the Olifants River near its mouth to the Limpopo River.

The famous Great Limpopo Transfrontier Park (GLTP) and Conservation Area is situated in the LRB in the area around Pafuri where the three countries

Mozambique, South Africa and Zimbabwe meet. The GLTP covers an area of 35,000 km^2, which is an area about the size of The Netherlands, and it comprises Mozambique's Limpopo National Park, South Africa's Kruger National Park, and Zimbabwe's Gonarezhou National Park.

The LRB is divided into three main geographic sections namely: the Upper Limpopo, the Middle Limpopo and the Lower Limpopo basins. The Upper Limpopo basin starts from the headwaters down to the Shashe River confluence at the Botswana, South Africa and Zimbabwe borders. The Middle Limpopo basin is located between the Shashe River confluence and Pafuri at the joint border between Mozambique, South Africa and Zimbabwe. The Lower Limpopo flows downstream from Pafuri to the mouth of the river onto the Indian Ocean (Chilundo et al., 2008; Hakala and Pekonen, 2008). The result of this division of the LRB means that the Lower Limpopo section of the basin lies entirely in Mozambique. The research in this thesis is centred on the Lower Limpopo section of the basin known as the LLRB.

According to Hakala and Pekonen (2008), it is postulated that the Mozambican part of the Limpopo River basin, that is, the Lower Limpopo, consists of three major climatic zones:

- The Coastal Xai-Xai Zone, which is situated in the lower portion of the region between Xai-Xai and Limpopo River's connection to the sea at Zongoene.
- The Lower Limpopo Valley, which is situated in the area between Xai-Xai and Macarretane Dam. This climatic zone includes the Chokwe area.
- The Upper Limpopo Valley, which is the area from Macarretane Dam in the main Limpopo River up to the Mozambican border at Pafuri.

The LRB suffers from extreme climate conditions, punctuated with frequent

droughts and floods. The mean annual rainfall in the basin varies considerably from 200 mm to 1,500 mm, while annual rainfall is highly seasonal with more than 95% received between October and April with peak mean monthly totals in February (WMO, 2012; Chilundo et al., 2008). The Limpopo River is characterised by very low (to negligible) flows during the dry season, resulting in severe droughts, while extreme flooding is common during the wet season. The river is not always perennial, in some years it can go dry for several months. It is estimated that only 10% of the measured flow at Chokwe is generated in the LLRB, the majority of the flow is generated beyond the borders of Mozambique particularly South Africa (WMO, 2012; Hakala and Pekonen, 2008).

The climate of the LRB varies from being arid in the west, semi-arid and temperate in central zones, and semi-arid in the east with a few sub-humid pockets in the centre (WMO, 2012; Chilundo et al., 2008; Hakala and Pekonen, 2008). The Mozambican coast is influenced by the Mozambique current that flows southward bringing warm water and humid air from the Equator and produces a humid, warm climate. It is generally warm in summer and mild in winter. Similar to annual rainfall patterns, annual air temperatures in the basin are seasonal, with highest temperatures recorded during early summer months and lowest temperatures during the cool and dry winter months (WMO, 2012; Hakala and Pekonen, 2008).

1.2.2 Benefits and drawbacks of the Limpopo River

Among the benefits of the Limpopo River, it supplies water to the biggest irrigation system in Mozambique, Chokwe Irrigation Scheme located in Chokwe district in the LLRB, through Massingir Dam in the Olifants tributary (Chilundo et al., 2008; Hakala and Pekonen, 2008). The supply of water by the Limpopo River to the irrigation system helps improve the standard of living for both the rural and urban communities throughout the region. The LLRB and the asso-

ciated Chokwe Irrigation Scheme do not only supply the much needed water for the communities and industry, but also help alleviate poverty in the region and thus contribute to one of the millennium development goals (MDGs) of poverty reduction in developing countries (UNDP Mozambique 2010).

Since the end of the Mozambican Civil War in 1992 after nearly two decades of conflict, the LLRB has attracted a wide range of researchers from diversified fields who have studied the region's climate and river system in an effort to better predict and therefore mitigate the damaging impact of the drought-flood cycle. These researchers include, (but are not limited to) non-governmental organisations, universities and independent researchers. In addition, donors such as the World Bank, British Department for International Development (DFID), United States Agency for International Development (USAID), Germany's Deutsche Gesellschaft f*ü*r Internationale Zusammenarbeit (GIZ) and African Development Bank provide financial and technical assistance in disaster risk management in the basin. These organisations are important stakeholders in the region and they work in collaboration with the Limpopo Watercourse Commission (LIMCOM) and the government of Mozambique (WMO, 2012).

The low-lying nature of the LLRB region across the coastal floodplain makes it susceptible to flooding during periods of high flows, posing a major problem in the lower part of the basin. The high incidence of flooding is mainly attributed to tropical cyclones that form in the Indian Ocean coast, as well as the fact that Mozambique is a downstream country through which all floods from neighbouring countries pass. According to historical records on natural disasters over the past 58-year period from 1956 through to 2013, Mozambique experienced about 11 droughts, 23 floods, 14 tropical cyclones, 18 epidemics and one earthquake (Musiya, 2013; WMO, 2012; Chilundo et al., 2008; Hakala

and Pekonen, 2008). In the most recent years 2014 and 2015, there were no major reported incidents of severe droughts, extreme floods and other natural disasters in Mozambique.

The most catastrophic and expensive of these reported natural disasters in the LLRB of Mozambique were the floods of the year 2000, in February and March, which caused the Limpopo River to reach levels never previously recorded. The river swelled from less than 100 m wide to between 10 km and 20 km wide for more than 100 km stretch and it inundated more than 1,400 km^2 of farmland (WMO, 2012; Jäger et al., 2009; Manuel and Vicente, 2002). This disaster, which was primarily a result of heavy rainfall brought by tropical cyclones Eline and Gloria, flooded the town of Chokwe and Xai-Xai, killed more than 700 people and caused economic damages estimated at US$500 million in the LLRB of Mozambique alone. It was also reported that one woman gave birth on a tree top in the year 2000 (WMO, 2012; Jäger et al., 2009; Manuel and Vicente, 2002).

The Limpopo River has a history of worst floods and droughts than all other national and international rivers in Mozambique. For example, in the recent past the river experienced extreme floods in 2013, 2012, 2010, 2000, 1999, 1997, 1985, 1977, 1975, 1972, and 1967 among other years (Jackson, 2013a,b; WMO, 2012). Each of these devastating floods in the LRB was attributed to cyclone activities and heavy rainfall events (WMO, 2012). During the recent 2013 flood it was reported that two women gave birth on rooftops in Chokwe district and over 100 people were killed (Jackson, 2013a,b; Musiya, 2013). The flood that occurred in February 1977 in the river was reported to be the worst that occurred in Mozambique until that date, but it was surpassed by the year 2000 flood which reached a maximum height of 13 m in Chokwe district, completely flooding the town of Chokwe and parts of Xai-Xai city in Mozambique and

shared water with Incomati River which had never happened before (WMO, 2012; Manuel and Vicente, 2002). The flood warning level at Chokwe is 4 m, which means that the 13 m flood height of the year 2000 was just over three times above the flood alert level (WMO, 2012). The river also experienced its worst droughts in the recent past in 2003, 2002, 1995, 1994, 1992, 1991, 1987, 1984, 1983, 1981, and 1980 (WMO, 2012).

1.2.3 Historical background of extreme value theory

According to Gumbel (1958) the informal history of EVT or statistics of extremes may be traced back to as far as the work of Nicholas Bernoulli in 1709 who discussed the mean largest distance from the origin, O, given a number of random points, n, that are lying on a straight line of fixed length, l. Without loss of generality, Rajaram (2006) argues that the first work to describe an application of EVT to flood flows was done in 1914 by Fuller (1914) and this was followed by the work of Griffith (1920) who studied the phenomena of rupture and flow in fluids in 1920, and brought out an application of EVT in the discussion. The work that may have stimulated a gradual development in EVT was written by von Bortkiewicz (1922) and initially published in the German language. According to Rajaram (2006), von Bortkiewicz's work introduced the concept of distribution of largest values for the first time in the history of EVT and, in particular, the work dealt with the distribution of range from normality (the normal distribution). In 1923 two respected mathematical statisticians Dodd (1923) and von Mises (1923) extended the work introduced by von Bortkiewicz (1922), in particular the distribution in his work, by evaluating its expected value (von Mises, 1923). Dodd (1923) evaluated the median of the distribution of large values introduced by von Bortkiewicz in his 1922 work and further discussed some non-normal parent distributions. All this early work up to mid-1920s contributed greatly to the gradual and systematic development of EVT, notwithstanding that the work had little direct relevance

to EVT (Rajaram, 2006).

The work that had more direct relevance to EVT was done in 1927 by another influential mathematical statistician Fréchet (1927) who studied the asymptotic distributions of largest values and discussed them in a paper published in French. The work of Fréchet was extended in 1928 by Fisher and Tippett (1928) who studied the same problem and went on to show that extreme limit distributions can only be one of three types. At that stage the three types of extreme limit distributions were not formally known but would later be known as the Fréchet, Gumbel and Weibull families of distributions. Following some other work in between the years, another most important contribution towards EVT development was made in 1936 by von Mises (1954), almost a decade after the work by Fisher and Tippett (1928). von Mises (1954) in his paper reproduced and published in 1954 extended the work by Fisher and Tippett (1928) in 1936 by studying each of the three types of limit distributions and presenting the *sufficient* conditions for the weak convergence of the largest order statistics to each of the three types of limit distributions (Rajaram, 2006, for more information).

According to Singpurwalla and Smith (2006), during the period 1940 to 1943 Gnedenko published a lot of work on limit distributions for the maxima of a series of random variables. Gnedenko (1943) published what is now known as the "*breakthrough paper*" in which he presented the most definitive results on the *necessary* conditions of weak convergence and the domains of attraction of each of the three types of limit laws introduced by Fisher and Tippett in 1928. In other words, Gnedenko (1943) provided the *necessary* and *sufficient* conditions for the limit laws of EVT. Gnedenko's work is reported to be "the first mathematically rigorous treatment of the fundamental limit theorems of extreme value theory" (Singpurwalla and Smith, 2006, p.82). Gnedenko was

awarded the Chebyshev prize in 1951 for his work on limit thoerems which he combined in 1949 into a monograph with Kolmogorov entitled "*Limit distributions for sums of independent random variables*" (Singpurwalla and Smith, 2006, p.82). The work by Gnedenko (1943) was very influential in probabilistic theory of extremes to the extent that it set the agenda for the three decades that followed. The work by Gnedenko was later refined and extended by many others including, notably, Mejzler (1949) who published his work entitled "*On a theorem of B.V. Gnedenko*" and much later on by de Haan (1970) who extended Gnedenko's work, in his PhD thesis, to regular variation and its application to the weak convergence of sample extremes. The work by de Haan (1970) and de Haan and Ferreira (2006) has also been very instrumental in recent research on probabilistic theory of extremes. More details on the historical development of EVT are also found in Berning (2010).

In 1974, Balkema and de Haan (1974) stimulated the theoretical developments of EVT regarding excesses over threshold in their work on residual life at great age. Pickands (1975) generalised the classic limit laws. The work by Balkema and de Haan (1974) and its extension and generalisation by Pickands (1975) gave rise to the generalised Pareto distribution (GPD). In the 1980s Leadbetter and others extended the theoretical developments of EVT to stationary processes. For example, Leadbetter, Lindgren and Rootzén (1983) published their work on the properties of sequences and processes related to EVT. In the 1990s focus on EVT research turned to multivariate and other techniques explored in order to improve inference. In the 2000s research interest in EVT was dominated by spatial and spatio-temporal applications, as well as applications in Finance. In the current decade, 2010-2020, while there is still on-going research interest in multivariate techniques in EVT, spatial and spatio-temporal applications, there has been growing interest in climate change induced extremes, financial risk management and a revisit to previous techniques and

approaches. For example, Ferreira and de Haan (2015) revisited the block maxima approach in EVT and justified, in theoretical terms, situations where the block maxima can be more preferable to peaks-over-threshold (POT).

While the early theoretical developments of EVT were in the 1920s and mid-1930s, much scholarly work dealing with various applications of the theory of EVT was published in the early 1940s. Emil Gumbel played a pivotal pioneering role in the 1940s and 1950s in the promotion of EVT as a tool for modelling tail behaviour of extreme events such as floods, droughts, and other disastrous rare events. Gumbel (1941) published his pioneering work in hydrology on the return period of river flows in which he argued that the rivers know the theory and only the engineers needed to be convinced of its validity. Later in the 1950s Gumbel published several other papers on theory and application of EVT (Gumbel, 1954) and his most popular book on statistics of extremes (Gumbel, 1958). The Gumbel distribution is named after Emil Gumbel which is clear evidence of his contribution towards EVT and its applications. Currently there are many excellent books on the theory and application of EVT. The books by Leadbetter et al. (1983) and de Haan and Ferreira (2006) are among excellent books that give a comprehensive mathematical background of the theory of EVT. The books by Reiss and Thomas (2007), Beirlant et al. (2004) and Coles (2001) are among the best books on the practical applications of EVT and data analysis in hydrology, insurance, finance and other related fields. The notation by Coles (2001) and Beirlant et al. (2004) are primarily followed in this thesis.

1.2.4 Problem statement

Mozambique is particularly vulnerable to climate change impacts due to its geographic location, and consequently faces a wide variety of natural disasters such as floods, droughts and wind storms. As a developing and emerging nation, Mozambique is not well-prepared for the catastrophic floods it is fre-

quently experiencing. The weak socio-economic infrastructure and extreme poverty of the population of Mozambique, with around 10 million people living in absolute poverty by the year 2007, render the country vulnerable to the flood phenomenon (Arndt and Simler, 2007). The current population of Mozambique is approximately 25.2 million with 70% of the population living in rural areas, and the national poverty rate stands at 54.7% (UNDP, 2015). Poverty in Mozambique contributes to the country's vulnerability to floods by driving a large number of people to build weak infrastructures in the floodplains in order to gain their livelihoods from agriculture and fisheries (Cuamba and Maúre, 2008). Floods in the southern region of Mozambique, mainly in the LLRB, routinely cause large-scale loss of life and catastrophic damage to cities, towns and villages (Jackson, 2013a,b; WMO, 2012). The communities that are most vulnerable to flood disasters are mainly those who reside in the rural areas where the majority of the poor live below UNDP poverty datum line of US$1.25 per day. Living below the UNDP poverty datum line implies that these communities have very limited capacity to be resilient against flood disasters in the LLRB and the floods worsen their poverty situation by washing away or damaging the little they have whenever they are caught unprepared (WMO, 2012). The general public such as travellers and investors who have their businesses in the flood prone areas are also affected by these flood disasters in the LLRB and other parts of the country when roads, bridges, telecommunications, farms, buildings and homes are either damaged or destroyed by floods.

Floods are generally of grave concern because they destroy crops cultivated in floodplains. However, floods can also have a positive impact as they may bring more water and increase availability of fish, which boosts the fisheries industry (one of the two main economic sectors in Mozambique). Floods also deposit topsoil and so increase soil fertility in the floodplains, boosting the agriculture industry, which is the other main economic sector in Mozambique (UNEP-

UNCHS, Sa). It is estimated that about 80% of the population of Mozambique gains its livelihood from agriculture (UNEP-UNCHS, Sa). Both the agriculture and fisheries industries depend on the state of the environment. Thus, flooding can have a tremendous beneficial or detrimental impact on the economy of Mozambique since it greatly influences the two main sectors of the economy. Given the positive and negative impact of the floods, it is easy to understand that normal floods can boost the economy whereas extreme floods can have devastating effects on the Mozambican economy.

The increased flooding and consequent social and economic losses in Mozambique pose a great threat to the country's ability to achieve the Sustainable Development Goals (SDGs) (formerly known as Millennium Development Goals (MDGs)) and economic development (UNDP, 2010). In addition to the socio-economic losses, a substantial amount of financial and other resources that were originally designated for development gets diverted to relief and rehabilitation assistance towards disaster-affected people each year.

According to the CII (2009) and Mondlane et al. (2013) climate change arising from extreme events such as floods and storms may result in unbearable losses in developed countries and prevent them from the introduction of insurance to developing countries. This situation can indeed leave developing countries such as Mozambique seriously exposed to extreme events. The progressive onset of climate change cost diagram in Mondlane et al. (2013, p.381) with further reference to the CII (2009) highlights that the cost of climate change starts with damage of infrastructures such as buildings, bridges and roads, among others; followed by operational cost caused by escalation of prices due to increased operating costs on consumers and businesses enhanced by weather patterns and increased sea level. Mondlane et al. (2013, p.381) states that the third climate change cost is uncertainty cost, that is "the opportunity cost will emerge; [aris-

ing from] the deferment of decisions due to uncertainty as the realization grows that climate change is a material issue", and lastly societal cost which arises due to environmental damage affecting the economically challenged vast majority of the population resulting in the creation of socio-economic stresses and escalation of social instability.

The problems highlighted in this thesis if not properly addressed or managed efficiently can pose serious challenges to the already economically challenged developing countries worldwide, particularly Mozambique whose disaster recovery budget is mainly dependent on donor funds and the country is just over two decades coming from a devastating civil war that lasted for nearly two decades resulting in hyperinflation.

1.3 Motivation of the study

The motivation for studying extreme floods in the LLRB of Mozambique is to reduce the associated risk and mitigate the deleterious impacts of these floods on humans and their property. It is argued by the International Federation of Red Cross (IFRC) that aid money buys more than four times as much humanitarian impact if spent before a disaster rather than on post-disaster relief operations (Redvers, 2009). Following the claim by the IFRC, it is therefore of paramount importance to reduce the predictive uncertainty of extreme floods, and thus reduce flood risk through the use of EVT techniques which have not been applied in the LLRB of Mozambique.

1.4 Purpose of the study

1.4.1 Research aim

The principal aim of this research is to reduce the predictive uncertainty of extreme floods in the vulnerable areas of Mozambique, the lower Limpopo River basin in particular, through the use of statistics of extremes. Hopefully these efforts will result in effective political and economic policies directed at these issues.

1.4.2 Objectives

The specific objectives of this research are to:

1. fit and compare several families of extreme value limit distributions for their goodness-of-fit to the annual maximum daily flood heights for the LLRB using Anderson-Darling and Kolmogorov-Smirnov tests, as well as other goodness-of-fit tests.

2. investigate whether there are significant differences between the annual maximum daily series and its corresponding annual moving sums in terms of measures of relative variability and other various river flow characteristics.

3. model the tail behaviour of extreme annual maximum daily flood heights for the LLRB.

4. model the tail behaviour of extreme daily flood heights that exceed a predetermined high threshold for the LLRB.

5. model the extremal behaviour of extreme flood heights (both annual daily maxima and excesses over a high threshold) in the presence of covariates for the LLRB.

6. improve and extend some of the extreme value models in the basin.

7. suggest areas for further research.

1.5 The hydrometric data

The flood height data used for this study were obtained from Mozambique National Directorate of Water (DNA) which is the authority responsible for water resource management in Mozambique under the Ministry of Public Works and Housing. The data cover all the key hydrometric stations for the LLRB. The flood heights (in metres) consist of daily readings with some readings dating back to 1940s. Each hydrometric station has a unique code starting with letter E, e.g. E35 for Chokwe.

The four key hydrometric stations for the Limpopo River are Chokwe, Combomune, Pafuri, and Sicacate. The LLRB has uninterrupted flood height data series for hydrometric stations Chokwe (E35) from 1951 to 2010, Sicacate (E36) from 1952 to 2010, and Combomune (E33) from 1966 to 2010, generating 60, 59 and 45 years of annual flood height data series, respectively (see Figures 1.1 to 1.6). The other hydrometric station, Pafuri (E31), has missing values that render the sample size too small for the purpose of this study except for regional flood frequency analysis.

All the hydrometric data used in this research are quantitative and continuous. The data are time series in nature and measured in metres (Figures 1.1 to 1.6). The raw data consist of daily flood heights recorded at most three times a day, that is, morning, afternoon and evening throughout the year (Figures 1.1 to 1.3).

Two fundamental data analysis procedures, block maxima and POT, are used

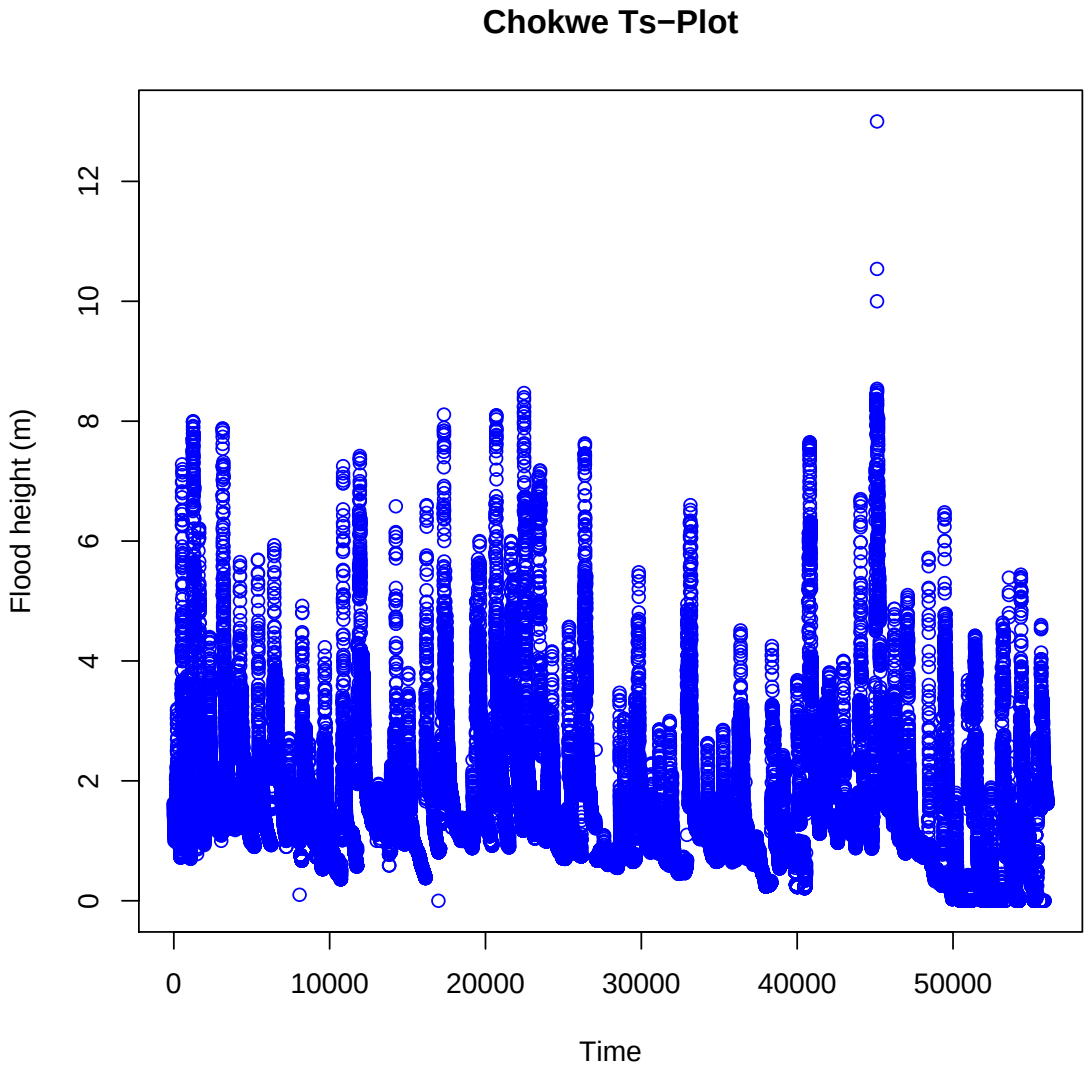


Figure 1.1: Time series plot of the instantaneous daily flood heights (in metres) at Chokwe hydrometric station(1951-2010) along the lower Limpopo River of Mozambique.

in this thesis. For the POT procedure, the data are not assumed to be independent because exceedances above a given threshold may be very close together, e.g. a number of high values mainly occur within a certain month or year while in other months or years there are no exceedances (Figures 1.1 to 1.3).

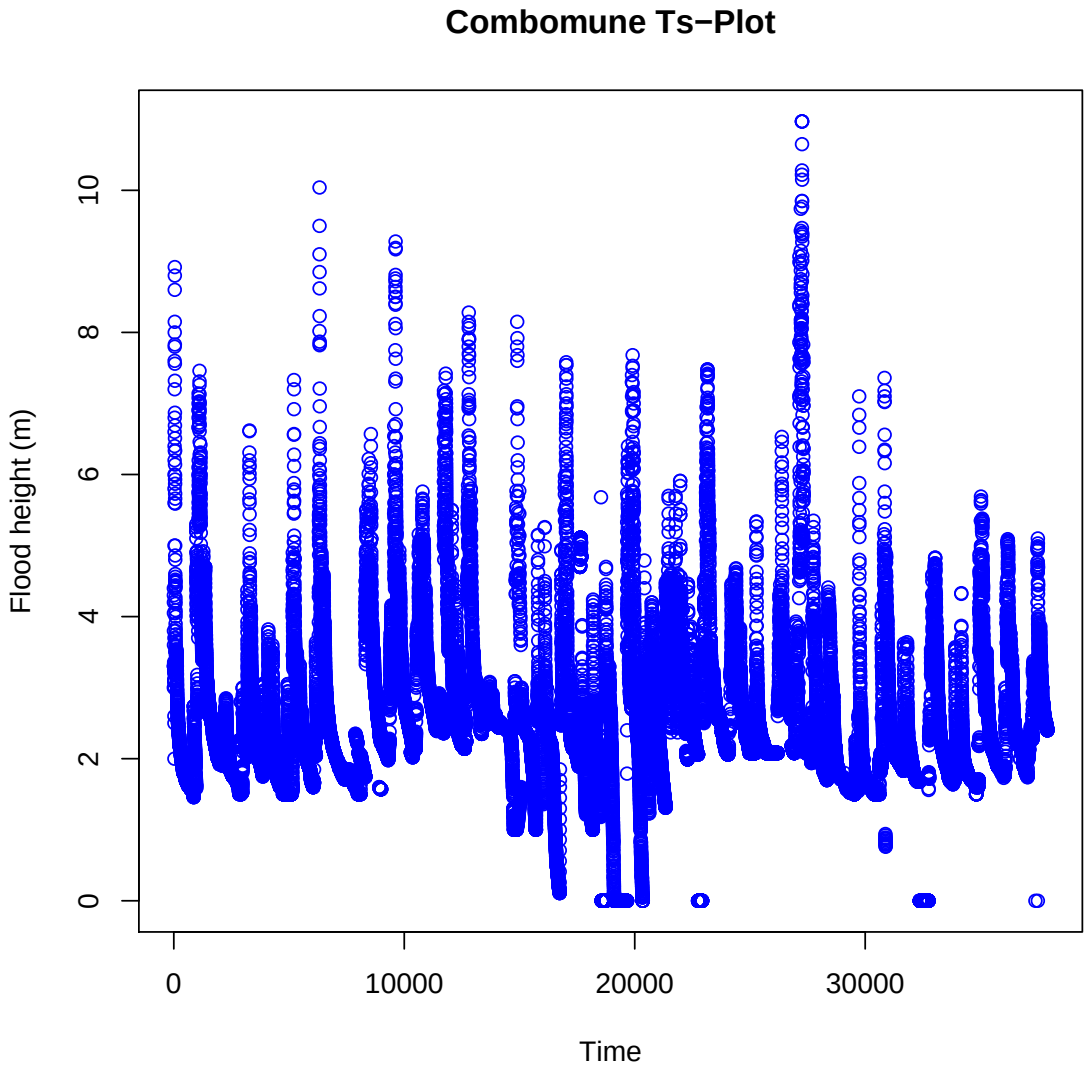


Figure 1.2: Time series plot of the instantaneous daily flood heights (in metres) at Combomune hydrometric station (1966-2010) along the lower Limpopo River of Mozambique.

In other words the exceedances may appear in clusters due to the existence of temporal dependence. The main reason for the existence of temporal dependence is that floods can compound each other. For instance, if a flood has recently occurred, another flood is more likely to follow before the water recedes to normal levels. In order to achieve independence or near independence

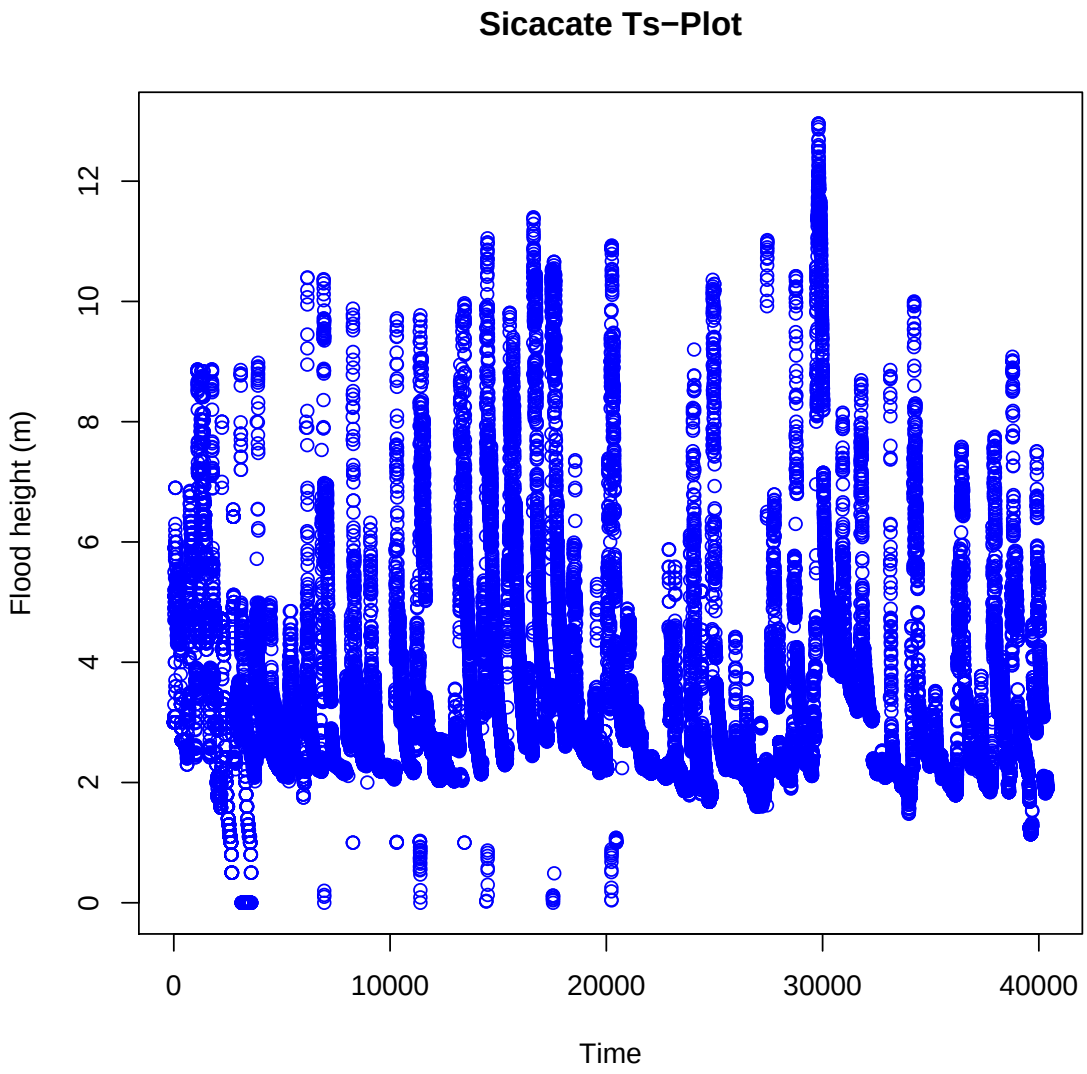


Figure 1.3: Time series plot of the instantaneous daily flood heights (in metres) at Sicacate hydrometric station (1952-2010) along the lower Limpopo River of Mozambique.

of data in the POT procedure a technique called declustering is used in this thesis.

The process of declustering involves filtering the dependent exceedances in the clusters while keeping the highest observation in each cluster such that the

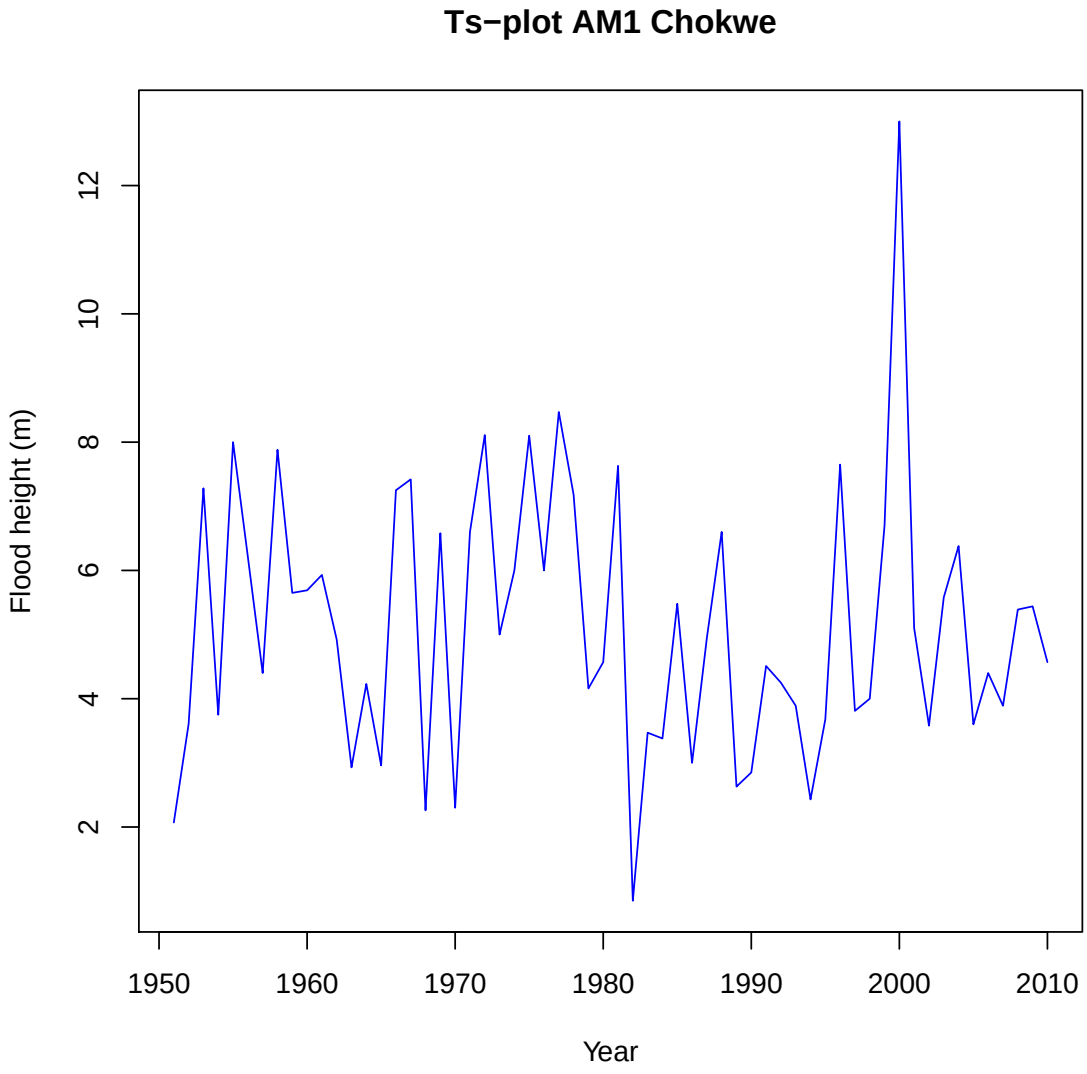


Figure 1.4: Time series plot of the annual maximum daily (AM1) flood heights at Chokwe hydrometric station (1951-2010) along the lower Limpopo River of Mozambique.

remaining observations are approximately independent (Cooley, 2005; Coles, 2001). The topic of declustering is further expanded in the next chapter.

For the block maxima approach, sequential steps are taken to select the highest daily flood height (water level) in each hydrological year (see Figure 1.4 to

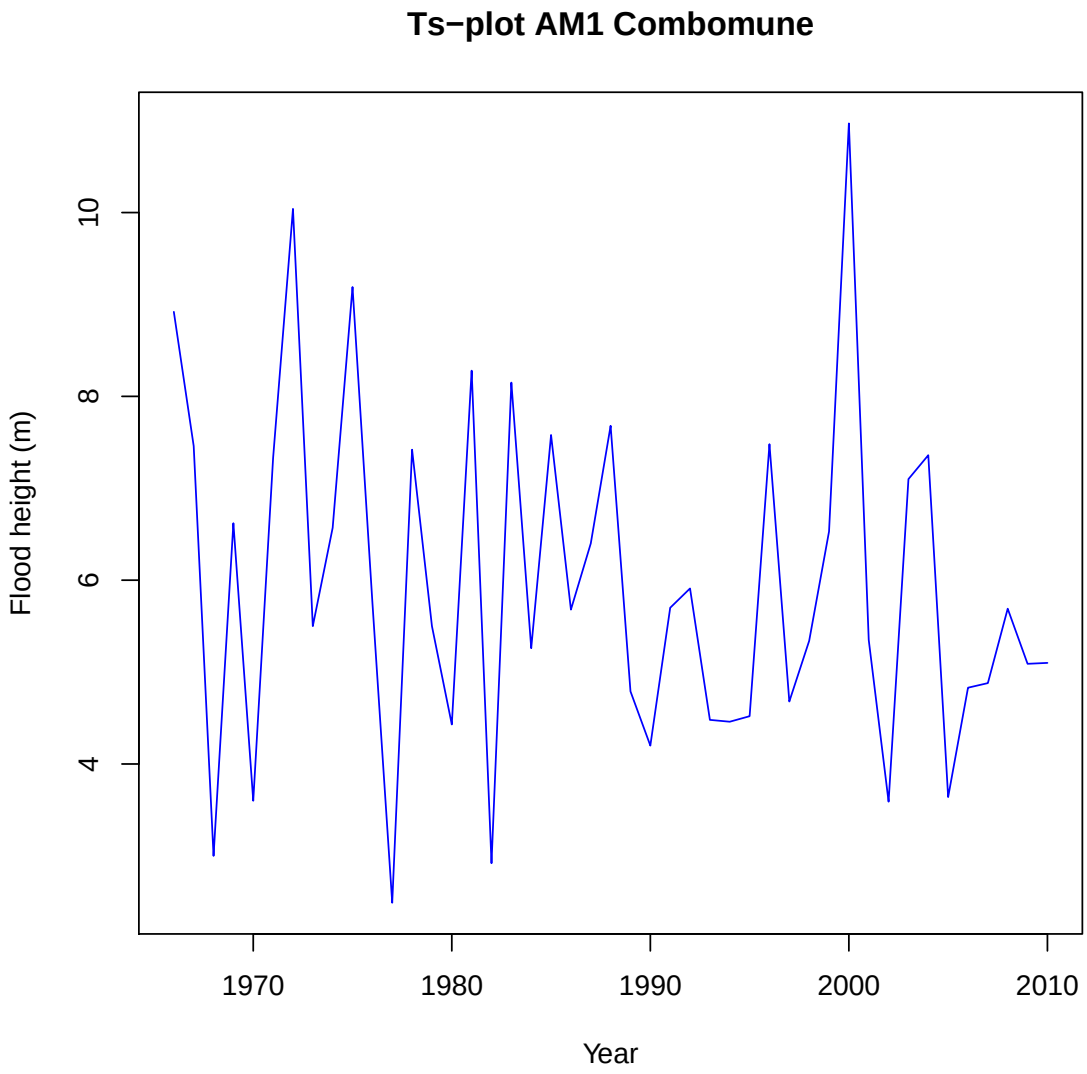


Figure 1.5: Time series plot of the annual maximum daily (AM1) flood heights at Combomune hydrometric station (1966-2010) along the lower Limpopo River of Mozambique.

1.6). The block maxima approach results in data series that are assumed to be independent and identically distributed (Figures 1.4 to 1.6).

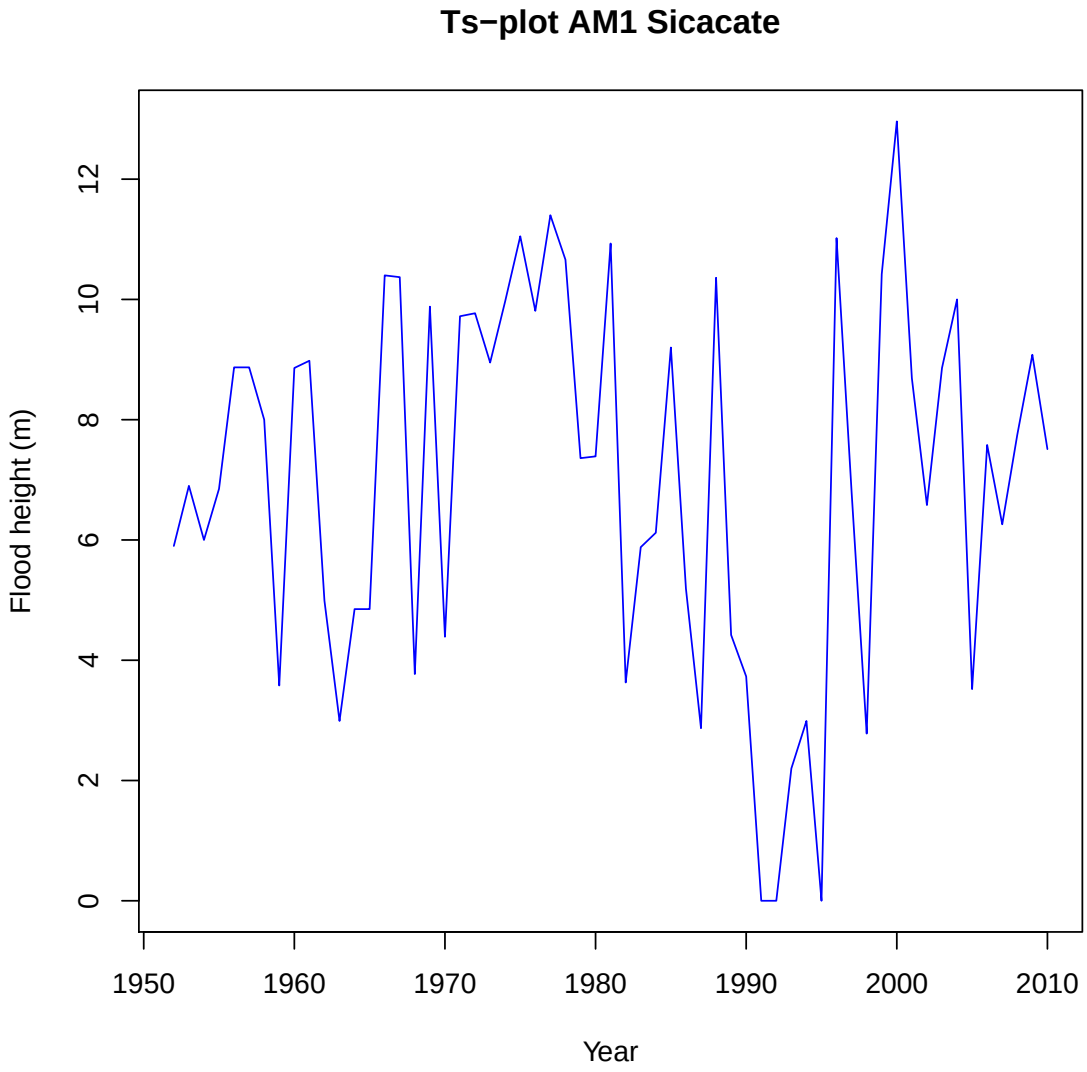


Figure 1.6: Time series plot of the annual maximum daily (AM1) flood heights at Sicacate hydrometric station (1952-2010) along the lower Limpopo River of Mozambique.

1.6 Importance of the study

The study is important for several reasons. If a community understands its vulnerabilities and its capacity, it can learn to support itself. In addition, if the predictive uncertainty of extreme floods is reduced, a substantial amount of financial and other resources can be saved.

This study is also important because it disseminates information on the occurrence and impact of one of the natural hazards (floods) in Mozambique. Mozambique is a country that is emerging from almost two decades of civil war and hyperinflation, and has seen the rate of economic development profoundly derailed by floods (Arndt, James and Simler, 2006), therefore the findings of this thesis will be crucial for the country's policy-makers in the fields of hydrology and water management. The Limpopo River basin has an impact on several other Southern African countries, hence the information disseminated from this research will also be useful to other countries in the SADC region.

The results of this research will serve as a benchmark for comparison with other models that are currently in use. This study will attempt to achieve some of the principal goals of Extreme Natural Hazards and Societal (ENHANS) International Workshop on Extreme Natural Hazards and Disaster Risk in Africa given in Chapter 2 which include among other goals, promoting studies on the prediction of extreme events and therefore reduce the predictive uncertainty, and it will lay the foundation for future research in modelling the tail behaviour of extreme floods for the LLRB. Knowledge of the distribution of observed flood heights (water levels) will be instrumental in the estimation and forecasting of future flood levels and resource allocation by policy and decision makers.

1.7 Scientific contributions of the study

The major contribution of this thesis is in applying statistics of extremes in modelling extreme daily and annual maximum daily flood heights for the LLRB of Mozambique. The specific contributions are as follows:

1. Offering for the first time, to the best of our knowledge, a rigorous extreme value theory treatment of the flood heights data in the LLRB of Mozambique.

2. Modelling the cumulative effect of the river flow by considering the moving sums of the daily flood heights for the LLRB of Mozambique. The findings obtained from the study of the cumulative effect of the river flow may not only be important to the LLRB of Mozambique but may be generalised to other rivers of similar characteristics and climatic conditions.

3. Construction of the flood frequency curves for the LLRB using the Bayesian and classical approaches, which will guide the policy-makers in the region in their long-term disaster risk reduction and management efforts.

4. Modelling the extremal behaviour of the daily flood heights and annual maximum daily flood heights for the LLRB of Mozambique in the presence of covariates.

5. Creating a benchmark for the LLRB of Mozambique for future research based on the application of extreme value theory in the basin.

6. Reducing the predictive uncertainty of extreme floods in the basin by proving the probabilities of return of extreme flood heights and their corresponding return periods. This is one of the principal goals of ENHANS project to promote studies on natural hazards in developing countries and reduce the predictive uncertainty of extreme natural hazards in these countries (Maree, 2011).

7. Extension and improvement of the existing models in the basin.

1.8 Outline of the thesis

The structure of this thesis is arranged such that Chapter 1 gives an introduction to the research study. Chapter 1[1] presents a comprehensive historical background of both the Limpopo River and theory and applications of EVT. The

[1]Fighting flooding in Mozambique. *ORMS Today*, **42**(2), April 2015 Edition, 2015.

problem statement, motivation of the study, broad aim and objectives of the study, the data used in the study, significance of the study and scientific contributions of the study are all given in Chapter 1. The highlights of the chapter are the early theoretical developments of EVT by Fréchet (1927), Fisher and Tippett (1928), the mathematical rigour treatment of EVT in the early 1940s and 1950s by Gnedenko, extension of Gnedenko's work by de Haan in the early 1970s and the pioneering work in the practical applications of EVT by Gumbel in the same period of Gnedenko's theoretical developments.

Chapter 2 reviews literature on the current flood forecasting techniques used in the LLRB, EVT techniques currently used in the basin, the gaps in the forecasting methods currently used in the basin, worldwide efforts towards reduction of natural disasters, efforts in the Southern African Development Community (SADC) region including LLRB governing authorities towards flood disaster risk reduction. The chapter also reviews current research focus in EVT in order to explore how the current worldwide research direction in hydrology and other related fields can be of benefit to the SADC region, particularly the south-eastern country of Mozambique.

Chapters 3 to 7 have been submitted individually for publication and therefore most of these chapters are self-contained. However, attempt is made in this thesis such that the notation remains consistent throughout the thesis. Chapters 3 to 8 model the tail behaviour of extreme daily flood heights and extreme annual maximum daily flood heights. These chapters also cover relevant recent literature for each chapter. The detailed chapter by chapter presentation is given in the next paragraph.

Chapter 3[1] compares ten candidate distributions for their goodness of fit to the

[1]Investigating the goodness-of-fit of ten candidate distributions and estimating high quantiles of extreme floods in the lower Limpopo River Basin, Mozambique. *Journal of Statistics*

annual maxima daily flood heights at the three sites Chokwe, Combomune and Sicacate under consideration in this study. The broad aim of Chapter 3 is to identify suitable families of extreme value limit distributions that can 'best' mimic the frequency of extreme flood heights at the three sites for the Limpopo River. The distributions identified in Chapter 3 are extended in the succeeding chapters by exploring various parameter estimation techniques including Markov chain Monte Carlo (MCMC) Bayesian techniques, as well as including climate change related factors in the models.

Chapter 4[1] deals with a comparative analysis of the moving sums of annual maxima daily flood heights time series models. The annual maximum daily series (AMDS) is compared to the annual 2-day maximum (AM2), annual 5-day maximum (AM5), annual 7-day maximum (AM7), annual 10-day maximum (AM10) and annual 30-day maximum (AM30) in order to explore if there are any significant differences between the AMDS and the corresponding annual moving sums in terms of variability and other various river flow characteristics.

Chapter 5[2] deals with the estimation of high quantiles of extreme flood heights using model based Bayesian MCMC techniques to estimate the parameters of the GEV distribution for the lower Limpopo River. The use of Bayesian technique enables the model to take into account uncertainty in the estimation of the parameters. In Chapter 5[3] the Bayesian MCMC inference procedure is also

and Management Systems, **17**(3): 265-283, doi:10.1080/09720510.2014.927602, 2014.

[1]Comparative analysis of annual maxima time series models along the lower Limpopo River basin of Mozambique. In: Awotona, A. (Ed.), (2016). Planning for Community-based Disaster Resilience Worldwide: Learning from Case Studies in Six Continents. *Ashgate Publishing Limited*, 2016.

[2]Estimating high quantiles of extreme flood heights in the lower Limpopo River basin of Mozambique using model based Bayesian approach. *Natural Hazards and Earth System Sciences, Discussion*, **2** (8): 5401-5425, doi:10.5194/nhessd-2-5401-2014, 2014.

[3]Construction of flood frequency curves in the lower Limpopo River basin of Mozambique using Bayesian and Markov chain Monte Carlo methods. *Proceedings of the 60th International Statistical Institute (ISI) World Statistics Congress*, 26-31 July 2015, Rio de Janeiro, Brazil.

used to develop and construct flood frequency curves for the LLRB.

In Chapter 6[1] the effect of climate change is considered in the model through the use of a time-heterogeneous GEV distribution to model annual maximum daily flood heights. The approach used in the chapter is that of Coles (2001) in which only the trend and scale parameters of the GEV distribution are allowed to change with time. The fundamental approach used in Chapter 6 is that of block maxima. The chapter also discusses the recent theoretical advances in block maxima after its recent revisit by Ferreira and de Haan (2015).

Chapter 7[2] is based on the POT fundamental approach. A time-heterogeneous GPD is used to model daily flood heights for the Limpopo River above a certain high threshold. In the chapter various techniques are used to identify a 'reasonably' high threshold for each given site. Again guided by Coles (2001), only the scale parameter of the GPD is allowed to vary with time in the non-stationary model. The chapter also discusses the theoretical link between GPD and GEV distribution. The findings in Chapter 6 and those in Chapter 7 are compared in order to investigate the closeness of the practical application results to the theoretical link between the GPD and GEV models.

Chapter 8 presents the r largest order statistics results. The results of the r largest order statistics are compared to those of the GEV distribution in the standard block maxima framework.

Chapter 9 concludes the thesis. The chapter consists of summary of the thesis and concluding remarks, summary of the key findings and contributions, limitations of the study and areas for future research directions. In general, the

[1]Modelling nonstationary annual maximum flood heights in the lower Limpopo River basin of Mozambique. *Jàmbá: Journal of Disaster Risk Studies*, 2016.

[2]Modelling extreme flood heights in the lower Limpopo River basin of Mozambique using a time-heterogeneous generalised Pareto distribution. *Statistics and Its Interface*, 2016.

chapter summarises what the thesis has covered and goes on to propose what may need to be covered in the future.

The main statistical package used throughout this thesis is R (Heffernan and Stephenson, 2015). EasyFit statistical package was also used in few occasions. Some of the R codes developed in this thesis are given in appendices at the end of each chapter and also end of thesis.

Chapter 2

Literature review

. . . Science is more than a body of knowledge; it is a way of thinking. The method of science, as stodgy and grumpy as it may seem, is far more important than the findings of science

... Carl Sagan

2.1 Introduction

The prediction of extreme flood levels and their corresponding return periods is very important for planning and decision making processes in government, construction engineering and hydraulic sectors. For instance, if a civil engineering structure such as a bridge is to be constructed, prior knowledge of the flood heights of the historic extreme floods in the particular river basin will be required, as well as knowledge of the recurrence period of such extreme floods in the basin.

While flood frequency information is important to engineering and hydraulic

structures such as bridges, dams, spillways, etc, it can also be used in land use and settlement control on flood-prone areas. According to Smithers (2012, p.635) flood frequency analysis (FFA) "remains a subject of great importance" because of its impact on the economy and the environment, and its "preservation of human life and property". However, Smithers (2012) and Cordery and Pilgrim (2000) argued that the demands for improved estimates of floods are not yet satisfied in terms of reliability of flood frequency peak flows and volume discharges. This is despite the increased understanding of fundamental processes in hydrology (Smithers, 2012). The issue of unreliable estimates of flood heights (or peak flows) and/or volume discharges brings serious economic challenges to hydrologists, engineers and government strategic planners. For instance, underestimation of flood heights has serious implications in that the structures built based on such estimates get easily destroyed during periods of floods. On the other hand, overestimation of flood heights results in wastage of resources needed to build the structures (Sigauke, 2014). It is therefore important to improve the accuracy of the estimates of flood heights in the LLRB to avoid these detrimental consequences.

The methods based on the analysis of observed floods can be grouped into empirical formulae, flood frequency analysis (FFA) and flood envelope curves (Smithers, 2012; Cordery and Pilgrim, 2000). The empirical formulae methods are based on algorithms which relate peak discharge to catchment size and its climatic characteristics among other catchment characteristics. Smithers (2012, p.635) and Cordery and Pilgrim (2000) condemned the use of the methods as "extremely hazardous" especially if the calibrations do not come from the catchment being considered. The catchment parameter method (CAPA) is among the empirical formulae methods. It is an index-flood type approach where the mean annual flood is estimated as a function of catchment mean annual precipitation (MAP), area, slope and catchment shape parameter (Smithers,

2012). In the CAPA approach the scaling factor is used as a function of MAP and exceedance probability of the design flood. It is advised that empirical and experience-based methods should be avoided when analysing observed floods and should only be used for checking other methods (Smithers, 2012; Cordery and Pilgrim, 2000).

The FFA methods are used to perform design flood estimation by analysing "observed flows where these are available and adequate in both length and quality" (Smithers, 2012, p.635). The FFA methods are categorised into two groups, namely at-site and regional approaches. Both at-site and regional FFA make use of the fitting of a probability distribution to the data (Blain and Meschiatti, 2014; Atroosh and Mustafa, 2012; Smithers, 2012). Smithers (2012) gave a detailed review of the two approaches of flood frequency analysis methods. This thesis is based on FFA methods. The two approaches, at-site and regional FFA, are reviewed in detail later in the coming sections of this chapter.

In the flood envelopes and regional maximum flood (RMF) methods of analysing observed floods, the largest observed flood discharges are usually plotted against catchment area on logarithmic axes (Smithers, 2012). All the data points are then included by sketching an envelope. In order to make estimates, data from catchments similar to the catchment of interest should be included in the design (Cordery and Pilgrim, 2000). The method allows for the determination of maximum peak discharges at ungauged sites. The size of the envelope is directly proportional to the increase in record length and observed larger floods (Smithers, 2012). According to Smithers (2012) the reliability of the this approach is limited to medium-sized catchments. The main shortcoming of the flood envelopes or RMF method is that there is no exceedance probability associated with the RMF (Smithers, 2012; Nortje, 2010).

In his concluding remarks of a thorough review of literature on the methods used for design flood estimation in South Africa and other parts of the world, Smithers (2012) acknowledged the gap between flood frequency research and practice emphasised by Cordery and Pilgrim (2000), where research is needed to improve the estimates of both deterministic and probabilistic floods. Although Smithers (2012) argued that such a gap between flood research and practice is not large in South Africa, it is arguably believed that such a gap can be large in economically challenged countries like Mozambique. Smithers (2012) advocated for the need to improve the existing methods and evaluate new methods adopted by other countries for design flood estimation with the view that these new methods may also work for the basins and catchments of the developing countries like Mozambique.

The rest of the chapter is organised as follows. Section 2.2 gives a review of relevant literature on flood forecasting and early warning systems in the LLRB of Mozambique, while a brief review of worldwide disaster risk reduction and flood management efforts is discussed in Section 2.3. Section 2.4 gives a discussion of statistical models currently used in the major river basins of Mozambique including the LLRB, while Section 2.5 reviews statistical models or distributions commonly used in flood frequency analysis. A discussion of the two approaches of flood frequency analysis, at-site and regional flood frequency analyses, appears in Section 2.6. Literature on modelling extreme flood flows or precipitation using EVT is discussed in Section 2.7, while literature on modelling of extreme flood levels using Bayesian Markov chain Monte Carlo (MCMC) techniques is discussed in Section 2.8. A review of literature on modelling nonstationary extremes in the presence of covariates is discussed in Section 2.9. A brief discussion of r-largest order statistics is discussed in Section 2.10 and finally Section 2.11 gives a summary of the chapter.

2.2 Flood forecasting and early warning systems in the lower Limpopo River basin

In the recent past attempts have been made with limited success to monitor Limpopo River flows and mitigate droughts and flood disasters. According to WMO (2012) weather and river gauging stations were installed in all the four riparian countries; Botswana, Mozambique, South Africa and Zimbabwe, in order to collect data for flood forecasting and early warning system (FFEWS) development. WMO (2012) reported that SADC – Hydrological Observing System (HYCOS) project which ran between 1998 and 2005 installed about 28 stations in the riparian countries with at least five in each country, except in Mozambique where there were only two. In an attempt to close the gap left by the SADC-HYCOS in Mozambique, the World Bank project installed 19 real time rain gauge stations in LLRB of Mozambique. Some flood forecasting and early warning system (FFEWS) models were also installed in Mozambique using Mike 11 'Flood Watch' and its upgrades and geo-spatial streamflow forecasting modelling (GeoSSFM). WMO (2012) further reported that less than 14% of the real time stations that were installed are still working and the (FFEWS) is struggling to be functional mainly due to inadequate project design, technical issues and maintenance problems.

WMO (2012) stressed the need for a fully operational and effective FFEWS which will benefit the vulnerable communities and the general public who are the chief stakeholders in the LLRB. WMO (2012) highlighted some flood risk management challenges existing in the current FFEWS in the LRB. These challenges include among other issues uncoordinated and incomplete FFEWS, inadequate water resources information and flood monitoring systems necessary for collecting data and information, limited data exchange and technical cooperation between National Weather Services and National Hydrological

Services within the country and among National Hydrological Services within the LRB, limited institutional and capacity development such as a regional centre that can coordinate telemetry work and lead in river flow of FFEWS development which can lead to the issuing of flood warning products.

WMO (2012) reported that the Mike 11 based FFEWS is experiencing problems due to data transmission issues from the telemetry system and problems related to model boundary conditions, leading the Regional Water Administration for the South in Mozambique (ARA Sul) to resort to using GeoSSFM based FFEWS with data collected from satellite and radars in South Africa and its own telemetry system in Mozambique's part of LRB as input rainfall or run-off modelling component. This excludes contributions from Botswana and Zimbabwe leading to the FFEWS results to be only indicative and not always reliable (WMO, 2012).

WMO (2012) recommended the use of flood forecasting models that are calibrated and operated to produce highly accurate and timely forecasts that can be used to develop actions intended to save lives, protect property and infrastructure from floods. The current prominent forecasting models in the LLRB are the National Weather Services' River Forecasting System (NWSRFS) and Mike 11 Flood Watch. WMO (2012) further recommended flood early warning and response to be the responsibility of a regional centre which should issue early warnings and advisory statements during the flood warning process (i.e., when the flood warning is in effect). More details on other hydrological models used in the LLRB are found in (Mabote, 2011).

The FFEWS efforts discussed in this section are meant for short-term forecasting of floods. Flood frequency analysis aimed for this thesis gives long-term forecasts. Given the abundance of problems experienced in the current

FFEWS in the LLRB of Mozambique which render the short-term forecasts based on the available geo-scientific models unreliable and untrustworthy, it is clear that statistical models which are probabilistic in nature are needed to complement the existing FFEWS in the basin. The combination of short-term and long-term forecasting can add value to the current flood forecasting situation in the basin.

2.3 Worldwide disaster risk reduction and flood management efforts

The risk of a flood is defined as the chance or likelihood of injury, loss, or damage (van Ogtrop et al., 2005). According to these authors, risk is directly proportional to the return probability and damage. In other words the risk, R, is given by: $R = p \times d$, where the damage, d, of a flood is expressed in terms of economic, environmental and social related costs while the return probability or probability, p, of a flood is the likelihood of a T-year flood level to occur once in T years, where T is the return period, e.g. 100 years. Attempts have been made with limited success worldwide to reduce the risk of natural disasters.

According to UN (2005) the World Conference on Disaster Reduction held in Kobe, Japan, in 2005, called for the establishment of a clear framework for action to "ensure that disaster risk reduction is a national and a local priority with a strong institutional basis for implementation; identify, assess, and monitor disaster risks and enhance early warning; use knowledge, innovation, and education to build a culture of safety and resilience at all levels; reduce the underlying risk factors; and strengthen disaster preparedness for effective response at all levels". As a follow up to this conference an Extreme Natural Hazards and Societal Implications (ENHANS) International Workshop on Extreme Natural Hazards and Disaster Risk in Africa was held in Pretoria, South

Africa in 2011 (Maree, 2011).

The principal goals of ENHANS International Workshop on Extreme Natural Hazards and Disaster Risk in Africa were as follows (Maree, 2011):

- "improving understanding of critical phenomena associated with extreme natural events and analysing impacts of the natural hazards on sustainable development of society,
- promoting studies on the prediction of extreme events reducing predictive uncertainty and natural hazards mitigation, and bringing these issues into the political and economic policies,
- disseminating knowledge and data on natural hazards for the advancement of research and education in general with an emphasis on developing countries,
- establishing links and networks with international organisations involved in research on extreme natural hazards and their societal implications, and setting up a consortium of experts International Council of Science Unions (ICSU) and several major intergovernmental and multi-national organisations involved in the ENHANS project".

ENHANS placed special emphasis on the importance of research on extreme natural hazards and disaster risk mitigation in the most vulnerable regions of the world, including, among others, Africa and Asia (Maree, 2011). Africa is an extremely vulnerable region that faces many forms of natural disasters such as floods, droughts, landslides, and earthquakes. The majority of African countries are developing and emerging economically, which places these countries in a particularly precarious and vulnerable position with regard to risk of loss as a result of a natural catastrophe such as extreme floods.

According to MunichRe (2011), the ten significant natural catastrophes in Africa from 1980 to 2010 indicated that Northern Africa is mainly affected by earthquakes and tsunamis, whereas floods and droughts are dominant in Southern Africa. For example, in December 2010 a global weather pattern known as La Niña caused increased rainfall over Southern Africa, leading to widespread flooding (OCHA-ROSA, 2011). Countries affected by the resulting series of floods included Botswana, Lesotho, Malawi, Mozambique, Namibia, South Africa, Zambia, and Zimbabwe. In January 2012 tropical storms Funso, Dando and Irina resulted in 44 deaths and displaced 108,048 people in Mozambique. These tropical storms also destroyed (damaged or flooded) an estimated 28,000 homes, 735 classrooms, 31 health units and affected 140,538 hectares of crops (OCHA-ROSA, 2012).

More recently in 2013 extreme floods hit the LLRB town of Chokwe and provincial city of Xai-Xai in Mozambique in just over a decade after the same cities were over-flooded by catastrophic floods of the year 2000 that killed more than 700 people and caused one woman to give birth in a tree (Jackson, 2013a; Musiya, 2013). The recent 2013 floods caused two women to give birth on rooftops and claimed the lives of more than 111 people from October 2012 to February 2013 with more than 70% of the people killed after January 2013 when the major floods started (Jackson, 2013a; Musiya, 2013; OCHA-ROSA, 2013).

Cuamba and Maúre (2008) wrote a chapter on the challenges to management of floods and droughts in transboundary river basins in Mozambique. They gave a detailed account of the rivers and water situation in Mozambique, and noted that the LRB has no large dams, making it very vulnerable to extreme floods. They also pointed out that the LRB is also very vulnerable to droughts, making it very difficult to construct large dams along the Limpopo River since it is

often dry most of the time until the rainy season. Manuel and Vicente (2002) studied the problem of catastrophic flooding in Mozambique in 2000 from a geo-scientific point of view and concluded that floods are part of the process of nature and human societies must live with them. They advocated for structural and non-structural measures to minimise the negative impacts of floods.

van Ogtrop et al. (2005) studied flood management in the lower Incomati River in Mozambique using two alternative strategies based on the advocated view of living with the floods and the traditional view of flood control. The authors argued that the traditional approach of flood control has been largely implemented in developed countries through the use of dams, dikes and other engineering structures to control floodwaters. The alternative approach by van Ogtrop et al. (2005) is based on society living with the floods by being disaster prepared with the right disaster reduction measures in place. The authors concluded that the resilient pathway which involves living with floods may be more appropriate not only to Mozambique, but also to other developing countries when the year 2000 floods are put in consideration.

In an attempt to monitor extreme flood events in the LLRB of Mozambique, Asante et al. (2007) used a geo-scientific model in the form of satellite-derived precipitation which has the ability to identify extreme flood events and extreme precipitation. The satellite-derived model is also able to identify the spatial extent of the extreme flood events as well as their relative intensities. The authors reported that water managers in Mozambique use this remotely sensed precipitation approach to make an independent flood hazard verification before issuing flood early warnings. They concluded that the developed scientific model greatly enhances the ability of water management authorities in the LLRB of Mozambique to monitor extreme flood events and issue communities at risk with flood early warning information.

In another effort to reduce the risk of floods in the LRB, Spaliveiro et al. (2014) conducted a geo-scientific field research aided by an extensive review of literature of the past evolution of the Limpopo River with the principal aim of analysing flood hazard in the basin. In other words, the authors were determined to answer the question of whether there exists a clear relationship between current flood hazard and past fluvial changes in the LLRB. These authors noted that human settlements have a tendency to concentrate closer to the river streams which can be attributed to the heavy dependence on agricultural activities and favourable climatic conditions in the floodplains. In their study, Spaliveiro et al. (2014) found that both the 2000 and 2013 floods followed the same dynamics and the floodwaters for both of these events were concentrated in the palaeo-delta which was identified by the authors in the LLRB. The palaeo-delta in the lower Limpopo is responsible for the drainage pattern of the floodplain in Mozambique and also influences the present flood dynamics of the LLRB (Spaliveiro et al., 2014). The authors noted that successive tectonic events have led to major changes in the basin from the ancient geomorphological history known as the jurassic period to the present time. They concluded that the high flood hazard in the basin was connected to the morphology of the river. These authors suggested that their method could be applied in other developing countries where the flood dynamics are complex.

At an international level, the International Disaster and Rick Conference (IDRC, henceforth referred to as IDRC Davos 2014) was held in Davos, Switzerland in 2014 with the principal aim of reducing the risk of disasters, particularly natural disasters (Stal et al., 2014). Among other important issues, the IDRC Davos 2014 was intended to give an in-depth preparatory discussion of the then forthcoming IDRC Sendai 2015 which was later held in Sendai, Japan in 2015 to mark the end of the UN's two international agreements, that is, the MDGs and the Hyogo Framework for Action (HFA) (Stal et al., 2014). The MDGs were

subsequently replaced by the SDGs in 2015. The contributions of the participants at the IDRC Davos 2014 were expected to address the following topics (Stal et al., 2014, p.1):

- Dominating and emerging trends in disasters and risks in the world,
- Type of international instruments that should be developed in the post 2015 framework process, that is, after the HFA,
- The principal issues that must be considered "for the future in disaster risk reduction and resilience, and how they should be tackled".

In his welcome speech address, Dr. Walter Ammann, the chairman of the IDRC Davos 2014 reported that in the first half of the year 2014 a number of natural disasters had already hit mother earth with a number of disastrous floods, catastrophic earthquakes, forest fires and many other human-induced crises and disasters of political, social and technological origin, which is a persistent reminder to us of the vulnerability of our communities, the limitations of our abilities to provide help and the difficulties of overcoming such situations (Stal et al., 2014).

Among the important topics presented at IDRC Davos 2014 sessions were topics on risk assessment methodologies, topics on the understanding and management of floods and droughts in the river basins with examples from Africa and Asia, topics on risk management approaches, community-based disaster risk reduction, and topics on prevention, preparedness and intervention.

The issues raised in this section revealed the gaps that are still existing in disaster risk reduction. In order to close these gaps that are still existing disaster risk reduction, statistical techniques or in particular, statistics of extremes which is concerned with the analysis of these extreme events will be

crucial in addressing disaster preparedness through the reduction of the predictive uncertainty of the natural disasters. This thesis uses statistical techniques to help reduce uncertainties in the prediction of floods in the LLRB of Mozambique. The use of statistical techniques in the basin will complement the FFEWS in the basin and it is hoped that the combination of methods in the basin can go a long way towards making improvements in disaster preparedness, management and intervention.

2.4 Statistical models currently used in the major river basins of Mozambique

Lucio (2007) revealed that the most common statistical distribution used in flood frequency analysis in the main river basins of Mozambique is the Gumbel distribution. The Gumbel distribution was used in flood risk analysis of extreme daily rainfall at Maputo station, south of LLRB and Save River, where the February-March 2000 maximum daily rainfall rose to 328 mm setting a new local record. The results by Lucio (2007) using the Gumbel distribution revealed that the 100-year flood level was 364 mm/day indicating that the 328 mm/day new local record at Maputo station was just above the 50-year flood level of 325 mm and way below the 100-year flood level. These results appear to underestimate the new local record, suggesting that other more appropriate statistical flood frequency models may be required in Mozambique to help in engineering design and mitigation of flood impacts. The February-March 2000 flood exceeded historical peaks by a factor of at least two in most rivers north of Maputo station including lower Limpopo River and Save River (Lucio, 2007).

In a workshop organised by UNDP Mozambique, Mozambique's National Institute for Disaster Management (INGC), Global Risk Identification Programme (GRIP) and Hepia/PMSA held in Maputo, on national flood risk assessment

and mapping in Mozambique, participants were trained on hydrological data processing using the data at Chokwe in the LLRB (UNDP, 2011). Among the exercises completed by the participants at the workshop were data control, statistical analysis using the Gumbel distribution and control of water levels (UNDP, 2011). Some of the files given to the participants at the workshop were on Gumbel distribution adjustment of extreme flood flows time series of Tete and Chokwe stations in the lower Zambezi and Limpopo River basin, respectively (UNDP, 2011).

UNDP (2011) reported that participants who attended the Maputo workshop were trained on the law build up, limitations and use of the Gumbel distribution, as well as extrapolation of flow or rainfall for a given return period. Training on adjustment of intensity duration frequency (IDF) curves from rainfall data, use and limitations of IDF curves were also covered at the 5-day workshop (UNDP, 2011). The knowledge gained by the participants at the workshop was intended for use at all the major basins in Mozambique namely Zambezi, Limpopo, Incomati, Licungo, Save, Buzi and Pungue (UNDP, 2011). The emphasis on the Gumbel distribution at the Maputo workshop revealed that the main statistical extreme flow frequency distribution model used in the main river basins in Mozambique is the Gumbel distribution (UNDP, 2011).

More recently Mondlane et al. (2013) wrote a paper on the application of extreme value distributions using Xai-Xai station in the Limpopo River part of Mozambique as a case study. The distributions used by Mondlane et al. (2013) were the Gumbel, Fréchet and Weibull distributions which are families of the generalised extreme value (GEV) distribution given by

$$G_{\xi}(x) = \exp\left\{-\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]_{+}^{-\frac{1}{\xi}}\right\}, \text{ for } \xi \neq 0 \tag{2.1}$$

where ξ is the extreme value index (EVI). When $\xi = 0$ equation (2.1) is the Gumbel family of distributions which is given by

$$G_{\xi}(x) = \exp\left\{-\exp\left[-\frac{x-\mu}{\sigma}\right]\right\}, \quad \xi = 0 \tag{2.2}$$

In (2.1), when $\xi > 0$ we have the Fréchet class of distributions and when $\xi < 0$ we have the Weibull class of distributions. Mondlane et al. (2013) found the two parameter Gumbel Min and Weibull distribution to provide a better fit to the simulated Xai-Xai data used in the analysis and the Gumbel Max provided a better fit to the collected Xai-Xai data. It should, however, be noted that the sample annual maximum rainfall series of 20 years used by Mondlane et al. (2013) to arrive at their conclusion was quite small.

The Gumbel distribution, which is unbounded both from below and from above, is heavily criticised in literature for its inappropriateness to model extreme hydrological events (Koutsoyiannis, 2004; Rowinski and Strupczewski, 2001). Using theoretical arguments, Koutsoyiannis (2004) and Rowinski and Strupczewski (2001) showed that the Gumbel distribution is quite unsuitable for application to hydrological extremes as it seriously underestimates the largest extreme flood heights or extreme rainfall amounts. Koutsoyiannis (2004) showed that long hydrological records (some decades of hydrological records) may display a misleading picture of a distribution which may suggest that the Gumbel distribution is of good fit when in fact it is not the appropriated model.

Koutsoyiannis (2004, with references therein) also questioned the appropriateness of the application of the Weibull (or Type III) distribution to extreme hydrological events due to its upper boundedness. The author, however, acknowledged that there is no general consensus in the appropriateness of the Weibull distribution and argued that “many [authors] regard an upper bound in natural quantities as reasonable” and that the Weibull distribution is the most

commonly found in nature. Koutsoyiannis (2004) went on to show, using "an extensive empirical investigation", that the Fréchet (or Type II) distribution which has a lower bound and unbounded from above is more consistent and appropriate for modelling flood heights or rainfall extremes. The application of the log-logistic distribution to extreme hydrological events has also been questioned in literature through theoretical arguments for its inability to model the tail part of the empirical distributions (Rowinski and Strupczewski, 2001). This existing gap gives room for other statistical extreme flood frequency distribution models to be explored in the main basins of Mozambique, in particular, the LRB, and compare with the prevailing Gumbel distribution for quantifying risks associated with extreme floods in the basin.

2.5 Statistical distributions commonly used in flood frequency analysis

Several probability frequency distribution models have been proposed in literature to model extreme floods or river flow peaks. In other words, the fitting of theoretical probability distributions to extreme flood events has received global attention in literature. In all FFA approaches the fitting of a probability distribution to the data is required (Ferreira and de Haan, 2015; Smithers, 2012). The methods used to estimate the parameters of the distributions are summarised in literature as methods of moments (MOM), maximum likelihood estimator (MLE), L-moments, probability weighted moments (PWM), Bayesian inference and non-parametric methods (Ferreira and de Haan, 2015; Smithers, 2012). According to Smithers (2012, with references therein) the L-moments method is reported to have relatively less bias as compared to other techniques and as a result its use in probability distribution fitting has been extensively covered in literature. The MLE and the PWM are also commonly used in applications and have both recently received a lot of attention in theoretical liter-

ature (Dombry, 2015; Ferreira and de Haan, 2015).

Haktanir et al. (2013) assessed the right-tail prediction ability of some distributions by Monte Carlo analysis. The Gumbel, three-parameter log-normal (LN3), GEV, three-parameter gamma (Ga3) and three-parameter log-gamma (LGa3) were compared from the aspect of predicting the right-tail quantiles of return periods in the range of years from finite-length samples by a Monte Carlo analysis. The authors estimated the parameters by MOM, MLE, PWM, zero-sample skewness and self-determined PWM. The study revealed that Ga3 estimated by PWM (or simply Ga3-PWM) performed better, followed by the LN3-MOM, LN3-PWM, Ga3-MOM, GEV-MOM and LN3-MLE for the ranges covered.

Abida and Allouze (2008) studied the probability distribution of flood flows in Tunisia. They used linear moments to identify regional flood frequency distributions for different zones in Tunisia. The selection of distributions was achieved through goodness-of-fit (GoF) comparisons in L-moment based regional test that compares observed to theoretical values of L-skewness and L-kurtosis for various candidate distributions. The distributions used by Abida and Allouze (2008) are among the most frequently used distributions in hydrological extreme variables, that is, GEV, Pearson type 3 (P3), generalised logistic (GLO), generalised normal (GN) and generalised Pareto distribution (GPD). Tunisia was divided into two homogeneous regions in the study. The GLO was appropriate for Northern Tunisia while Central/Southern Tunisia was best fitted by GLO and GEV. The findings in Abida and Allouze (2008) revealed distinguished spatial trends with respect to the best fit flood frequency distribution.

In Bangladesh, Ahammed and Hewa (2012) developed hydrological tools using

extreme rainfall data for Dhaka. They performed statistical analysis of annual daily maximum rainfall of Dhaka through the use of the Gumbel distribution function and scaling theory to produce the probable return periods of extreme rainfall events and intensity duration frequency (IDF) curves, respectively. The authors recommended the outcomes of their paper to be used in better designing of hydraulic structures in Bangladesh.

In South Africa several authors have studied extreme floods. Alexander (2002) demonstrated that widespread and severe floods were caused by infrequent, but not rare, meteorological phenomena including tropical cyclones. The author concluded that direct statistical analysis methods seriously under-estimate the frequency of occurrence of extreme floods. He recommended the widely used log-Pearson type 3 (LP3) distribution, which uses conventional moment estimators, as the preferred method for the statistical analysis of hydrological and meteorological data in South Africa. Pointing to the gap that exists between research and practice, Alexander (2002) remarked that different countries use different statistical analyses to estimate flood discharges in their flood design guidelines; for example, Australia, Canada, the UK and the USA, each uses a different statistical method.

Mélice and Reason (2007) studied the return period of extreme rainfall at George in South Africa. The authors used the longest available daily rainfall series and the Gumbel distribution to conclude that the 01 August 2006 flood was an extremely rare event. Nortje (2010) used a relatively new approach termed the Regional Estimation of Extreme Flood Peaks by Selective Statistical Analysis (REFSSA) to analyse data within similar hydrological regions. The suitability of the REFSSA method was demonstrated for the estimation of extreme flood peaks with very low annual exceedance probabilities of between 0.001 and 0.0001. The author suggested that the method would be more appli-

cable to climates experiencing outlier type of extreme flood peaks.

Vogel et al. (1993a) argued that many investigators have suggested alternative models such as GEV as an improvement to the LP3 used in Australia and USA. Vogel et al. (1993a) used flood flow data at 383 sites in Southwestern USA to explore the suitability of various flood frequency models using L-moment diagrams. The authors also repeated the experiment performed in the original Water Resources Council report (Bulletin 17B, issued in 1982) which led to the LP3 mandate. Among the candidate distributions considered were LP3, GEV, LN2, LN3, P3, normal and Gumbel distributions. Their findings revealed that the LP3, GEV, LN2 and LN3 provide a good approximation to flood flow data in the Southwestern USA region. Other models such as P3, normal and Gumbel distributions were shown to perform poorly. The main concluding remarks by Vogel et al. (1993a) were that index-flood procedures need not be restricted to the GEV distribution only because other models such as the LN2, LN3 and LP3 appear to offer suitable alternatives, for instance, in the Southwestern USA.

In another similar article Vogel et al. (1993b) used L-moment diagrams to select a set of suitable flood frequency distributions for modelling annual maxima flood flows in Australia. Among the distributions considered GPD, LP3, LN3, GEV and Wakeby provided acceptable approximations to the probability distribution structure of flood flows in Australia. Vogel et al. (1993b) reported that the GPD provided the best description of the distribution of flood flows in the most densely populated regions of Southwestern coastal region, while the GEV provided the best approximation to the empirical distribution flood flows in the winter-dominated rainfall regions of Tasmania and Southwest coast. The other flood frequency distributions LP3, LN3 and Wakeby also performed credibly well across the regions considered. Vogel et al. (1993b) also reported that P3, normal, Gumbel, uniform and exponential distributions performed poorly

in the study. Similar to Vogel et al. (1993a), these authors revealed that index-flood procedures should not be restricted to a single distribution such as GEV because other distributions such as GPD and LN3 perform significantly better in most densely populated regions of Australia. In the present study, the researcher follows the recommendation by both Vogel et al. (1993a) and Vogel et al. (1993b) by considering several candidate distributions before restricting the study to GEV or GPD.

Mujere (2011) analysed the frequency of Nyanyadzi River floods in Zimbabwe using the Gumbel distribution, arguing that although meteorological forecasts such as FFEWS provide accurate short term forecasts they do not often give enough time for disaster preparedness to reduce the impact of flood events. The author also reported that the incidence of false alarms in the region makes people not to take meteorological forecasts seriously thereby justifying the need to complement these forecasts with flood frequency models.

Mujere (2011) defined FFA as the process of fitting of a probability distribution model to a sample of historically recorded annual flood peaks for a particular site or catchment in a given region. The author argued that reliable flood frequency estimates are important for floodplain management which include protection of the public against loss of life and property, minimising flood related costs to government and private enterprises, designing and locating hydraulic structures and assessing hazards related to the development of flood plains. The author reiterated that although several studies have employed several statistical distribution models to explain the likelihood and intensity of floods none of these models have gained worldwide acceptance and is specific to a particular country or site. The author's views are also shared by Singo et al. (2012) and Olofintoye et al. (2009). Mujere (2011) attributed reasons for applying the Gumbel distribution to Nyanyadzi River flows to the homogeneous

and independent nature of peak flow which is free from long-term trends, unregulated river flow that is not affected by reservoir operations, diversions and urbanisation and long record good quality flow data. The MOM was used to estimate the parameters and the Chi-square test was used for GoF test leading the author to recommend the Gumbel distribution as a reliable distribution model for the Nyanyadzi floods.

In the present study, the researcher will extend the work of Mujere (2011) by considering several candidate distributions including the Gumbel and GEV as well as considering several parameter estimating techniques such as MLE, PWM, Bayesian inference and many others. The Chi-square test will also not be used in this thesis for the goodness-of-fit test, instead, the researcher will use the Anderson-Darling, Kolmogorov-Smirnov and diagnostic probability plots as goodness-of-fit tests.

Smithers et al. (2001) assessed the severity of the February 2000 floods from a probabilistic perspective, spatial variability of the extreme precipitation and flooding which occurred in north-eastern parts of South Africa, Mozambique and Zimbabwe. The analysis was performed for events ranging from 1 to 7 days in duration using the GEV distribution for the Sabie River catchment part of South Africa. The findings by Smithers et al. (2001) revealed that the floods experienced in the Sabie catchment in February 2000 were a result of unusual rainfall with return period in excess of 200 years in some parts of the catchment. The results also revealed that the extent of the extreme rainfall was associated with longer durations. Smithers et al. (2001) reported that many gauging stations did not function and several gauging structures were inundated due to the high magnitudes of the February 2000 floods. The spatial variability of the return periods of the simulated runoff depths was reported to vary greatly within the Sabie catchment. The study by Smithers et al. (2001)

shares similarities with the current study in that the February 2000 floods also affected the Limpopo River and that the Sabie catchment is part of the upper LRB. The current study extend the duration of the rainfall events to include 10 days and 30 days moving sums.

In a recent study similar to some parts of the present study, Blain and Meschiatti (2014) used multi-parameters distributions Wakeby, Kappa and GEV to assess the return probability of extreme rainfall data. The authors compared the performance of the Wakeby, Kappa and GEV by fitting the annual maximum of daily, 2-day and 3-day rainfall amounts collected over the period 1890-2012 obtained from Campinas weather station in São Paulo, Brazil. The presence of serial autocorrelation and climate trends were tested using autocorrelation function (ACF) and Mann-Kendall tests and the results showed no presence of serial autocorrelation and climate trends in the rainfall series considered. Based on the GoF tests, these authors concluded that the Kappa and GEV distributions provided the best fit to the series. In the present study, the GEV distribution together with other candidate distributions are used and the flood height annual maxima data series is extended to the 5-day, 7-day, 10-day and 30-day.

Another interesting study is that of Singo et al. (2012) who studied the extreme floods of the Livuvhu River catchment in South Africa which is part of the middle Limpopo River basin and the Livuvhu River is a tributary of the Limpopo River. Singo et al. (2012) used the GEV, Gumbel, LN2 and LP3 to fit the historical annual maximum data recorded for the Livuvhu River catchment. These authors concluded that the Gumbel and LP3 distributions provided the best fit to the data used in the study. The present study also considers the distributions used in Singo et al. (2012) in addition to other distributions for the lower Limpopo River of Mozambique. It will be of great interest to find out if the

conclusions reached in Singo et al. (2012) are consistent with the downstream Limpopo River in Mozambique.

Most recently Izinyon and Ehiorobo (2015) performed a FFA of Owan River at Owan site in Benin Owena River basin in Nigeria using the GEV, GPD and GLO. The L-moments method was used to estimate the parameters of the distributions and four GoF tests for the candidate distributions were used, namely the root mean square error (RMSE), relative root mean square error (RRMSE), mean absolute deviation index (MADI) and probability plot correlation coefficient (PPCC). These authors concluded that the GPD provided the best fit to the annual maximum flood series at the site. The authors further remarked that the L-moments and L-ratios are convenient summary statistics when analysing sample flood series.

Sukla et al. (2014) performed a FFA of the daily rainfall amount in the Mahanadi Delta region in India. The data used by the authors were daily rainfall volume data collected during the rainy season for a period of 28 years. Sukla et al. (2014) fitted 17 probability distribution models to the data and tested for their GoF using Kolmogorov-Smirnov, Anderson-Darling and Chi-square tests. These authors noted that the GPD and GEV emerged as the best and second best distributions at all the sites considered. They concluded that the GPD prevailed as the best distribution for modelling daily rainfall data at the study site.

In a separate study in India, Gohil and Chowdhary (2013) used the annual peak discharge data for the Tan River at Amba station in Gujarat to perform the FFA for the river. The authors used four statistical methods for their analysis namely, Foster method, Gumbel distribution, Ven Te Chow method and LP3 distribution. These authors found the Gumbel distribution to outperform the rest of the distributions and concluded that the Gumbel distribution may

be recommended for practical use at the site.

Alam and Khan (2014) noted that hydrologists use different methods to select the best distribution at a particular site or region because of the existence of a large number of distributions and also possibly due, partly, to availability of advanced computer software packages. Alam and Khan (2014) performed a FFA of five peripheral rivers Buriganga, Turag, Tongi, Balu and Lakhaya around Dhaka city using data for annual maximum, annual minimum water level and discharge volumes. These authors fitted the data with 62 probability distributions and tested for their GoF using Kolmogorov-Smirnov, Anderson-Darling and Chi-square. The overall rank of a distribution was obtained by considering the ranks from each of the GoF tests and taking the median of the three values as the overall rank of the particular distribution.

The study by Alam and Khan (2014) revealed that Dagum (4P) (i.e. 4-parameter Dagum distribution) was the best fit for annual maximum water level, Chi-square (2P) was the best fit for annual minimum water level, Cauchy best fitted the annual maximum discharge and Johnson B best fitted the annual minimum discharge. The frequently used distributions in FFA namely, GEV, LP3, log-normal and Gumbel performed poorly when compared against the best fit distributions. The authors recommended the method they used to have potential applications in other rivers in Bangladesh and other countries. In the present study the researcher will consider 10 candidate distributions commonly used in FFA and test their GoF using Kolmogorov-Smirnov, Anderson-Darling and probability diagnostic plots. Unlike Alam and Khan (2014), the present study will apply the Chi-square method for GoF testing and will also not use the median to obtain the overall rank, instead the researcher will use the total of the ranks from the two tests to obtain the overall rank of a particular distribution.

Neykov et al. (2014) examined stochastic daily precipitation models which use two components: the binary logistic regression to model the occurrence of the data and the other component models the intensity series using the continuous-valued distributions such as log-normal, gamma or Weibull. These standard models then use the joint density and standard software for generalised linear models to model the precipitation series. Neykov et al. (2014) noted that these standard models tend to underestimate the frequency of extreme rainfall events. In an attempt to overcome these shortcomings, Neykov et al. (2014) used hybrid distributions namely hybrid gamma-generalised Pareto and hybrid Weibull-generalised Pareto to develop a stochastic precipitation model for daily rainfall at Ihtiman in Western Bulgaria. The authors used simulation results to compare the models based on hybrid distributions and those based on standard distributions. These authors also reported on the difficulties that could be encountered with the hybrid models when using hourly precipitation data.

Kochanek et al. (2013) performed a thorough data-based comparison of the FFA methods used in France. These authors gave two distinct families of FFA approaches existing in France; the first involving the estimation of parameters of a pre-specified distribution (commonly either the Gumbel or GEV distribution) and the second approach uses a continuous simulation "where a rainfall generator is coupled with a rainfall-runoff model to generate long hydrological series from which extreme quantiles can be inferred" (Kochanek et al., 2013, p.4447). Within both families of approaches parameter estimation can be done at local scale, regional scale or local-regional scale (i.e. a combination of both local and regional scale). These authors noted that unlike other countries that have national FFA guidelines, France does not have a specific national FFA guideline that is officially recommended or prescribed by regulation. Kochanek et al. (2013) argued that the diversity of approaches in FFA in France raises ques-

tions concerning the optimal ambits of each family of approach and for that reason these authors performed a national comparison of the predictive performances of the main FFA implementations used in France.

Kochanek et al. (2013) considered eight main implementations that correspond to local, regional and local-regional parameter estimation of the Gumbel and GEV, as well as the local and regional estimation of the continuous simulation approach. The authors used daily river flow data from over 1,000 gauging stations in France to check the predictive performance of the eight competing implementations in terms of their stability and reliability. These authors found that two implementations emerged to dominate their competitors in their predictive performances. The two dominating implementations were the local-regional version of a GEV distribution and the local version of the continuous simulation approach. Kochanek et al. (2013) made the following three important specific conclusions:

- "the Gumbel distribution is not suitable for Mediterranean catchments, since this distribution demonstrably leads to an underestimation of flood quantiles;
- the local estimation of a GEV distribution is not recommended, because the difficulty in estimating the shape results in infrequent predictive failures;
- all the purely regional implementations evaluated in [the] study displayed a quite poor reliability, suggesting that prediction in completely ungauged catchments remains a challenge." (Kochanek et al., 2013, p.4446, see also p.4464 for more information).

Kochanek et al. (2013) showed further that the local Gumbel or mixed local-regional Gumbel or GEV perform relatively well in oceanic-influenced catch-

ments. The findings in Kochanek et al. (2013) can add value to the present study. The present study uses data from the LLRB whose catchments are also oceanic-influenced and the part of the basin is characterised by frequency tropical cyclones resulting in extreme floods.

One other important aspect in FFA is the incorporation of paleo-flood and historic peaks in the analysis. The incorporation of historic flood peaks and/or paleo-flood (i.e. past or ancient floods that occurred before the systematic gauged record) in FFA has received considerable attention in literature (Nguyen, 2009; van den Brink and Können, 2009; Reis and Stedinger, 2005; Baker, 2003). Reis and Stedinger (2005) asserted that these studies generally show that the use of historical information can be of great value in reducing uncertainty in flood quantile estimators. However, several authors including Reis and Stedinger (2005) and Sho et al. (2000) have raised concern related to measurement and recording errors in historic flood peaks and paleo-flood. In the present study, the researcher will not use paleo-flood or historic flood peaks for two reasons: the possible presence of measurement and recording errors in the ancient floods and that the historic flood peaks and paleo-flood data are not readily available in the basin under study.

2.6 At-site and regional flood frequency analyses

There are two fundamental paradigms in FFA, namely at-site and regional FFA (Smithers, 2012). At-site FFA refers to direct frequency analysis of observed river flows which include selecting and fitting a suitable theoretical probability distribution model to the observed data at a single site. On the other hand, regional FFA involves pooling data from similar and nearby locations (or other local basins within the same region) when performing FFA in

order to get efficient estimates of parameters of the chosen distribution.

A lot of criticism has been levelled against the use of at-site FFA (Smithers, 2012, with references therein). The at-site FFA approach is only highly recommended to be performed for large sample sizes (or long records of river flows or rainfall amounts) (Smithers, 2012). In just over two decades ago, Vogel et al. (1993b) pointed out that most studies were skewing towards the use of regional index-flood procedures which are presumed to be more accurate and robust than at-site FFA type procedures.

A study by Abida and Allouze (2008) clears the mist on the appropriate use of at-site (or single site) flood frequency analysis and regional frequency analysis approaches. At-site or single site FFA is appropriately performed when extensive historic peak-flow data are available whereas regional FFA is appropriately used when there is little or no historic flow data at a particular site. It should also be noted that regional FFA is also equally appropriate for large sample sizes of historic data records. Furthermore, in regional analysis, "... all data from other local basins within the same region are pooled to get an efficient estimate of parameters of a chosen distribution and hence a more robust quantile estimate" (Abida and Allouze, 2008).

Smithers (2012, p.635, with references therein) gave a list of the limitations of at-site (or direct) statistical approach:

- "The correct distribution of the flood peaks is unknown and different probability distributions may give acceptable fits to the data, but result in significantly different estimates of design floods when extrapolated.
- The records of gauged runoff are generally short and the calibration of the gauging structures may not be very robust. Hence the sample only represents a small distribution of the floods at the site and the fitted dis-

tribution may be further biased by gauging errors.

- The frequency of flood-producing rainfalls and land-use characteristics may have changed during the period of historical measurement.
- The fitted distribution does not explicitly take into account any changes in the runoff generation process for higher magnitude events."

Although these limitations are meant to refer to at-site FFA, a critical analysis of these limitations would show that only the second bullet point refers directly to at-site FFA, the rest of the bullet points may also refer to regional FFA. One of the main advantages of regional FFA over at-site FFA is its ability to tackle the issue of short or no historic flow data at a particular site under study. Thus when the historic flow records are long or sample size is large, the at-site FFA is very comparable to the regional FFA in performance (Smithers, 2012).

In regional FFA it is assumed that the standard variate is uniformly distributed throughout all sites in the selected region and this assumption allows data from a particular region to be combined to form "a single regional flood, rainfall, frequency curve that is applicable anywhere in the region with appropriate site-specific scaling." (Smithers, 2012, p.365,with refrences therein). When performing regional FFA of short records of annual maximum flood heights or rainfall the shape parameter of the parent distribution is estimated from the pooled regional data while the scale parameter is estimated locally from the site (Smithers, 2012, with references therein).

In the context of FFA, regionalisation refers to the identification of regions that are homogeneous in their response to flood events and the selection of a suitable frequency distribution for the identified region (Rostami, 2013; Smithers, 2012). Regional FFA can be defined as the concept of supplementing ungauged or time-limited sampling records by incorporating spatial randomness and use

data from different sites in the homogeneous region (Smithers, 2012). When dealing with ungauged sites precaution must be taken to ensure that the site comes from the region in which the method was developed or outside the range of observations from which the method was developed.

It is postulated that in almost all practical situations regional FFA outperforms at-site FFA and numerous respected authors generally agreed that regionalisation greatly improves flood quantile estimation (Rostami, 2013; Smithers, 2012; Cordery and Pilgrim, 2000; Hosking and Wallis, 1997; Alexander, 1990). Cordery and Pilgrim (2000) stated that a well conducted regionalisation would lead to improved flood prediction. This was echoed by Alexander (1990) who argued that regional FFA provides a basis for improved estimates of parameters of both ungauged sites and gauged sites with short historical records.

Smithers (2012) stated that geographic proximity does not imply that there is hydrologic similarity. In these recent years it is argued that much research has now focused towards the identification of homogeneous regions. The index flood-based L-moments method developed by Hosking and Wallis (1997) is reported to be a robust method for identifying homogeneous regions. The method uses a cluster analysis of site characteristics to identify regions that are potentially homogeneous, and this enables the independent testing of the at-site data for homogeneity (Smithers, 2012). According to Smithers (2012, citing Hosking and Wallis, 1997) the L-moments based methods are used to estimate the frequency of floods, screen for discordant data and test clusters for homogeneity.

Following a literature review by Kachroo et al. (2000), a recent extensive literature by Smithers (2012) showed that there is no universally accepted method of regionalisation. In South Africa, Alexander (1990) noted that there were no

comprehensive studies on regional statistics of extremes methods made since the early 1970s. The author outlined a generalised regional statistical analysis procedure which is based on plotting scaled growth curves and rejecting those stations whose growth curves are inconsistent with the remaining stations. Alexander (1990) argued that overseas methods used to identify homogeneous regions were not valid in South Africa where the distribution of gauging stations is too sparse for the pre-determination of hydrologically homogeneous regions.

Smithers (2012) criticised Alexander's generalised regional statistical analysis approach citing inconsistencies in results and duplication of efforts by users. Nortje (2010) developed the REFSSA method in South Africa which uses similar recorded annual maximum flood peaks or flood heights from a catchment to define 'similar hydrological regions' or homogeneous regions. The homogeneous regions are transformed to the site that is being investigated "in proportion to the ratio of the square root of the respective catchment areas." (Smithers, 2012). However, (Nortje, 2010) cautioned against using the REFSSA method outside the regions where it was developed and recommended the method to be appropriate for regions that experience one or two extreme peaks in a historical record which is typical in most catchments in Southern Africa. The region considered in the present study is a typical region that experiences one or two extreme outliers in the hydrological record. The approach by Nortje (2010) is also subject to inconsistencies in results arising from subjectivity in delineating homogeneous regions (Smithers, 2012).

Mkhandi et al. (2000) used the L-moments to identify regional distributions in their FFA of Southern Africa including Mozambique. The authors identified 41 homogeneous regions and concluded that the possible underlying distributions from the L-moments of the observed data were P3, LN3, GPD and GEV.

However, simulation studies indicated that P3-PWM and log-Pearson type 3 (LP3) estimated by MOM, were suitable procedures for the Southern African region. Most recently Rostami (2013) used the L-moments approach to perform a regional FFA of west Azarbayjan province basins. The author used the Wald hierarchical cluster method to identify homogeneous regions. Four homogeneous regions were identified by Rostami (2013) and the best fitting distributions per region were GPD for region A; GEV, P3 and log-normal for B; P3, log-normal and GEV for C; and log-normal and GEV for D. It is clear that the GEV is prevalent in all regions except for region A which is an indication of its dominance in FFA.

In this section the two fundamental approaches in FFA, at-site and regional FFA, were extensively reviewed. The limitations of each of the two fundamental approaches were outlined, notably small sample size records or ungauged sites for at-site approach, and methods for the identification of 'similar hydrological regions' for regional FFA. The advantages of regional FFA over at-site (or direct methods) are evident from the majority of the studies reviewed.

2.7 Modelling observed extreme flood events using extreme value theory

In the previous chapter we defined extreme value theory (EVT) as the branch of statistics that deals with extremely low or high values in the distribution tails. EVT is categorised into two widely used fundamental approaches namely, block maxima and peaks-over-threshold (POT) (Ferreira and de Haan, 2015). The block maxima approach consists of organising the data into equal-sized nonoverlapping periods (known as blocks) and restricting attention to the selection of the maximum of the observations in each block (Ferreira and de Haan, 2015). When dealing with hydrological data such as flood heights a block is

naturally a year (i.e. observations are naturally blocked by years). The new observations formed (i.e. the block maxima) follow approximately an extreme value distribution, G_ξ for some real ξ, under extreme value conditions (Ferreira and de Haan, 2015; Farada et al., 2011; Coles, 2001). Without loss of generality, the GEV distribution arises naturally when dealing with block maxima, especially for large sample sizes (Dombry, 2015; Ferreira and de Haan, 2015; Coles, 2001).

In the POT approach in EVT a certain high threshold is determined and attention is restricted to the selection of those observations from the initial data set that exceed the given threshold (Ferreira and de Haan, 2015, 2014; Beirlant et al., 2004; Coles, 2001). The new observations (i.e. the exceedances) follow approximately a GPD, under extreme value conditions (Ferreira and de Haan, 2015, 2014). In both block maxima and POT approaches parametric (and sometimes nonparametric) statistical techniques for the extreme value distributions are applied to the selected observations (Ferreira and de Haan, 2015).

According to Ferreira and de Haan (2015), the block maxima approach is the older one (Gumbel, 1958) while the POT approach was developed much later by Pickands (1975) who also supplied the theoretical framework and formulated the statistical tools. However, although the block maxima is the older of the two approaches, it has not been thoroughly studied as much research has been devoted to POT (Ferreira and de Haan, 2015; Farada et al., 2011). Most recently Ferreira and de Haan (2015) performed an extensive review of literature of the comparison in performance between block maxima and POT, and came up with the following two main conclusions based on the reviewed literature: for large sample sizes, the POT and block maxima are comparable in performance, and secondly, when the number of exceedances is larger than the number of blocks (at least 1.65 times the number of blocks for $\xi = 0$) the POT

method is comparably more efficient in performance than the block maxima in a number of circumstances.

Several authors have recommended EVT as the most powerful and reasonably robust framework to use when modelling the behaviour of the tails of a wide-ranging class of distributions (Ferreira and de Haan, 2015; Sigauke, 2014; de Wet et al., 2012; de Wet, 2004; Gencay and Selcuk, 2004; Coles, 2001). In the present study, the tail behaviour of extreme flood heights is modelled using GEV, GPD and other families of distributions. A comprehensive overview of EVT is given in a number of books and research journal articles (Dombry, 2015; Dombry and Ribatet, 2015; Ferreira and de Haan, 2015, 2014; Cooley, 2009; Reiss and Thomas, 2007; de Haan and Ferreira, 2006; Beirlant et al., 2004; Coles, 2001; de Haan, 1970; Fisher and Tippett, 1928). The conditions for block maxima and POT are presented in the following subsections together with GEV and GPD, respectively.

2.7.1 Block maxima and generalised extreme value distribution

The block maxima approach, having been almost neglected in theoretical research, has received a lot of attention in recent years from a number of respected authors in the field (Bücher and Segers, 2016; Dombry, 2015; Ferreira and de Haan, 2015; Farada et al., 2011). Most recently Ferreira and de Haan (2015) revisited the block maxima approach and showed theoretically that the block maxima approach is more efficient than the POT approach in terms of lower asymptotic variances of the EVI and quantile estimators. These authors also noted that the asymptotic minimal mean square error is lower for block maxima as compared to POT, while the optimal number of blocks is generally lower than the number of exceedances. These theoretical findings by Ferreira and de Haan (2015) signify a major breakthrough in block maxima research

which has often been regarded as less efficient in literature in favour of the POT approach.

When dealing with block maxima the two widely used estimators are MLEs and PWM estimators (Ferreira and de Haan, 2015). Most recently Dombry (2015) proved the theoretical existence and consistency of MLEs within the block maxima framework, while Ferreira and de Haan (2015) proved the consistency of PWM estimators within the block maxima framework.

The theoretical framework of the estimators and their properties is reviewed as follows. Consider $(X_i)_{i\geq 1}$ iid random variables with common distribution function $F \in D(G_\xi)$ and corresponding normalisation sequences of constants $a_m > 0$ and b_m such that

$$\lim_{m\to+\infty} F^m(a_m x + b_m) = G_\xi(x), \quad x \in \mathbb{R}, \tag{2.3}$$

where ξ is the extreme value index (Dombry, 2015; Coles, 2001). In other words, it can be said that the distribution function F satisfies the extreme value condition with extreme value index ξ (Dombry, 2015; Ferreira and de Haan, 2015). The necessary and sufficient conditions that F must satisfy are found in de Haan and Ferreira (2006, chapter 1) and the revised conditions based on PWM within the block maxima framework are in Ferreira and de Haan (2015). The latter conditions shall be presented shortly in this section.

Consider $M_{k,m} = \max(X_{(k-1)m+1}, ..., X_{km}),\ k \geq 1$,, this implies that $n = m \times k$ observations are divided into k blocks of size m, where n is the total number of observations. Then there exists a non-degenerate function G such that for a fixed block length $m \geq 1$ the variables $(M_{k,m})_{k\geq 1}$ are iid with distribution

function F^m such that

$$P\left(\frac{M_{k,m}-b_m}{a_m}\leq x\right)=\lim_{m\to+\infty}F^m(a_mx+b_m)\to G_\xi(x),\ \text{ as }\ n,k\to+\infty, \tag{2.4}$$

where $G_\xi(x)$ is the extreme value distribution given by

$$G_\xi(x)=\exp(-(1+\xi x)^{-1/\xi}),\quad \xi\in\mathbb{R},\ 1+\xi x>0. \tag{2.5}$$

According to Ferreira and de Haan (2015) it is usually taken for granted that the block maxima follow exactly (or very well) an extreme value distribution, which is not always true. The maxima only follow an extreme value distribution G_ξ approximately when the sample sizes are large. Ferreira and de Haan (2015) accounted for this misspecification by quantifying it in terms of the second-order condition in order to account for the bias that may arise in distribution specification since $G_\xi(x)$ is not the exact distribution.

When the blocks are of large size, $G_\xi(x)$ approximates the GEV distribution (Dombry, 2015),

$$G(\mu,\sigma,\xi;x)\ =\ \begin{cases}\exp\left(-\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right), & \text{for } 1+\xi\left(\frac{x-\mu}{\sigma}\right)>0,\ \xi\neq 0,\\ \exp\left(-\exp(-\frac{x-\mu}{\sigma})\right), & x\in\mathbb{R},\ \ \xi=0,\end{cases} \tag{2.6}$$

where μ,σ and ξ are the location, scale and shape parameters, respectively. In (2.6), the distribution is the heavy-tailed Fréchet (also called type II) class of distributions if $\xi>0$, the short-tailed Weibull (or type III) class of distributions if $\xi<0$, the light-tailed Gumbel (or type I) class of distributions if $\xi=0$ (Coles, 2001). In the previous sections the researcher has reviewed literature which showed how common these distributions are used in FFA. The extensive use of the GEV distribution families in FFA was revealed in the previous sections. Two important theorems concerning the GEV distributions follow (Coles,

2001):

Theorem 2.1. *(Non-degeneracy) (Coles, 2001, p.48).*
If there exist sequences of constants $a_m > 0$ *and* b_m *such that*

$$P\left(\frac{M_{k,m} - b_m}{a_m} \leq x\right) \to G(x), \quad \textit{as} \quad n = m \times k \to +\infty, \tag{2.7}$$

for a non-degenerate distribution function G*, then* G *is a member of the GEV family in (2.6).*

The proof of Theorem 2.1 is trivial and is given in Coles (2001, p.48). The theorem can be interpreted as suggesting the application of the GEV family of distributions for modelling the distribution of block maxima of long data sequences.

The next theorem pertains to max-stable distributions.

Theorem 2.2. *(Max-stability) (Coles, 2001, p.50).*
A distribution is max-stable if, and only if, it is a generalised extreme value distribution in (2.6).

The proof of Theorem 2.2 is given in (Coles, 2001, p.50).

Attention is now drawn to the two conditions within the block maxima framework known as first-order and second-order conditions in the univariate case (Ferreira and de Haan, 2015).

CONDITION 2.1**:** (First-order condition) (Ferreira and de Haan, 2015, p.279). If (2.4) and (2.5) are combined, then (2.4) can be re-written as

$$\lim_{m\to\infty} \frac{1}{m} \frac{1}{-\log F(a_m x + b_m)} = (1 + \xi x)^{1/\xi}, \tag{2.8}$$

this is an equivalence to the convergence of the inverse functions

$$\lim_{m\to\infty} \frac{V(mx) - b_m}{a_m} = \frac{x^\xi - 1}{\xi}, \quad x > 0, \tag{2.9}$$

with $V = (-1/\log F)^{\leftarrow}$. Hence, we can choose b_m to be $V(m)$ (Ferreira and de Haan, 2015). This is called the first-order condition. Ferreira and de Haan (2015) gave a second order condition (expansion) in their analysis of block maxima.

CONDITION 2.2**:** (Second-order condition) (Ferreira and de Haan, 2015, p.279-280). Consider some positive function a and some other positive or negative function A with $\lim_{t\to\infty} A(t) = 0$,

$$\lim_{t\to\infty} \frac{(V(tx) - V(t))/a(t) - (x^\xi - 1)/\xi}{A(t)} = \int_1^x s^{\xi-1} \int_1^s u^{\rho-1} du ds = H_{\xi,\rho}(x), \tag{2.10}$$

$\forall x > 0$, with function $|A|$ regularly varying with index $\rho \leq 0$ (Ferreira and de Haan, 2015; de Haan and Ferreira, 2006).

In order to obtain the PWM estimators within the block maxima framework the procedure is as follows. Suppose $X_{1,k}, ..., X_{k,k}$ is the order statistics of block maxima $X_1, ..., X_k$. The statistics $\beta_0 = \frac{1}{k}\sum_{i=1}^k X_{i,k}$ and

$$\beta_r = \frac{1}{k}\sum_{i=1}^{k} \frac{(i-1)...(i-r)}{(k-1)...(k-r)} X_{i,k}, \quad r = 1, 2, 3, ..., \ k > r, \tag{2.11}$$

are unbiased estimators of $EX_1F^{rm}(X_1)$ (Ferreira and de Haan, 2015, with references therein). The PWM estimators of the shape parameter ξ, as well as the location parameter b_m and scale parameter $a_m = a([m])$, are simple functionals of β_0, β_1 and β_2. The estimator $\hat{\xi}_{k,m}$ for ξ is defined as the solution of the

equations

$$\frac{3^{\hat{\xi}_{k,m}}-1}{2^{\hat{\xi}_{k,m}}-1}=\frac{3\beta_2-\beta_0}{2\beta_1-\beta_0}, \tag{2.12}$$

$$\hat{a}_{k,m}=\frac{\hat{\xi}_{k,m}}{2^{\hat{\xi}_{k,m}}-1}\frac{2\beta_1-\beta_0}{\Gamma(1-\hat{\xi}_{k,m})}, \tag{2.13}$$

and

$$\hat{b}_{k,m}=\beta_0+\hat{a}_{k,m}\frac{1-\Gamma(1-\hat{\xi}_{k,m})}{\hat{\xi}_{k,m}}, \tag{2.14}$$

where $\Gamma(x)=\int_0^\infty t^{x-1}e^{-t}dt,\ \ x>0$ (Ferreira and de Haan, 2015, with references therein).

The following theorem is the basis for the analysis of estimators in block maxima approach (Ferreira and de Haan, 2015). This is known as the asymptotic normality for the univariate case within the block maxima framework. Let $\lceil u \rceil$ represent the supremum of u (i.e. the least upper bound greater than or equal to u).

Theorem 2.3. *(Asymptotic normality) (Ferreira and de Haan, 2015, p.280). Assume that $F \in D(G_\xi)$ and that Condition 2.1 holds. Furthermore, assume that Condition 2.2 also holds. Let $m=m_n\to\infty$ and $k=k_n\to\infty$ as $n\to\infty$, in such a way that $\sqrt{k}A(m)\to\lambda\in\mathbb{R}$. Let $0<\varepsilon<1/2$ and $\{X_{i,k}\}_{i=1}^k$ be the order statistics of the block maxima $X_1,X_2,...,X_k$. Then, given that $\{E_k\}_{k\geq 1}$ is an appropriate sequence of Brownian bridges (Ferreira and de Haan, 2015),*

$$\begin{aligned}\sqrt{k}\left(\frac{X_{\lceil ks\rceil,k}-b_m}{a_0(m)}-\frac{(-\log s)^{-\xi}}{\xi}\right) &= \frac{E_k(s)}{s(-\log s)^{1+\xi}}+\sqrt{k}A_0(m)H_{\xi,\rho}\left(\frac{1}{-\log s}\right)\\ &\quad+\left(s^{-1/2-\varepsilon}(1-s)^{-1/2-\xi-\rho-\varepsilon}\right)o_P(1),\end{aligned} \tag{2.15}$$

as $n\to\infty$, where the $o_P(1)$ term is uniform for $1/(k+1)\leq s\leq k/(k+1)$ (Ferreira and de Haan, 2015). The functions $a_0(m)$ and $A_0(m)$ are as defined in Ferreira and de Haan (2015, Lemma 4.2). The proof of Theorem 2.3 is referred to Ferreira and de Haan (2015, Section 4).

As for the MLEs method within the block maxima approach, consider the distribution of $M_{k,m}$, for large k (the number of blocks or years in this case) and $m \geq 1$, (see (2.4)) which is approximately a GEV distribution with parameters (a_m, b_m, ξ). According to Dombry (2015), it is common practice to estimate these unknown parameters by the MLEs method. The log-likelihood for large k or a large sample of size $n = m \times k$ (i.e. $M_{1,m}, ..., M_{k,m}$) is

$$L_k(\mu, \sigma, \xi) = \frac{1}{k} \sum_{i=1}^{k} \ell_{(\mu,\sigma,\xi)}(M_{k,m}). \tag{2.16}$$

According to Dombry (2015) L_k does not have a global maximum and therefore $(\hat{\mu}_k, \hat{\sigma}_k, \hat{\xi}_k)$ is an MLE if L_k has a local maximum at $(\hat{\mu}_k, \hat{\sigma}_k, \hat{\xi}_k)$. The MLEs are obtained from the solution of the likelihood equations

$$\nabla L_k = 0 \quad \text{with } \nabla L_k = \left(\frac{\partial L_k}{\partial \mu}, \frac{\partial L_k}{\partial \sigma}, \frac{\partial L_k}{\partial \xi} \right). \tag{2.17}$$

The MLE is any solution of (2.17) with a negative definite Hessian matrix (Dombry, 2015). The emphasis in both Dombry (2015) and Ferreira and de Haan (2015) is to control the block size, that is, to let the block size (or length of block) depend on the sample size in order to meet the asymptotic requirements. In the present study the block size is naturally one year since it is natural for floods to follow annual maxima. The major findings concerning the existence of consistent MLEs by Dombry (2015) are summarised in the following theorem.

Theorem 2.4. *(Existence of consistent MLEs) (Dombry, 2015, p.4-5).*
Suppose $F \in D(G_\xi)$ *with* $\xi > -1$ *and assume that*

$$\lim_{n \to +\infty} \frac{m(n)}{\log n} = +\infty. \tag{2.18}$$

Then there exists a sequence of estimators $(\hat{\mu}_n, \hat{\sigma}_n, \hat{\xi}_n)$ *and a random integer*

$N \geq 1$ *such that*

$$\mathbb{P}[(\hat{\mu}_n, \hat{\sigma}_n, \hat{\xi}_n) \ \textit{is a MLE for all} \ \ n \geq N] = 1 \tag{2.19}$$

and

$$\hat{\xi}_n \to \xi, \ \ \frac{\hat{\mu}_n - b_m}{a_m} \to 0 \ \textit{and} \ \ \frac{\hat{\sigma}_n}{a_m} \to 1 \ \ a.s. \ \textit{as} \ \ n \to +\infty. \tag{2.20}$$

The condition for $\xi > -1$ has been recommended by several authors and the likelihood in (2.17) has no solution for $\xi < 1$ and no consistent MLE exists when $\xi < -1$ (Dombry, 2015, with references therein). In general, both MLEs and PWM are said to be consistent for $-1 < \xi < 1/2$, which is a usual range in many applications (Dombry, 2015; Ferreira and de Haan, 2015).

Bücher and Segers (2016) established the asymptotic normality of the MLE in the multivariate case within the block maxima framework. Bücher and Segers (2016) used the multivariate approach to show that the three parameter GEV distribution family is not differentiable in quadratic mean. The following proposition (Proposition 2.7.1) and its proof in Bücher and Segers (2016) show the existence of strong consistency of MLEs for the GEV distribution within the block maxima framework.

Proposition 2.7.1. *(Consistency) (Bücher and Segers, 2016, p.6).*
Let $X_1, X_2, ...$ *be an independent random sample from the* G_{θ_0} *distribution, with* $\theta_0 = (\mu_0, \sigma_0, \xi_0) \in (-1, \infty) \times \mathbb{R} \times (0, \infty)$. *Then there exists a compact set* $\Theta \subset (-1, \infty) \times \mathbb{R} \times (0, \infty)$ *such that* θ_0 *is in the interior of* Θ *and for any estimator sequence* $\hat{\theta}_n$ *such that* $\mathbb{P}_n \ell_{\hat{\theta}_n} = max_{\theta \in \Theta} \mathbb{P}_n \ell_\theta$*, such maximisers always existing, we have* $\hat{\theta}_n \to \theta_0$ *almost surely as* $n \to \infty$.

The detailed proof of this proposition is given in Dombry (2015) Theorem 2 and Bücher and Segers (2016, p.6-7).

Proposition 2.7.2. *(Asymptotic normality) (Bücher and Segers, 2016, p.7). Let $X_1, X_2, ...$ be independent and identically distributed (iid) random variables with common GEV distribution G_{θ_0}, with $\theta_0 = (\mu_0, \sigma_0, \xi_0) \in (-1/2, \infty) \times \mathbb{R} \times (0, \infty)$. Then for any compact parameter set $\Theta \subset (-1/2, \infty) \times \mathbb{R} \times (0, \infty)$, any sequence of MLEs over Θ is strongly consistent and asymptotically normal:*

$$\sqrt{n}(\hat{\theta}_n - \theta) = I_{\theta_0}^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \dot{\ell}_{\theta_0}(X_i) + o_{\mathbb{P}}(1) \rightsquigarrow N_3(0, I_{\theta_0}^{-1}), \quad n \to \infty. \tag{2.21}$$

The block maxima has been thoroughly reviewed in this section including the existence of consistency from commonly used estimators such as MLEs and a discussion of the formulation of the PWM estimators in the block maxima framework in Condition 2.2. The following section reviews the POT approach as well as a comparative discussion of the two approaches.

2.7.2 Peaks-over-threshold and generalised Pareto distribution

Until most recently the POT method was regarded as more efficient than the block maxima method discussed in the previous section because of its ability to use more of the available data than block maxima since the number of exceedances is greater than the number of blocks for most of the situations (Ferreira and de Haan, 2015, with references therein). However, most recent studies have shown through functional analysis theory that the block maxima approach is more efficient than POT when the sample size is large (Ferreira and de Haan, 2015). The set up for the POT method is the following.

Consider $X = X_1, ..., X_n$ to be iid random variables representing an extreme event e.g. flood heights. Let F be the distribution function (commonly unknown) of the extreme event X, then the conditional excess $(X - u)$ of the

distribution function is

$$\begin{aligned} F_u(y) &= P(X-u \leq y | X > u) \\ &= \frac{P(X-u \leq y \text{ and } X > u)}{P(X>u)} \\ &= \frac{F(y+u)-F(u)}{1-F(u)}, \quad 0 \leq y \leq x_F - u, \end{aligned} \tag{2.22}$$

where u is the threshold, $y = x - u$ are the excesses and $x_F < \infty$ is the right endpoint of F (Scarrott and MacDonald, 2012; Magadia, 2010; Coles, 2001).

Based on Pickands (1975) and Balkema and de Haan (1974) theorems, the conditional excess distribution function F, for large enough u is well approximated by a GPD

$$H(\sigma, \xi; y) = \begin{cases} 1 - \left(1 + \frac{\xi}{\sigma} y\right)^{-1/\xi}, & \text{for } \xi \neq 0,\ 0 \leq y \leq x_F - u, \\ 1 - \exp(-\frac{y}{\sigma}), & \text{for } \xi = 0, \end{cases} \tag{2.23}$$

where σ and ξ are the scale and shape parameters, respectively. These parameters are usually estimated by the MLE, PWM and other parameter estimating methods. When $\xi < 0$, the distribution in (2.23) has an upper end point, i.e. it is bounded from above, when $\xi > 0$ the distribution is bounded from below and it belongs to the ordinary Pareto family, while if $\xi = 0$ the distribution is exponential (or belongs to the Gumbel family), (Ferreira and de Haan, 2014; Magadia, 2010; Beirlant et al., 2004; Coles, 2001). The complete formulation of the Pickands-Balkema-de Haan result is given in Beirlant et al. (2004, Proposition 2.1, p.73; with references therein). In EVT the GPD plays the role of a natural distribution model for excesses over a reasonably high threshold. In other words, the GPD plays the same role for POT as the GEV plays for block maxima.

In the previous sections in this chapter several areas of applications of the GPD were reviewed: ranging from flood events, electricity demand and supply, to environmental and life sciences, among others (Dombry and Ribatet, 2015; Ferreira and de Haan, 2014; Sigauke, 2014; Rajaram, 2006; Cooley, 2005). The GPD has recently been used in energy supply to model electricity daily peak demand (Sigauke, 2014). While modelling electricity demand for South Africa, (Sigauke, 2014) noted that estimating the parameters of a GPD by the frequentist methods such as PWM and MLEs may be complicated by the requirement of regularity conditions in some situations and thus suggested the use of Bayesian inference which does not depend on regularity conditions.

Smith (1987) studied the problem of violation of the regularity conditions when using the MLEs method to estimate the parameters of the GEV and GPD and came up with the following findings

- when $\xi < -1$, the MLEs are not likely to be obtainable,
- when $-1 < \xi < -1/2$, the MLEs are normally obtainable, but they do not have the usual asymptotic properties,
- when $\xi > 1/2$, the second and higher order moments cease to exist.
- when $\xi > -1/2$, the usual asymptotic properties of asymptotic efficiency, consistency and asymptotic normality hold for these distributions GEV and GPD (Sigauke, 2014; Rajaram, 2006).

From the results above, we can deduce that the regularity conditions can always be met with relative easy for the range $-1/2 < \xi < 1/2$ for the GPD and GEV. Dombry (2015) and Ferreira and de Haan (2015) have recently shown that the range $-1/2 < \xi < 1/2$ is permissable for both MLEs and PWM methods in theory which is also usually the situation in many practical applications.

While performing a theoretical comparison between block maxima and POT using PWM estimators, Ferreira and de Haan (2015) introduced PWM-POT estimators for ξ and $a(n/k)$ which can be taken as an approximate of σ, and k is the number of exceedances or selected order statistics $\{X_{n-i,n}\}_{i=0}^{k-1}$, from the universal sample $X_i, i = 1, 2, ..., n$. The following statistics

$$P_n = \frac{1}{k}\sum_{i=0}^{k-1} X_{n-i,n} - X_{n-k,n} \text{ and } Q_n = \frac{1}{k}\sum_{i=0}^{k-1} \frac{1}{k}\left(X_{n-i,n} - X_{n-k,n}\right) \tag{2.24}$$

are estimators for $a(n/k)(1-\xi)^{-1}$ and $a(n/k)(2(2-\xi)^{-1})$, respectively. As a consequent, the PWM estimators are

$$\hat{\xi}_{k,n} = 1 - \left(\frac{P_n}{2Q_n} - 1\right)^{-1} \text{ and } \hat{a}(n/k) = P_n\left(\frac{Pn}{2Q_n} - 1\right)^{-1}. \tag{2.25}$$

The quantile estimator is

$$\hat{x}_{k,n} = X_{n-k,n} + \hat{a}(n/k)\frac{(k/(np_n))^{\hat{\xi}_{k,n}} - 1}{\hat{\xi}_{k,n}}. \tag{2.26}$$

Asymptotic normality for PWM-POT estimators hold under the conditions equivalent to those of block maxima (Ferreira and de Haan, 2015). In the block maxima case, k is the number of blocks whereas in POT k is the number of selected high exceedances or top order statistics. Thus in both cases k refers to the number of observations selected. The choice of the number of blocks in block maxima method and the threshold choice in POT method both involve the same issue of a trade-off between bias and variance. In making theoretical comparisons between the PWM-block maxima and PWM-POT, Ferreira and de Haan (2015) confined the extreme value index ξ to the range $\xi \in [-1, 1/2)$, which is argued to be also the case in numerous practical applications. The present study shall also be guided by these asymptotic normality conditions.

Unlike in block maxima where the observations of interest (maxima or minima) are easily picked once a block is defined, in the POT approach the threshold needs to be expertly selected first. Selecting a desired threshold is not as easy a task as selecting a block. Several techniques have been suggested in literature to solve the problem of threshold selection (Scarrott and MacDonald, 2012, with references therein). The other issue in POT is that the exceedances are not iid if the original observations are dependent and thus a technique called declustering is required to make the exceedances iid before analysis is done. In the next two subsections these two issues, threshold section and declustering, are reviewed.

Threshold selection techniques

When modelling extreme events using the POT approach and GPD, the first step is to identify a technique for threshold selection. Once threshold selection is done then probability distribution fitting may commence after meeting the iid assumption requirements. Currently the common methods used for threshold selection include, but are not limited to, mean residual life plot, threshold choice or stability plot, L-moment plot, dispersion index plot, Pareto quantile plot, Hill plot and mixture models, among others (Scarrott and MacDonald, 2012; Magadia, 2010; Ribatet, 2006).

The main goal behind threshold section is to have a trade-off between bias and variance, that is, the threshold needs to be as small as possible in order to make the variance small, but at the same time it has to be as high as possible to avoid bias that may arise due to the selection of those among the initial observations that are not considered extreme points. In other words, the main goal in threshold selection is to satisfy the model asymptotic basis and reduce quantile estimation uncertainty (Reiss and Thomas, 2007; Beirlant et al., 2004; Coles, 2001).

Scarrott and MacDonald (2012) classified threshold estimation techniques as classical fixed, tail fraction, mixture models and re-sampling based techniques. Classical fixed approaches include the mean residual life plot, threshold choice plot, L-moment plot and dispersion index plot (Beirlant et al., 2004; Coles, 2001). These classical fixed approaches use the graphical plots to select a threshold and therefore have the benefit of allowing the practitioner to inspect the data graphically, understand the features and make an assessment of model fit (Scarrott and MacDonald, 2012). The disadvantages of the classical fixed methods are that a substantial amount of expertise is required and such methods can be very subjective. The other drawback is that once the threshold is selected it becomes fixed and in subsequent inferences this subjectivity or uncertainty is usually ignored and yet sensitivity analysis of the threshold may suggest a different choice of threshold (Bernadara et al., 2014; Scarrott and MacDonald, 2012).

The tail fraction estimation threshold selection techniques use graphical diagnostics that are based on asymptotic optimality (Scarrott and MacDonald, 2012). These include, among others, Pareto quantile plot, and the Hill plot (Scarrott and MacDonald, 2012; Berning, 2010; Goegebeur et al., 2008; de Haan and Ferreira, 2006; Beirlant et al., 2004). The main drawbacks of these methods include the complications associated with estimating parameters in higher order characteristics (i.e. power law of the function that is slowly varying) and the failure to account for (or explain) threshold uncertainty in the event of subsequent inferences (Scarrott and MacDonald, 2012). In line with the tail fraction estimation threshold selection techniques, Berning (2010) developed a threshold estimation technique based on a measure which quantifies the stability of the parameter estimates over a range of thresholds. The threshold selection technique developed by Berning (2010) can be used to obtain a range of thresholds which is characterised by the most stable estimates.

The re-sampling based techniques can also be referred to as computational approaches and these are algorithms mainly based on bootstrapping methods and adaptive Hill estimators (Scarrott and MacDonald, 2012; Beirlant et al., 2004). These methods are not widely applicable due to their restrictive assumptions (Scarrott and MacDonald, 2012).

The main disadvantage of most of the threshold estimation techniques aforementioned is that they cannot account for the associated threshold choice uncertainty. In order to overcome this problem mixture models were proposed in the last decade (Scarrott and MacDonald, 2012). This procedure considers both the distribution of excesses and the bulk distribution (i.e. the distribution of non-exceedances or non-extreme events). 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. In most cases the mixture models naturally account for uncertainty associated with threshold choice in subsequent inferences (Scarrott and MacDonald, 2012). The major drawbacks of the mixture models are that the asymptotic properties of their impromptu heuristics are not yet well understood, their behaviour at the threshold point is still unclear (i.e. the question of whether the fitted density is continuous) and whether the bulk and tail distribution fits are comparatively robust to each other. More details on mixture models and other threshold choice techniques are obtained in Bernadara et al. (2014, with references therein), MacDonald (2012, with references therein) and Scarrott and MacDonald (2012, with references therein).

In this thesis the commonly used classical fixed techniques will be applied since the recently developed mixture models still leave a lot to be desired in terms of their computational abilities. Among the classical fixed techniques, the present study will apply the mean residual plot and the threshold stability plot concur-

rently to identify a threshold for a particular data set (Yilmaz et al., 2014). In order to chose a threshold, u, using the mean residual life plot, the plot ought to approximate linearity in u above a high threshold, u_0, at which the GPD gives a valid approximation to the distribution of excesses (Yilmaz et al., 2014; Coles, 2001). When using the threshold stability (or choice) plot the basis for the choice of a threshold is that at the threshold, u_0, at which the GPD gives a valid approximation to the distribution of excesses, the shape parameter ξ should approximate a constant value, while the estimates of the scale parameter should approximate linearity in u (Coles, 2001). The subsection that follows reviews the declustering techniques associated with the POT approach and GPD distribution.

Declustering techniques

One of the shortcomings of POT approach is that the threshold exceedances are not usually iid if the original observations are dependent and they occur in clusters, that is, the exceedances have a dependence structure (Bernadara et al., 2014; Coles, 2001). In order to achieve the assumption of iid, a technique called declustering is used (Bernadara et al., 2014; Ribatet, 2006; Coles, 2001). This technique involves filtering the dependent exceedances from the clusters to obtain a data set of approximately independent threshold excesses. According to Coles (2001) declustering works as follows:

- it uses an empirical rule to make a definition of clusters of exceedances;
- it identifies the maximum excess in each of the clusters;
- it assumes that cluster maxima are independent, with the conditional excess distribution modelled by the GPD;
- it fits the GPD to the cluster maxima.

According to Ribatet (2006), with the aid of R statistical programming software, the clusters are are identified as follows: the first exceedance value starts the first cluster, the first observation below the threshold ends the cluster, unless the time condition for independence (denoted tim.cond in programming) does not hold, the next exceedance initiates a new cluster unless the time condition for independence does not hold, and the process is repeated as required. The declustering approach has been criticised in literature for being sensitive to the run length, r, that is chosen arbitrarily in cluster determination, and the discarding of all data to leave out cluster maxima only is also thought as wastage of data (Beirlant et al., 2004; Coles, 2001).

The issue of the time condition for independence has various views. In Ribatet (2006), the time allowed between two flood events to be considered independent is 8 days, while recently Yilmaz et al. (2014) used 24 hours as the time condition for two flood events to be considered independent, i.e., POT values a day prior and a day after the flood event are removed from the series. Ferro and Segers (2003) developed what is now famously known as the Ferro-Segers declustering technique which is based on the estimation of the extremal index. The method by Ferro and Segers (2003) uses an automatic selection of the run length, r, used to identify the independent clusters. More details on the Ferro-Segers method are given in Chapter 7. Recently Bernadara et al. (2014) discussed the extremal index method and defined it as representing the reciprocal of the mean size of the clusters of the events. Literature is abundant on the application of extremal index on autocorrelated data (Bernadara et al., 2014; Beirlant et al., 2004; Coles, 2001; Leadbetter et al., 1983).

In this thesis the researcher will apply the extremal index procedure for declustering based on Ferro-Segers algorithm. More details on the method are given in Chapter 7.

2.8 Bayesian modelling of extreme floods

In Bayesian analysis of extreme floods it is possible to combine local or at-site flood data with additional information such as historic floods temporal information, spatial information on neighbouring basins or catchments floods, flood processes causal information from experts and many other information from experts who have past experience or from related studies (Viglione et al., 2013). This additional information based on expert past experience or previous studies is modelled through the introduction of a prior distribution $\pi(\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a parameter vector, that is, $\boldsymbol{\theta} = \{\theta_1, ..., \theta_p\} = \{\mu, \sigma, \xi\}$ in this thesis. The priors are grouped into informative and non-informative priors. Priors based on expert information are called informative priors, while the other priors which do not depend on expert information are called non-informative priors.

After the first step of taking the prior information into account, the second stage in Bayesian analysis involves the collection of data to form the likelihood function denoted by $\pi(\boldsymbol{x}|\boldsymbol{\theta})$ where $\boldsymbol{x} = \{x_1, ..., x_k\}$ is the observations vector. The likelihood in Bayesian analysis describes how the vector of observations, $\boldsymbol{x}$, depends on the parameter vector $\boldsymbol{\theta}$ (Viglione et al., 2013; Renard et al., 2006; Kuczera, 1999). The final step in Bayesian analysis combines the likelihood and the prior distribution to form the posterior distribution. The posterior distribution is computed using Bayes' Theorem

$$\pi(\boldsymbol{\theta}|\boldsymbol{x}) \quad = \quad \frac{\pi(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int_{\boldsymbol{\Theta}} \pi(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) d\boldsymbol{\theta}} \tag{2.27}$$

which is usually written as

$$\pi(\boldsymbol{\theta}|\boldsymbol{x}) \quad \propto \quad \pi(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) \tag{2.28}$$

where $\boldsymbol{x}$ is a vector observations, $\boldsymbol{\theta}$ is a parameter vector, $\pi(\boldsymbol{\theta})$ is the prior dis-

tribution, $\pi(\boldsymbol{\theta}|\boldsymbol{x})$ is the posterior distribution, $\pi(\boldsymbol{x}|\boldsymbol{\theta})$ is the likelihood function, $\pi(\boldsymbol{x})$ is the normalisation constant and $\boldsymbol{\Theta}$ is the space parameter.

The prediction of future values of the vector of observations, $\boldsymbol{x}$, can be done using the posterior predictive distribution $\pi(x_{k+1}|\boldsymbol{x})$ given by

$$\pi(x_{k+1}|\boldsymbol{x}) = \int_{\boldsymbol{\Theta}} P(\boldsymbol{x_{k+1}}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{x})d\boldsymbol{\theta}. \tag{2.29}$$

The provision of the full posterior distribution of the parameters in Bayesian framework is considered to be one of the main advantages of Bayesian approach over frequentist methods, and this implies that no approximation is needed to evaluate uncertainties (Reis and Stedinger, 2005; Kuczera, 1999). It is also argued that the credible intervals for parameters provided by a Bayesian framework are much easier to interpret as compared to the confidence intervals in frequentist approach (or classical statistics) (Reis and Stedinger, 2005). In simpler cases it is possible to choose a conjugate prior in order to obtain an analytically traceable posterior and in such cases the normalisation constant is easily computed (Reis and Stedinger, 2005, with references therein). However, in more complicated cases where the parameter vector, $\boldsymbol{\theta}$, is large it can be computationally infeasible to compute the normalisation constant and such complications had hindered the application of Bayesian approach for a long time. This problem was facilitated by the rapid development of computers in the last three decades and the introduction MCMC methods (Gaume et al., 2010; Kwon et al., 2008; Renard et al., 2006; Reis and Stedinger, 2005; Kuczera, 1999; Tierney, 1994).

According to Reis and Stedinger (2005, p.101) MCMC algorithms are able to sample the "values of the parameters from the posterior distribution without computing the [normalisation] constant". Two most popular MCMC algorithms are the Gibbs sampler and Metropolis-Hastings algorithms (Reis and

Stedinger, 2005; Tierney, 1994). More details on the theoretical description of MCMC are given in Tierney (1994) and Reis and Stedinger (2005, with references therein).

Reis and Stedinger (2005) performed a Bayesian MCMC flood frequency analysis to evaluate the posterior distributions of flood quantiles, flood risk and parameters of the log-normal and LP3. The authors showed that Bayesian MCMC provides a complete representation for measurement and discharge errors of large flood records and historical flood information. These authors showed that the traditional non-informative priors based upon the Fisher Information matrix are limited in application. The study by Reis and Stedinger (2005) revealed the superiority of Bayesian MCMC over other methods including MLEs method in providing a full description of uncertainty in quantiles and parameters.

One of the merits of Bayesian inference is that it does not depend on regularity conditions, which is often the case in classical statistics (Reis and Stedinger, 2005; Martins and Stedinger, 2000). In a study on hydrologic data, Martins and Stedinger (2000) used the generalised maximum likelihood (GML) with a Bayesian prior distribution to estimate the parameters of the GEV distribution and showed that the GML estimator performed substantially better than the MLEs, MOM and L-moment quantile estimators for $-0.4 \leq \xi \leq 0$.

Gaume et al. (2010) used data sets from Slovakia and South of France with some historic flood events at ungauged sites in a RFFA approach. These authors used the Bayesian MCMC to estimate the values of parameters of the regional distributions. They argued that the Bayesian MCMC method was chosen due to its rigorousness. The results in Gaume et al. (2010) showed that the Bayesian MCMC is able to substantially narrow down the confidence intervals

provided the ungauged extremes come from the comprehensive sampling over the region selected. These results were also shown to be consistent with at-site flood frequency results based on paleo-flood and historic flood information. In a separate study, Viglione et al. (2013) showed that when using Bayesian analysis, the uncertainty in the estimates reduce substantially with an increase in prior information used.

Gaioni et al. (2010) defined the flood height or water level as not only a function rainfall and snow melt (where applicable), but also a function of the geometry of the river and other numerous surrounding characteristics such as soil permeability, and extent of human development in the surrounding area. The authors acknowledged that quantification of these characteristics and incorporation into a mathematical model is very challenging in some instances, and the data is also not readily available in some other instances. These authors encouraged the use of information from an expert who is familiar with river flow in such cases. In most cases however, such experts may also not be available or they may not have sufficient information to give meaning to the desired study. Gaioni et al. (2010) propose a novel approach which starts from the assessment of the features of the model. According to Gaioni et al. (2010, p.75) the method "starts from the quantiles of the parametric model, translates them into values of the parameters of interest, and uses them to specify a prior distribution."

The present study will use Bayesian MCMC approach to estimate the parameters of the GEV distribution for both at-site and regional flood frequency analysis for the LLRB of Mozambique using flood height data. The next section reviews the modelling of extreme flood events in the presence of covariates.

2.9 Modelling nonstationary extremes in the presence of covariates

In statistics of extremes most of the mathematical arguments are based on the assumption that the underlying process consists of a sequence of iid random variables. Stationarity is the most natural generalisation that encompasses a sequence of independent random variables. According to Coles (2001), stationarity is a more realistic assumption to a large number of physical processes and it refers to a physical process whose random variables may be mutually dependent, but have stochastic properties that are constant (or homogeneous) through time. A great number of stationary series satisfies the property that, two events $X_i > u$ and $X_j > u$, are approximately independent if u is large enough and the separation between the time points i and j is large.

Emil Gumbel, (Gumbel, 1941, p.71), the pioneer in the application of statistics of extremes, made the following comment on stationarity: "In order to apply any theory we have to suppose that the data are homogeneous, i.e., no systematical change of climate and important change in the basin have occurred within the observation period and that no such change will take place in the period for which such extrapolations are made."

Indeed even in those ancient times, Gumbel was concerned that the then recently developed EVT techniques could apparently, only be applied under the stationarity assumption (Katz, 2010). It was quite obvious that this strong assumption would not hold much longer with the changes in geographic features over time and variability in climatic conditions. Stimulated by Gumbel's work and his book on statistics of extremes (Gumbel, 1958), extensive development has since been made in statistics of extremes dating from the time of the quote of Gumbel (Katz, 2010). Several changes have occurred in recent studies in-

cluding the recent introduction of nonstationary models (Yilmaz et al., 2014; Coles, 2001). A recent international conference on international disaster and risk studies (IDRC) held in Davos in 2014 concluded that the frequency and intensity of natural disasters are increasing and are anticipated to continue to increase in the near future (Stal et al., 2014).

Nonstationary processes are characterised by changing systematically through time. When dealing with environmental processes such as floods, nonstationarity is often due to seasonality (possibly due to climate changing with months) and trends (possibly attributed to long-term climate changes) (Coles, 2001). In the case of nonstationarity, standard models in block maxima and POT are usually modified to accommodate the changes in the parameters. For instance, the GEV model with a quadratic trend in the location parameter will be denoted by $G(\mu_t, \sigma, \xi)$ where $\mu_t = \mu_0 + \mu_1 t + \mu_2 t^2$ (Yilmaz et al., 2014; Coles, 2001).

Yilmaz et al. (2014) argued that the increased frequency and magnitude of floods and other extreme events such as high temperatures and severe rainfall, mainly attributed to climate change, put to question the assumption of stationarity. Yilmaz et al. (2014) used nonstationary GPD models to investigate trends and other nonstationarity characteristics in extreme rainfall events, as well as investigating the potential climate change impacts and variability on intensity-frequency-duration (IFD) relationships at an observation station in Melbourne, Australia. These authors found a statistically significant trend to be present in the storm durations of half-an hour, 3 hours, and 48 hours. However, there was no evidence on nonstationarity in the GPD models for all storm durations and Yilmaz et al. (2014) concluded that stationary GPD models could be used to model the extreme rainfall data at the site for all storm durations. These authors also found, through IFD analysis, that urban flash floods that produce hourly rainfall had increased over time for Melbourne.

A similar study involving IFD was conducted by Verdon-Kidd and Kiem (2015). These authors stated that, over the last 15 years, the validity of the assumption of climate stationarity has been questioned, particularly in Australia, where there is increased evidence of nonstationarity in annual maxima rainfall time series. Verdon-Kidd and Kiem (2015) demonstrated that IFD relationships depend on the length of the period of the rainfall data set used in the development of the IFD information, or in other words, IFD relationships are dependent on the sample size of the rainfall data series used in the design. These authors recommended for nonstationarity in annual maxima to be given serious consideration in ongoing revisions of Engineers Australia's guide for Australian rainfall and runoff and also recommended that clear guidelines be provided on how to deal with nonstationarity in extreme events. Katz (2010) recommended that in order to address the challenges associated with impact of climate change on extreme events increased collaboration is required between climate scientists and statisticians.

In a separate study concerning IFD curves, Cheng and AghaKouchak (2014) showed that in the presence of nonstationarity, a stationary assumption may substantially underestimate precipitation extremes by up to 60%, which consequently increases the chance of failure risk in infrastructure systems and flood risk. These authors used Bayesian inference to present a generalised framework that can be used for estimating nonstationary IFD curves and their uncertainties.

Statistics of nonstationary extremes has been applied in various fields and situations including extreme wave climate (Vanem, 2015a,b), surface level ozone (Eastoe and Tawn, 2009), climate change related flash floods (Velasco et al., 2013), economic development based on climate uncertainty (Arndt and Thurlow, 2014), trends in extreme rainfall indices (Saidi et al., 2013), annual cycle

of heavy precipitation (Otto et al., 2014; Maraun et al., 2009), climate data sparse regions (Tramblay et al., 2014), continental Spanish rivers climate and reservoir indices external covariates (López and Francés, 2013), spatial data exhibiting multidimensional covariate effects (Jonathan et al., 2014), climate covariates (Vasiliades et al., 2013), and annual maxima nonstationary random sequences (Ribereau et al., 2008).

Thomas et al. (2014) stated that the frequency of severe natural disasters over the last 40 years has been on the rise, particularly in Asia-Pacific countries. These authors investigated the extent to which the frequency of severe natural disasters is related to the increase in the number of people who are exposed to natural hazards, changes in the vulnerability of people to hazards, and precipitation and temperature extremes. These authors found that hydrometeorological intense natural disasters are strongly associated with extreme precipitation and rising population exposure, while climatological intense natural disasters are strongly associated with changes in temperatures. Thomas et al. (2014) generalised their conclusion by attributing the frequency of severe natural disasters observed in the region of Asia-Pacific to man-made climate change, since the climate change could be a result of greenhouse gas emissions that alter the climate system in the atmosphere.

Arnell et al. (2014) studied the impacts of climate change across the globe using a multi-sectoral assessment approach in a geoscientific framework. These authors found, using a narrative assessment, that approximately 450 million people would be exposed to increased frequency of river flooding, 1.3 million additional people would be flooded in coastal floods every year, the demand for residential energy would be reduced due to reduced heating demands and the production of crops would reduce in most regions. However, most of these global impacts on flooding and water stress would be experienced in Asia, and

this would extend to the Middle East and North Africa. The study by Arnell et al. (2014) also revealed that in 2050 the changes in temperature and sea level would still be similar and differ substantially in 2080.

In a separate study involving the LLRB of Mozambique, Aich et al. (2014) used a geoscientific approach to study the impacts of climate change on streamflow using four large African basins. The four basins included the Limpopo, Niger, Oubangui, and Upper Blue Nile. These authors used an eco-hydrological model named Soil and Water Integrated Model (SWIM) which they set up to each of the four basins. The validation of the models showed results that were adequately good. In order to assess the impact of climate, the authors compared "the trends in the mean discharges, seasonality and hydrological extremes in the 21^{st} century" (Aich et al., 2014, p.1305). The authors found the uncertainty of results to be high for all the basins, and the uncertainty of projections to be lowest in the Upper Blue Nile which is most likely to experience an increased streamflow. The Limpopo and Niger basins were found to experience high magnitudes of trends accompanied by a wide range of uncertainty. The results for Oubangui basin showed the least impact of climate change. The results by Aich et al. (2014) are of great interest to the researcher since the Limpopo River is also being studied in this thesis. The existence of high magnitude of trends in the mean and hydrological extremes, and wide range of uncertainty will be investigated in the nonstationary context using statistics of extremes in this thesis. The interest in the present study will be to find out if the two different approaches can yield similar findings that may lead to the same conclusions.

2.10 A brief review of r largest order statistics

There are two main generalisations or fundamental approaches in EVT; one based on exceedances over a high threshold and the other based on block max-

ima in which the r largest order statistics is selected from a block, for small r (Beirlant et al., 2004; Coles, 2001). The idea of r largest order statistics selected from a block implies that the standard block maxima approach in which only one observation is selected from a block (usually a year in floods) arises when $r = 1$ (Beirlant et al., 2004). One of the shortcomings of block maxima, which is usually evident with small samples, is that only a few observations are retained. This problem can be solved by selecting a small number of r largest observations from the block, for $r \geq 1$ (Coles, 2001). This approach, however, brings to question the issue of independence of observations, just as is the case with POT.

Consider $X_1, X_2, ...$ to be a sequence of iid random variables with common distribution function $F(x)$ such that $F(x) < 1 \ \forall x \in \mathbb{R}$ and suppose $M_n^{(r)} = r^{th}$ largest order statistics of $\{X_1, X_2, ...X_n\}$ for r a fixed positive integer and $n \geq r$. Thus, $M_n^{(r)}$ is the r^{th} largest of the n observations (Li and Tomkins, 1991). Then from Theorem 2.2, for r fixed, the following theorem arises.

Theorem 2.5. *(r largest order) (Coles, 2001, p.66-67).*
From the result of Theorem 2.2, let there be sequences of constants $a_n > 0$ and b_n such that

$$P\left\{\frac{M_n - b_n}{a_n} \leq x\right\} \to G(x), \quad \textbf{as } n \to +\infty$$

for some non-degenerate function G such that G is the GEV distribution function in (2.6), then for r fixed,

$$P\left\{\frac{M_n^{(r)} - b_n}{a_n} \leq x\right\} \to G_r(x), \quad \textbf{as } n \to +\infty$$

where

$$G_r(x) = \exp\{-\tau(x)\} \sum_{s=0}^{r-1} \frac{\tau(x)^s}{s!}, \tag{2.30}$$

with

$$\tau(x) = \left[1 + \xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-1/\xi}.$$

In Theorem 2.5, if the unknown scaling constants are absorbed into the location and scale parameters of the model, then for n large, it follows that the approximate distribution of $M_n^{(r)}$ is within the family of the distribution of (2.30) (Beirlant et al., 2004; Coles, 2001; Li and Tomkins, 1991). The situation in Theorem 2.5 is equivalent to having each of the largest r order statistics within each of the blocks, for example, when $r = 5$ for each of the blocks (or year in this thesis) we have 5 largest numbers for each year. Algebraically, the complete vector is

$$\boldsymbol{M}_n^{(r)} = (M_n^{(1)}, ..., M_n^r) \tag{2.31}$$

for each of the numerous blocks (Coles, 2001). Although Theorem 2.5 provides the approximate distribution for each of the distribution of $\boldsymbol{M}_n^{(r)}$ it does not give the joint distribution of $\boldsymbol{M}_n^{(r)}$. It can also be noted that the components are not independent e.g. $M_n^{(3)}$ cannot be great than $M_n^{(1)}$. The joint probability density function (pdf) is given in the following theorem.

Theorem 2.6. *(r largest order) (Coles, 2001, p.68).*
Again from Theorem 2.2, let $a_n > 0$ and b_n be sequences of constants, such that

$$P\left\{\frac{M_n - b_n}{a_n} \leq x\right\} \to G(x), \quad \textbf{\textit{as}}\ n \to +\infty$$

for some non-degenerate function G, then for r fixed, the limiting distribution, as $n \to +\infty$, of

$$\tilde{M}_n^{(r)} = \left(\frac{M_n^{(1)} - b_n}{a_n}, ..., \frac{M_n^{(r)} - b_n}{a_n}\right)$$

falls within the family of distributions with pdf

$$f(x^{(1)},...,x^{(r)}) = \exp\left\{-\left[1+\xi\left(\frac{x^{(r)}-\mu}{\sigma}\right)\right]^{-1/\xi}\right\} \times \prod_{k=1}^{r}\sigma^{-1}\left[1+\xi\left(\frac{x^{(r)}-\mu}{\sigma}\right)\right]^{-1/\xi-1}, \quad \forall\xi\neq 0 \tag{2.32}$$

where $-\infty<\mu<\infty,\ \sigma>0$ *and* $x^{(r)}\leq x^{(r-1)}\leq ...\leq x^{(1)}$
and $x^{(r)}: 1+\xi(x^{(k)}-\mu)/\sigma>0$ *for* $k=1,...,r$.

When $r=1$, (2.32) reduces to the GEV family density functions (Coles, 2001). In the case $\xi=0$, or the limiting form of (2.32) as $\xi\to 0$ is the family of density functions

$$f(x^{(1)},...,x^{(r)}) = \exp\left\{-\exp\left[-\left(\frac{x^{(r)}-\mu}{\sigma}\right)\right]\right\} \times \prod_{k=1}^{r}\sigma^{-1}\exp\left[-\left(\frac{x^{(r)}-\mu}{\sigma}\right)\right], \quad \forall\xi=0 \tag{2.33}$$

which reduces to the Gumbel family when $r=1$ (Coles, 2001). The proof of Theorem 2.6 is in Coles (2001, Chapter 7). The theoretical arguments on how far away from the maximum r can go are given in Beirlant et al. (2004). According to Coles (2001) the likelihood of the r largest order has parameters that correspond to those of the block maxima GEV. Thus, the parameters are interpreted in the same way as in block maxima GEV, although it is argued that the precision of estimates improves due to the inclusion of additional information (Coles, 2001).

The choice of r is based on a trade-off between variance and bias. Too large a value of r will violate the asymptotic theory leading to bias and too small a value of r would lead to large variance (An and Pandey, 2007; Beirlant et al., 2004; Coles, 2001). Extensive theoretical proofs for the complete stability of the large order statistics are provided in Li and Tomkins (1991). Detailed theoret-

ical stochastic comparisons of the largest order statistics using the generalised gamma distribution family which encompass Weibull, gamma and exponential random variables are provided in Kochar and Torrado (2015). Zhao and Balakrishnan (2015) furthered the topic of stochastic comparisons of largest order statistics to multiple-outlier gamma models. These authors, Zhao and Balakrishnan (2015), made their comparisons in terms of different stochastic ordering that included the likelihood ratio order, hazard rate order, star order and dispersive order. They also presented a general sufficient condition for the star order.

In terms of application of the r largest order statistics, just like block maxima, the r largest order statistics has found many applications which include, among others, sea level (Coles, 2001), wind speed (An and Pandey, 2007), and wave height (Soares and Scotto, 2004). Coles (2001) applied the r largest order statistics to the Venice sea level data and reported that the standard errors of the estimates decrease with increasing values of r, which corresponds to increased precision. Coles (2001) argued that if the asymptotic approximation is valid for any choice of r then the estimates of the parameters ought to be stable when the model is fitted with reasonably fewer order statistics. Coles (2001) noted that there was no evidence of stability for estimates of the parameters of Venice sea level data and concluded that the validity of the model may need further scrutiny for values of $r \geq 5$. The diagnostic checks also indicated lack of fit of the model for the Venice sea level data.

An and Pandey (2007) used annual r largest order statistics to model extreme wind speed. They used the joint GEV distribution of r largest order statistics derived from the Poisson process theory. These authors used the MLEs method to estimate the parameters of the distribution of annual wind speed. Based on the findings in Tawn (1988) that the results for the range $r = 3$ up to 7 are very

stable and that the method gives very consistent results when r is within that range, An and Pandey (2007) chose $r = 5$ for the analysis of the wind speed data. The wind speed data used in their study was collected at 30 stations in Ontario, Canada. An and Pandey (2007) concluded, using the r largest order statistics, that the wind speed data for Ontario could be suitably modelled by a Gumbel distribution.

In Portugal, Soares and Scotto (2004) applied the r largest order statistics for the extreme wave height long-term predictions. These authors used the limiting joint GEV distribution of r largest order statistics model to estimate the return levels of extreme wave height. Soares and Scotto (2004), based on the same arguments as in An and Pandey (2007) and Tawn (1988) raised in the preceding paragraph, with the support of some sensitivity analysis results based on the likelihood ratio test, chose $r = 5$ for the analysis of the wave height data. Soares and Scotto (2004) secured the independence of the observations by adopting a method that filters the observations to extract the r largest independent values. The findings in Soares and Scotto (2004) revealed that the estimates of the parameters and quantiles of the joint GEV were more accurate than the classical annual block maxima method. These authors concluded in favour of the joint GEV distribution model based on the r largest order statistics ahead of the classical annual block maxima GEV model due to its limited number of data points.

The study of flood heights in Mozambique using both block maxima and r largest order statistics may indeed add value in terms of increasing the precision of the estimates of the parameters, (i.e. if the r largest order statistics is worth enough to apply in the basin). The theoretical arguments raised in this section on the limiting joint GEV distribution of the r largest order statistics are very important in complementing the application of the approach in this

thesis.

2.11 Summary of the chapter

In this chapter extensive relevant literature was reviewed. In the first part of the chapter the existing methods in LLRB of Mozambique were reviewed, and the new and existing methods in other countries, particularly developed countries, were discussed. Implementations from worldwide disaster conferences were discussed as well as some disaster risk reduction efforts being made worldwide to reduce the deleterious effects of these natural disasters on humans and property.

The literature ranged from discussion of the fundamental approaches of FFA to fundamental approaches of EVT. Regarding the two fundamental approaches of FFA, at-site and regional flood frequency analysis, the approaches were clearly distinguished, their merits and demerits were reviewed and probability fitting methods used to fit the data in each of the approaches were thoroughly reviewed. The parameter estimation techniques used in these FFA approaches were also discussed. The probability models currently used in the LLRB were reviewed as well as the FFEWS in the basin. A lot of gaps were exposed in literature including the scarcity (or lack of) statistical models in the LLRB under study.

In terms of the two fundamental approaches in EVT, the block maxima and POT approaches were extensively reviewed. The recent articles by Bücher and Segers (2016), Dombry (2015), Ferreira and de Haan (2015) and Ferreira and de Haan (2014) form the backbone of the theories applied in this thesis. Although the main focus of this thesis is in application of these recently developed techniques (or revisited), the theorems stated and proved in this thesis help our

understanding of the theoretical developments which greatly improves our application of the methods. The r largest order approach, which is an extension of the block maxima was also extensively reviewed. Based on theoretical arguments of Ferreira and de Haan (2015) it was concluded that the block maxima approach is more efficient in a number of situations as compared to the POT approach. This result by Ferreira and de Haan (2015) is a major turning point in current research as majority of the previous research would claim that the POT method is more efficient than the block maxima since it optimises the use of available data. Literature reviewed in this thesis also tends to favour the r largest order approach ahead of the classical block maxima which often result in limited sample size.

The convergence of block maxima and POT in functional spaces using both PWM and MLEs methods in the univariate case were discussed using examples from Dombry (2015), Ferreira and de Haan (2015) and Ferreira and de Haan (2014). Asymptotic convergence in multivariate case was also extensively reviewed with examples from Bücher and Segers (2016, with references therein). Bayesian MCMC modelling was also reviewed in this chapter with a view to apply it in order to account for more information in the model. Bayesian inference has the advantage of not depending on regularity conditions, which is often the case with classical methods.

In general, literature reviewed in this chapter has revealed several gaps in theory and applications. Several authors have advocated for the improvement of the existing methods in literature and the need to evaluate the new methods adopted in other countries (usually developed countries), with the view that these methods may also work in the developing countries' rivers or catchments, given different operating characteristics and climatic conditions.

Chapter 3

Investigating the goodness-of-fit of ten candidate distributions and estimating high quantiles of extreme floods in the lower Limpopo River Basin of Mozambique.

3.1 Introduction

The 21^{st} century has been characterised by an unexpected number of natural disasters, for instance, earthquakes in Nepal in 2015, floods in Mozambique in 2013, flooding and landslides in Brazil in 2011, a combination of the gi-

ant earthquakes and devastating tsunamis in Japan in 2011, Haiti in 2010, Indonesian Islands in 2004 and devastating floods in Mozambique and most parts of Southern Africa in 2000 (CNN, 2015; Wikipedia, 2015; Smithers, 2012; WMO, 2012; Maree, 2011; Smithers et al., 2001). The most common devastating natural hazard in Southern Africa is flooding which causes great economic damages in the region (MunichRe, 2013, 2011). In Chapter 2 an extensive literature on flood frequency analysis (FFA) was reviewed including the two fundamental methods of FFA; at-site and regional, and the two fundamental approaches of extreme value theory (EVT); block maxima and peaks-over-threshold (POT). In this chapter the focus is on fitting the candidate distributions and making comparisons based on their goodness-of-fit (GoF).

Hosking and Wallis (1997) defined FFA as an estimation of how often a particular event, usually extreme, will occur. Flood height magnitude is a very important hydrologic parameter in, for instance, water resources engineering, storm water management, floodplain control and urban planning. Reliable estimation of extreme flood heights and their frequency of occurrence are essential for the proper design of hydraulic structures in and across a river and it helps in the identification of flood risk area (Izinyon and Ehiorobo, 2015). Reliable estimation of the magnitude and frequency of extreme flood heights at the site of interest also help in the design of hydraulic structures such as dams, drains, spillways, irrigation ditches and culverts (Izinyon and Ehiorobo, 2015; Abida and Allouze, 2008). The advantages of FFA given here are some of the main reasons behind the motivation for this chapter and the thesis in general.

A crucial problem in hydrological studies is the choice of a frequency distribution function that fits the extreme flood series in a particular river basin or region (Abdul-Karim and Chowdhury, 1995). This is despite the improved understanding of hydrological processes and advanced computer programming

software (Smithers, 2012). Several probability frequency distribution models have been developed to describe the frequency distribution of extreme floods (Izinyon and Ehiorobo, 2015; Alam and Khan, 2014; Blain and Meschiatti, 2014; Neykov et al., 2014; Sukla et al., 2014; Gohil and Chowdhary, 2013; Haktanir et al., 2013; Baratti et al., 2012; Smithers, 2012; Hosking and Wallis, 1997). However, major problems arise when selecting the best distribution to use since there is no general agreement on which method or distribution should be used for the flood frequency analysis of extreme hydrological events (Singo et al., 2012; Smithers, 2012; Olofintoye et al., 2009; Vogel et al., 1993a,b). The selection of an appropriate distribution, therefore, depends solely on the characteristics of available data at a particular site or catchment (Smithers, 2012; Olofintoye et al., 2009). The emphasis in this gap (or problem) is summarised in a quotation found in Izinyon and Ehiorobo (2015, p.1) "Many probability distributions have been suggested in the hydrological and statistical literature to model extreme hydrological events but no particular model is considered superior for all practical applications hence WMO (2009) suggests that available models be screened based on the problem to be solved and nature of available data". The present situation faced by the researcher in this chapter is to exploit this existing gap in literature and identify appropriate flood frequency distributions for the three sites in the LLRB of Mozambique, given that there is scarcity of literature on distribution fighting in the basin.

Most recently Izinyon and Ehiorobo (2015) performed a FFA of annual maximum flood peaks at Owan site along the Owan River in Benin Owena River basin of Nigeria in an attempt to determine the best fit probability distribution that is applicable to the site. The GEV, GLO and GPD whose parameters were estimated by the L-moments method were used as candidate distributions. These authors tested the GoF of the candidate distributions using RMSE, RRMSE, MADI and PPCC, and the GPD emerged as the best fit distri-

bution for the site.

Baratti et al. (2012) investigated the estimation of flood frequency distribution at annual and seasonal time scales. These authors designed an approach that distinguishes between annual and seasonal distributions, where the annual distribution is the product of the seasonal cumulative distribution functions. Baratti et al. (2012) applied the approach to the Blue Nile River daily flows at the Ethiopia-Sudan border to estimate the seasonal and annual flood quantiles. The seasons were divided into dry season, pre-flood season, flood season and post-flood season. According to these authors the method performed well when tested through sensitivity analysis. This procedure allows users to introduce subjective weights into the components of the objective function. However in the view of the researcher, this subjectivity may pose problems to the method with users ending up having completely different results for the same data if their weights are different.

In another distribution fitting study, Sukla et al. (2014) fitted 17 types of probability distributions to the daily rainfall amount in Mahanadi Delta of Odisha in order to find an appropriate probability distribution for the delta. The fitting of such a large number of distributions was facilitated through the use of a software called EasyFit 5.5 Professional Version. The Kolmogorov-Smirnov (K-S), Anderson-Darling (A-D) and Chi-square were used to test for the GoF of the fitted distributions and hence identify the most appropriate one. These authors found the GPD and GEV distributions as the best and second best fitted distributions for the delta, respectively.

The approach used by the researcher in this chapter draws some similarities from the work of Sukla et al. (2014). Similar to Sukla et al. (2014), the present study applies the K-S and A-D tests to test for GoF of the candidate distribu-

tions. Unlike Sukla et al. (2014), the present study does not apply the Chi-square GoF test since it does not show much sensitivity to either the centre or the tails of the fitted distribution. The approach in this chapter goes a step further to perform some simulation studies on the fitted distribution in order to investigate if the proposed distributions can mimic the characteristics of the observed data at the sites. The choice of the best fitting distribution using the ranks of the K-S and A-D is also made simpler to understand in this chapter than in Sukla et al. (2014).

Some other recent studies on distribution fitting in FFA were, as discussed in the previous chapter, performed for Dhaka city by Alam and Khan (2014) and for Tan River in Gujarat by Gohil and Chowdhary (2013). Among other recent studies discussed in Chapter 2, Blain and Meschiatti (2014) fitted flood frequency distributions at Campinas weather station in Sao Paulo and Neykov et al. (2014) used hybrid distribution to analyse daily precipitation. Mehrannia and Pokgohar (2014) recently reviewed the performance, ability and statistical tools of the EasyFit software. The package is used both for analysis and simulation of data. The researcher uses the EasyFit software for analysis and simulation in this chapter, as well as R programming language.

In this chapter ten candidate flood frequency distributions are fitted and compared for their GoF to the at-site data series. Estimation of high quantiles is then performed for the expected return periods of extreme floods at Chokwe, Combomune and Sicacate hydrometric stations in the LLRB of Mozambique. The distribution that is currently used for statistical FFA in the LLRB of Mozambique is the Gumbel distribution (UNDP, 2011; Lucio, 2007).

The rest of the chapter is arranged such that Section 3.2 discusses the research methodology, Section 3.3 presents the results and discussion of the chapter,

while Section 3.4 gives the concluding remarks of the chapter. Section 3.5 gives a summary of the chapter and finally Appendix 3.1 which presents the diagnostic plots of the chapter is given at the end of the chapter.

3.2 Research methodology

The data used in this chapter is hydrometric data series obtained from the Mozambique National Directorate of Water (DNA). Annual maximum daily flood height data series (in metres) recorded for the Limpopo River at Chokwe (1951-2010), Combomune (1966-2010) and Sicacate (1952-2010) were used in this chapter (see Fig. 1.4, 1.5 & 1.6, in Chapter 1). The data in its raw form composed of daily flood heights (or water levels) recorded at least once a day. Sequential steps were taken to select the highest (or peak) flood height in each hydrological year.

The block maxima approach in which years are taken as independent and identically distributed blocks was used in this chapter. Ten candidate hydrological frequency distributions commonly used in FFA were fitted to the data and checked for their GoF to the data. The ten candidate distributions fitted were the GEV, two-parameter Weibull (Weibull 2P), three-parameter Weibull (Weibull 3P), Gumbel, generalised gamma (GG), two-parameter gamma (Ga2), Three-parameter gamma (Ga3), two-parameter lognormal (LN2), three-parameter lognormal (LN3) and log-Pearson Type 3 (LP3). These frequency distributions are well-known in literature (Blain and Meschiatti, 2014; Sukla et al., 2014; Singo et al., 2012; Khodabin and Ahmadabadi, 2010; Reiss and Thomas, 2007; Beirlant et al., 2004; Coles, 2001; Vogel et al., 1993a,b). The distribution functions for the proposed candidate models are presented in the subsequent parts of this section.

3.2.1 Block maxima probability framework

The two fundamental approaches of EVT are block maxima and POT (Ferreira and de Haan, 2015). Block maxima consists of selecting the highest flood height in each hydrological year (called block), whereas the POT approach consists of observations that exceed a certain high threshold (Ferreira and de Haan, 2015, 2014; Reiss and Thomas, 2007; Beirlant et al., 2004; Coles, 2001). In this chapter the block maxima approach is used. The probability framework of block maxima is presented in Chapter 2, Subsection 2.7.1.

The following subsection presents the extreme value distributions and some of their families commonly applied in FFA and other natural hazards.

3.2.2 Flood frequency distributions

A number of flood frequency distributions commonly used in the estimation of flood events (e.g. flood heights, storms, high precipitation) and their corresponding frequency of occurrence were discussed in Chapter 2. The following subsections present the candidate distributions used in this chapter to analyse the flood height data series for comparative purposes.

Generalised extreme value distribution model

The GEV distribution is a flexible three-parameter distribution model that combines the Gumbel, Fréchet, and Weibull extreme value family of distributions (Beirlant et al., 2004; Coles, 2001). The GEV cumulative distribution function (CDF) is as in (2.6) given by

$$G(\mu,\sigma,\xi;x) \;=\; \begin{cases} \exp\left(-\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right), & \text{for } 1+\xi\left(\frac{x-\mu}{\sigma}\right)>0,\ \xi\neq 0, \\ \exp\left(-\exp(-\frac{x-\mu}{\sigma})\right), & x\in\mathbb{R},\ \ \xi=0, \end{cases} \tag{3.1}$$

where μ, σ and ξ are respectively location, scale and shape parameters estimated by the method of L-moments.

The estimates of the extreme quantiles are obtained from the quantile function, X_p, given by

$$X_p = G^{-1}(1-p) \quad = \quad \begin{cases} \mu + \frac{\sigma}{\xi}\left[(-\ln(1-p))^{-\xi} - 1\right], & \xi \neq 0, \\ \mu - \sigma \ln(-\ln(1-p)), & \xi = 0, \end{cases} \tag{3.2}$$

where p is the exceedance probability, which means the chance or likelihood that a given flood height is going to be equaled or exceeded at least once in a given return period.

Weibull distribution model

The Weibull distribution, also known as extreme value type III (EV3), has its two-parameter CDF given by

$$F(x) = 1 - \exp\left(-\left(\frac{x}{\beta}\right)^{\alpha}\right), \tag{3.3}$$

and the three-parameter Weibull CDF is given by

$$F(x) = 1 - \exp\left(-\left(\frac{x-\gamma}{\beta}\right)^{\alpha}\right), \tag{3.4}$$

where $\gamma,\ \beta$ and α are the continuous location, scale and shape parameters, respectively ($\gamma \equiv 0$ yields the two-parameter Weibull distribution). The Weibull distribution is also very common in hydrological studies, but it is more common in reliability studies (Reiss and Thomas, 2007).

The parameters of the Weibull 2P distribution were estimated by the least squares method, and those of the Weibull 3P distribution were estimated by

the MLE method.

Gumbel distribution model

The Gumbel distribution is in two forms: Gumbel Max (Maximum extreme value) and Gumbel Min (Minimum extreme value) used to model right-skewed and left-skewed data, respectively. The Gumbel distribution (also referred to as EV1) is one of the most commonly used distributions in annual flood flows (Alam and Khan, 2014; Singo et al., 2012; Abdul-Karim and Chowdhury, 1995). To emphasise its popularity in FFA, Gumbel (1958) commented that the rivers knew the theory and that only engineers needed to be convinced of the validity of the analysis. The CDF for the Gumbel Max is given by

$$F(x) = \exp\left(-\exp\left(-\frac{x-\mu}{\sigma}\right)\right), \tag{3.5}$$

and the CDF for the Gumbel Min is given by

$$F(x) = 1 - \exp\left(-\exp\left(-\frac{x-\mu}{\sigma}\right)\right), \tag{3.6}$$

where μ is the continuous location parameter and $\sigma(>0)$ is the continuous scale parameter. The parameters of both the Gumbel Max and Gumbel Min were estimated by the MOM.

Gamma distribution model

The two-parameter gamma CDF is given by

$$F(x) = \frac{\Gamma_{(x/\beta)}(\alpha)}{\Gamma(\alpha)}, \tag{3.7}$$

and the three-parameter Gamma CDF is given by

$$F(x) = \frac{\Gamma_{(x-\gamma)/\beta}(\alpha)}{\Gamma(\alpha)}, \tag{3.8}$$

where γ is a continuous location parameter ($\gamma \equiv 0$ yields the two-parameter gamma distribution), $\beta(> 0)$ is a continuous scale parameter, $\alpha(> 0)$ is a continuous shape parameter, Γ is the gamma function, and Γ_z is the incomplete gamma function where $z = (x/\beta)$ in (3.7) and $z = (x - \gamma)/\beta$ in (3.8).

The parameters of the Ga2 distribution were estimated by the MOM, while the Ga3 distribution parameters were estimated by the MLE method. The gamma distribution is among the commonly used distributions in FFA (Khodabin and Ahmadabadi, 2010) and it is in the class of distributions in the Gumbel domain (Beirlant et al., 2004).

Generalised gamma distribution model

The three-parameter generalised gamma (GG) has its CDF given by

$$F(x) = \frac{\Gamma_{(x/\beta)^k}(\alpha)}{\Gamma(\alpha)}, \tag{3.9}$$

where the superscript $k(> 0)$, is another shape parameter and all the other parameters are as defined for the gamma distribution.

It is also important to note that the Ga2 distribution is a subfamily of the GG distribution model (Khodabin and Ahmadabadi, 2010). The parameters of the GG distribution were estimated by the MLE method.

Log-Pearson type 3 distribution model

The LP3 distribution is popular in modeling flood flows (Sukla et al., 2014; Singo et al., 2012; Abdul-Karim and Chowdhury, 1995). The CDF of the LP3 is given by

$$F(x) = \frac{\Gamma_{(\ln(x)-\gamma/\beta)}(\alpha)}{\Gamma(\alpha)}, \tag{3.10}$$

where γ is a continuous location parameter, $\beta(\neq 0)$ is the continuous scale parameter and $\alpha(> 0)$ is the continuous shape parameter. These parameters were estimated by the MOM.

Log-normal distribution model

The log-normal distribution is among the commonly used distributions in extreme flood frequency analysis (Sukla et al., 2014; Singo et al., 2012; Abdul-Karim and Chowdhury, 1995) and the distribution is also in the Gumbel domain (Beirlant et al., 2004). The CDF for LN2 is given by

$$F(x) = \Phi\left(\frac{\ln x - \mu}{\sigma}\right), \tag{3.11}$$

and the CDF for LN3 is

$$F(x) = \Phi\left(\frac{\ln(x-\gamma) - \mu}{\sigma}\right), \tag{3.12}$$

where Φ is a Laplace integral, γ is a continuous location parameter ($\gamma \equiv 0$ yields the two parameter lognormal), $\sigma(> 0)$ is a scale parameter and μ is the continuous shape parameter. The parameters of both LN2 and LN3 were estimated by the MLE method.

3.2.3 Parameter estimating methods

The parameter estimation methods used in this study are the L-moments, method of moments (MOM), least squares estimators (LSE) methods and maximum likelihood estimators (MLE) methods. The L-moments, MLE and PWM were discussed in detail in Chapter 2. Literature is abundant on the extensive usage of L-moments mainly attributed to its little bias when compared to other methods (Smithers, 2012). The MLE is also a very common parameter estimation method in EVT mainly attributed to its flexibility and consistency (Dombry, 2015). The PWM has also recently been discussed in block maxima approach (Ferreira and de Haan, 2015). The MOM and LSE methods are also common methods in statistics although not as frequently used as the MLE, L-moments and PWM in EVT.

L-moments and PWM methods

The PWM estimators for the GEV in block maxima approach are given in this thesis in Chapter 2 pages 69 and 70. According to Izinyon and Ehiorobo (2015) the L-moments are a linear combination of PWM estimators, and both PWM and L-moments are used in summarising theoretical probability distributions and observed samples. They are commonly used for parameter estimation, interval estimation and hypothesis testing and these methods have found useful application in FFA (Izinyon and Ehiorobo, 2015; Shabri and Ariff, 2009; Vogel et al., 1993b).

The PWM for the k^{th} order of a distribution function is given by (Sukla et al., 2014; Shabri and Ariff, 2009, with references therein):

$$\beta_k = \int_0^1 x(F)F(x)^k dF, \quad k = 0, 1, 2, ... \tag{3.13}$$

where $x(F)$ is the quantile function, that was previously denoted as X_p in

(3.5), and it corresponds to the inverse of the CDF, $F(x)$, and $F(x)$ is the non-exceedance probability.

The first four PWM unbiased estimators, β_j, for any distribution are computed as follows (Sukla et al., 2014; Shabri and Ariff, 2009; Hosking and Wallis, 1997):

$$\begin{aligned}
\beta_0 &= \frac{1}{n}\sum_{i=1}^{n} X_{(i)}, \\
\beta_1 &= \sum_{i=1}^{n-1}\left[\frac{(n-i)}{n\,(n-1)}\right] X_{(i)}, \\
\beta_2 &= \sum_{i=1}^{n-2}\left[\frac{(n-i)\,(n-i-1)}{n\,(n-1)\,(n-2)}\right] X_{(i)}, \\
\beta_3 &= \sum_{i=1}^{n-3}\left[\frac{(n-i)\,(n-i-1)\,(n-i-2)}{n\,(n-1)\,(n-2)\,(n-3)}\right] X_{(i)},
\end{aligned} \tag{3.14}$$

where $X_{(i)}$ is the ranked annual maximum flood height series in which $X_{(1)}$ is the largest flood height and $X_{(n)}$ is the smallest.

The L-moments, λ_k, of a random sample X is defined in terms of the PWM (Sukla et al., 2014; Shabri and Ariff, 2009; Hosking and Wallis, 1997). The general form of the L-moments is given by

$$\lambda_{k+1} = (-1)^{k-r}\sum_{r=0}^{k} P^*_{k,r}\beta_r, \tag{3.15}$$

where the coefficients $P^*_{k,r}$ are defined as

$$P^*_{k,r} = (-1)^{k-r}\binom{k}{r}\binom{k+r}{r} = \frac{(-1)^{k-r}\,(k+r)!}{(r!)^2\,(k-r)!}.$$

The first four L-moments unbiased estimators, $\lambda_4 = (\lambda_1, ..., \lambda_4)$, are therefore

given by

$$\begin{aligned}
\lambda_1 = \ell_1 &= \beta_0, \\
\lambda_2 = \ell_2 &= 2\beta_1 - \beta_0, \\
\lambda_3 = \ell_3 &= 2\beta_2 - 6\beta_1 + \beta_0, \\
\lambda_4 = \ell_4 &= 20\beta_3 - 30\beta_2 + 12\beta_1 - \beta_0.
\end{aligned} \tag{3.16}$$

The L-moments ratios which are used for expressing the parameter estimates are computed as follows:

$$\begin{aligned}
\tau_2 &= \frac{\lambda_2}{\lambda_1}, \\
\tau_3 &= \frac{\lambda_3}{\lambda_2}, \\
\tau_4 &= \frac{\lambda_4}{\lambda_2},
\end{aligned} \tag{3.17}$$

where $\tau_2,\ \tau_3$ and τ_4 are the L-coefficient of variation (L-CV), the L-skewness and L-kurtosis, respectively (Sukla et al., 2014; Shabri and Ariff, 2009; Hosking and Wallis, 1997). The parameter estimation is done by equating the sample L-moments to distribution L-moments (Sukla et al., 2014). The distribution L-moments are clearly given in Hosking and Wallis (1997). For example, the parameter estimates for the GEV distribution when using L-moments are

$$\begin{aligned}
\sigma &= \frac{\ell_2 \xi}{\Gamma(1+\xi)\,\Gamma(1-2^{-\xi})} \\
\mu &= \ell_1 + \frac{\sigma\left(\Gamma(1+\xi) - 1\right)}{\xi}, \\
\xi &= 7.8590C + 2.9554C^2, \\
C &= \frac{2}{3+\tau_3} - \frac{\ln 2}{\ln 3}.
\end{aligned} \tag{3.18}$$

Maximum likelihood estimators method

The MLEs method has already been extensively discussed in the previous chapter for the block maxima approach. Consider the distribution of annual maxima flood heights, $M_{k,m}$, for large k (the number of blocks or years in this case), which is approximately a GEV distribution with parameters (a_m, b_m, ξ) (Dombry, 2015). The general form of the log-likelihood for large k or a large sample of size $n = m \times k$ (i.e. $M_{1,m}, ..., M_{k,m}$) is

$$L_k(\mu, \sigma, \xi) = \frac{1}{k} \sum_{i=1}^{k} \ell_{(\mu,\sigma,\xi)}(M_{k,m}). \tag{3.19}$$

According to Dombry (2015) L_k does not have a global maximum and therefore $(\hat{\mu}_k, \hat{\sigma}_k, \hat{\xi}_k)$ is an MLE if L_k has a local maximum at $(\hat{\mu}_k, \hat{\sigma}_k, \hat{\xi}_k)$. The MLEs are obtained from the solution of the likelihood equations

$$\nabla L_k = 0 \quad \text{with } \nabla L_k = \left(\frac{\partial L_k}{\partial \mu}, \frac{\partial L_k}{\partial \sigma}, \frac{\partial L_n}{\partial \xi} \right). \tag{3.20}$$

The MLE is any solution of (3.20) with a negative definite Hessian matrix (Dombry, 2015). The MLE method is not only limited to annual maxima, but it can also be applied to other types of data such as threshold excesses (by replacing the block maxima with ordered exceedances) and order statistics.

For example, consider $\boldsymbol{X}_i = \{X_1, X_2, ..., X_n\}$ to be a random vector of observations whose joint density function is $f(x_1, x_2, ..., x_n|\theta)$. Let the likelihood of θ be defined as

$$l(\theta) = f(x_1, x_2, ..., x_n|\theta).$$

In general, it is often easier and preferred to maximise the natural log of the likelihood function. If the $X_i's$ are iid then the log-likelihood takes the general

form

$$l(\theta) = \log f(x_i|\theta) = \sum_{i=1}^{n} f(x_i|\theta). \tag{3.21}$$

Then the maximum likelihood estimate, $\hat{\theta}$, of the parameter θ is the value of θ that maximises the log-likelihood function.

The requirement when using the MLE approach in block maxima is that the block size needs to be controlled, that is, to let the block size (or length of block) depend on the sample size in order to meet the asymptotic requirements (Dombry, 2015; Ferreira and de Haan, 2015). This requirement, however, does not present problems in the analysis of flood heights since it is natural to block by years.

Method of moments

The MOM is one of the oldest methods of finding, θ, point estimators. It is a simple procedure for finding an estimator for population parameters. Consider $\mu_k^{'} = E\left[X^k\right]$ to be the k^{th} moment about the population of a random variable, X, if it exists. Let $m_k^{'} = \frac{1}{n}\sum_{i=1}^{n} X_i^k$ be the corresponding k^{th} moment from a selected sample. Then by the MOM, $m_k^{'}$ is the estimator of $\mu_k^{'}$.

The MOM is based on equating sample moments to corresponding population moments using the following procedure. Let there be ℓ parameters, $\theta = (\theta_1, ..., \theta_\ell)$, then:

1. find the ℓ population moments, $\mu_k^{'}, \ \forall\ k = 1, ..., \ell$, where $\mu_k^{'}$ contains one or more parameters $\theta_1, ..., \theta_\ell$;
2. obtain the corresponding ℓ sample moments, $m_k^{'}, \ \forall\ k = 1, ..., \ell$;
3. form the system of equations $\mu_k^{'} = m_k^{'}, \ \forall\ k = 1, ..., \ell$, and solve for the parameter θ to obtain the moment estimator, $\hat{\theta}$.

For example, the first population moment is $\mu_1^{'} = E(X)$ and the first moment of a sample is $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. Therefore moment estimator of μ_1 is $\bar{X}$.

Least squares methods

The least squares method is based on minimising the square of the errors or discrepancies between the observed and the expected values. Consider Y_i to be a random variable with finite mean μ_{y_i} and standard deviation σ_y, where the form of $f_{Y_i}(y:\theta)$, may not be known. Let n observations be a sample from this population, then the least squares estimator(s) of θ are those values of θ, θ_{ls}, that minimise

$$S^2 = \sum_{i=1}^{n} \left[(y_i - E\left[y_i : \boldsymbol{x}_i, \theta\right])^2\right], \tag{3.22}$$

where $\boldsymbol{x}_i$ is the vector of observed dependent variables. In order to find the least squares estimate of θ, it requires that the form of $E[y_i : \boldsymbol{x}_i, \theta]$ is specified, however, in least squares estimation it is not a requirement to specify the form of $f_{Y_i}(y_i, \boldsymbol{x}_i, \theta)$, as is a typical requirement in MLE method, which automatically implies $E[y_i : \boldsymbol{x}_i, \theta]$. In order to obtain the estimators, the partial derivatives of S^2 are with respect to the variable of interest and the resulting expression is equated to zero, then solve for the variable of interest.

3.2.4 Goodness-of-fit tests

Two goodness-of-fit (GoF) tests K-S and A-D were tested at 5% level of significance. Let $x_1, x_2, ..., x_n$ be a sample of n annual maximum flood heights observed and suppose the CDF of the random variable X is $F(.)$, then the K-S and A-D tests are as presented in the next two subsections.

Anderson-Darling test

The A-D test is based on a comparison between the fit of an observed CDF to a theoretical (expected) CDF (Sukla et al., 2014). The test statistic, denoted, A^2, is defined by

$$A^2 = -n - \frac{1}{n}\sum_{i=1}^{n}(2i-1)\left[\ln F(x) + \ln\left(1 - F_n(x)\right)\right]. \tag{3.23}$$

where $F_n(x)$ is the empirical CDF and $F(x)$ the theoretical CDF and the data x are ordered.

The hypothesis for GoF under the A-D test procedure is
H_0: The data follow a specified distribution, versus, H_1: The data do not follow the specified distribution.

The mathematical formulation of the hypothesis for GoF under the A-D test procedure is

$$H_0 :\ F(x) = F_0(x;\boldsymbol{\theta})\ \ vs.\ \ H_1 :\ F(x) \neq F_0(x;\boldsymbol{\theta}).$$

where F_0 is a specified (hypothesised) distribution such as GEV and $\boldsymbol{\theta}$ is a vector of unknown parameters.

The H_0 for this A-D test is rejected at 5% level of significance if $F(x)$ is very different from the hypothesised distribution $F_0(x;\boldsymbol{\theta})$, i.e. if A^2 calculated is greater than the tabulated or critical value (Sukla et al., 2014).

The A-D test is more sensitive to the tails of the distribution than the K-S test. In other words, more weight is given to the tail of the distribution in A-D test than in K-S test (Sukla et al., 2014). This implies that when comparing

candidate distributions for their goodness-of-fit of an empirical extreme value distribution, the A-D test is more appropriate and thus more weight in tie-breaking must be given to the A-D test. This approach is employed in this thesis.

Kolmogorov-Smirnov test

The K-S test is based on a comparison between the largest vertical distance, D_{max} between the empirical (observed) CDF, $F_n(x)$, and the theoretical CDF, $F(x)$. The test statistic of the K-S test is given by

$$D_{max} = \mathbf{Max}_x \left| F_n(x) - F(x) \right|. \tag{3.24}$$

The hypothesis for the GoF under the K-S test procedure is

$$H_0 : \ F(x) = F_0(x; \boldsymbol{\theta}) \ \ vs. \ \ H_1 : \ F(x) \neq F_0(x; \boldsymbol{\theta}).$$

where F_0 is a specified distribution such as GEV and $\boldsymbol{\theta}$ is a vector of unknown parameters.

The H_0 for this test is rejected at 5% level of significance if D_{max} calculated is greater than the tabulated value $D_{0.05} = 1.36/\sqrt{n}$ (Sukla et al., 2014). The K-S test is more sensitive to the centre of the distribution than the A-D test.

Diagnostic plots

The diagnostic plots consists of probability density function (PDF) plot, probability-probability (P-P) plots and quantile-quantile (Q-Q) plots, among others (Berning, 2010). The Q-Q plots are extensively discussed in Berning (2010). A Q-Q plot is a graphical tool that is used to assess the GoF, that is, whether the underlying distribution F is the appropriate distribution for random variable

X, whose observations are $x_1, x_2, ..., x_n$. Berning (2010) defined a Q-Q plot as follows:

Definition 3.1. : *Q-Q plot (Berning, 2010, p.8).*
A Q-Q plot is a graphical tool used to visually assess the goodness-of-fit of a distribution. A plot of $(Q(i/(n+1)); x_{i,n}))\,, \quad i = 1, 2, ..., n$*, should be approximately linear if* $x_1, x_2, ..., x_n$ *are from a distribution with quantile function* Q*.*

One main advantage of a Q-Q plot is that the location and scale parameters of the model being fitted do not need to be known in advance. In a normal Q-Q plot the location, μ, and scale, σ, parameters of the distribution can be estimated from the normal Q-Q plot. This leads to the definition of a normal Q-Q plot.

Definition 3.2. : *Normal Q-Q plot (Berning, 2010, p.8).*
Consider observations $x_1, x_2, ..., x_n$*. Then a plot of* $(\Phi^{-1}(i/(n+1)); x_{i,n}))\,, \quad i = 1, 2, ..., n$ *is a normal Q-Q plot of the data, where* $\Phi(.)$ *is the standard normal distribution function. If the observations* $x_1, x_2, ..., x_n$ *are from a normal distribution with mean,* μ*, and standard deviation,* σ*, then the slope should be approximately linear with gradient* σ *and intercept* μ*.*

Similar to the Q-Q plot, the definition of P-P plot follows.

Definition 3.3. : *P-P plot.*
A P-P plot is defined as a graphical tool used to visually assess the GoF of a specific distribution. It is based on a plot of the empirical CDF values against the theoretical CDF values. Like the Q-Q plot, the P-P plot is used to assess how well a specified distribution fits the observed data. The P-P plot should be approximately linear if the specified theoretical distribution is the appropriate model for the random variable $X = \{x_1, x_2, ..., x_n\}$*.*

The only difference between a Q-Q plot and a P-P plot is that in the Q-Q plot the observed values (quantiles) are used, whereas in the P-P plot cumulative

probabilities are used in plotting the graph.

A PDF can be defined for both discrete and continuous random variables. However, in this thesis the focus is on the continuous random variables and hence the PDF for a discrete random variable will not be defined. The definition for the PDF of a continuous random variable follows.

Definition 3.4. : *PDF plot.*
Consider a continuous random variable X*. Then the PDF of* X *is a function* $f(x)$ *such that for any two data points (numbers)* a *and* b *(*$a \leq b$*),*

$$P(a \leq x \leq b) = \int_a^b f(x)dx. \tag{3.25}$$

That is, the probability that X *takes on a value in the interval* $[a, b]$ *is the area above this interval and under the graph of a density function. The name usually given to the graph of* $f(x)$ *is the density curve.*

A PDF is usually plotted by superimposing it on a histogram, and it gives a visual display of the symmetry (or asymmetry) and kurtosis of the empirical distribution.

3.3 Results and discussion

This section presents results for the study in this chapter. The section also presents a brief discussion of the results for this chapter.

3.3.1 Characteristics of annual maximum flood heights at the three study sites

A summary of the descriptive statistics of the observed data at the three study sites Chokwe, Combomune and Sicacate are presented in Table 3.1. The re-

sults in Table 3.1 show that the distribution of the data is positively skewed for both Chokwe and Combomune, while for Sicacate the distribution of the data is negatively skewed. These findings are consistent with the results of the measures of central tendency based on the means and medians of these three distributions, that is, for both Chokwe and Combomune the mean of annual maximum flood heights is greater than the median confirming right-skewness for the distributions of these two sites (see Table 3.1). On the other hand, the mean of annual flood heights for Sicacate is less than the median which is a confirmation of the left-skewness of the distribution at the site and this is again consistent with the results of skewness (Table 3.1). The 13 m flood height that occurred at both Chokwe and Sicacate in February-March 2000 was very extreme in comparison to the rest of the data recorded at the two sites in Figures 1.4 & 1.6 in Chapter 1. However, the 95^{th} percentile of the annual maximum flood height data at Chokwe (8.11 m) and Sicacate (11.05 m), as well as Combomune (9.81 m) indicates that the 13 m flood height of the year 2000 was far more severe at Chokwe than all the other sites, since the annual maximum flood heights at Chokwe are generally low.

In Table 3.1 relative variability as measured by the coefficient of variation (CV) is lowest at Combomune compared to the other two sites. On the other hand, the highest relative variability in annual maximum flood heights is experienced at Sicacate, followed closely by Chokwe. Absolute variability as measured by interquartile range (IQR), range and standard deviation is also lowest at Combomune, followed by Chokwe and then Sicacate (Table 3.1). In general, the annual maximum flood heights are more consistent at Combomune than at Chokwe and Sicacate, where high variability in annual maximum flood heights is experienced.

Table 3.1: Descriptive statistics of the characteristics of annual maxima flood heights at the three sites

Statistic	**Chokwe**	**Combomune**	**Sicacate**
Sample size	60	45	59
Min	0.85	2.49	0
Max	13.00	10.97	12.96
Range	12.15	8.48	12.96
Mean	5.14	5.95	6.99
Median	4.94	5.68	7.39
Variance	4.36	3.59	9.93
Std. dev.	2.09	1.89	3.15
CV	0.41	0.32	0.45
Std. Error	0.27	0.28	0.41
Skewness	0.85	0.53	-0.44
Excess kurtosis	2.00	0.11	-0.58
Lower quartile (Q1)	3.63	4.60	4.42
Upper quartile (Q3)	6.60	7.39	9.77
95^{th} percentile	8.11	9.81	11.05
IQR	2.97	2.79	5.35

3.3.2 Parameter estimation

Table 3.2 presents results for the estimates of the parameters for all the three sites. Notably a number of three-parameter distributions did not fit the data at Sicacate due to the continuous location parameter that was identically equal to zero. This failure of fit could be attributed to the existence of zeroes in the annual maximum flood height data at the site which occurred during the years of drought. Singo et al. (2012) stressed that a distribution with a larger number of flexible parameters would be able to model the data more accurately than a distribution with lesser number of flexible parameters. For instance, GEV and GG would likely model the data better at Sicacate than LN2 (Table 3.2). At Chokwe Ga3, GEV and LN3 had large values for their flexible parameters (Table 3.2).

Table 3.2: Summary of the parameter estimates for the candidate distributions at the three study sites

Distribution	Chokwe	Combomune	Sicacate
GEV	$\xi = -0.115$	$\xi = -0.104$	$\xi = -0.464$
	$\sigma = 1.829$	$\sigma = 1.686$	$\sigma = 3.370$
	$\mu = 4.283$	$\mu = 5.131$	$\mu = 6.197$
GG	$\kappa = 0.992$	$\kappa = 1.003$	$\kappa = 8.895$
	$\alpha = 6.017$	$\alpha = 9.919$	$\alpha = 0.224$
	$\beta = 0.843$	$\beta = 0.603$	$\beta = 11.272$
Weibull 2P	$\alpha = 2.714$	$\alpha = 3.646$	$\alpha = 2.917$
	$\beta = 5.684$	$\beta = 6.464$	$\beta = 8.177$
Weibull 3P	$\alpha = 2.372$	$\alpha = 2.177$	
	$\beta = 5.278$	$\beta = 4.378$	
	$\gamma = 0.464$	$\gamma = 2.065$	$\gamma \equiv 0$
Gumbel	$\sigma = 1.624$	$\sigma = 1.477$	$\sigma = 2.405$
	$\mu = 4.212$	$\mu = 5.093$	$\mu = 8.427$
Ga2	$\alpha = 6.110$	$\alpha = 9.852$	$\alpha = 5.469$
	$\beta = 0.843$	$\beta = 0.603$	$\beta = 1.330$
Ga3	$\alpha = 10.882$	$\alpha = 9.352$	
	$\beta = 0.621$	$\beta = 0.617$	
	$\gamma = -1.607$	$\gamma = 0.173$	$\gamma \equiv 0$
LN2	$\sigma = 0.442$	$\sigma = 0.325$	$\sigma = 0.464$
	$\mu = 1.552$	$\mu = 1.732$	$\mu = 1.890$
LN3	$\sigma = 0.216$	$\sigma = 0.215$	
	$\mu = 2.217$	$\mu = 2.139$	
	$\gamma - 4.251$	$\gamma = -2.742$	$\gamma \equiv 0$
LP3	$\alpha = 4.475$	$\alpha = 39.191$	
	$\beta = -0.211$	$\beta = -0.053$	
	$\gamma = 2.495$	$\gamma = 3.789$	$\gamma \equiv 0$

3.3.3 The simulation procedure

A simulation procedure was performed using the estimates of the parameters provided in Table 3.2. This was done with the aid of a statistical software specifically designed for distribution fitting called EasyFit Professional, version 5.6, that was recently released in early 2015. The simulation could also be done using R software, but for the purpose of this chapter EasyFit was found to be more appropriate (Mehrannia and Pokgohar, 2014; Southworth and Hef-

fernan, 2013; Ribatet, 2006).

The simulation procedure proceeds as follows. The first stage involved the estimation of the parameters of the candidate distributions. These parameters were then used in a routine procedure to simulate 30 samples. Each of the simulated samples was compared with the empirical sample through a correlation analysis procedure. The empirical sample would represent the control sample as the 31^{st} sample. The tails of the simulated samples were also checked for their closeness in mimicking the characteristics of the empirical sample distribution. More weight was given to a candidate distribution that produced samples which were able to mimic, in particular, the tails of the empirical sample distribution. The mean correlation coefficient of the 30 samples, relative to the empirical sample, was calculated and recorded (see Tables 3.3, 3.4 & 3.5).

It was found that although most of the mean correlation coefficients were significantly high (> 0.70), quite a large number of distributions were not able to produce samples that mimic the tails of the empirical sample distribution. This was particularly true in the case of distributions with lesser number of flexible parameters, that is, in most distributions with two or less parameters (Singo et al., 2012). This implies that the large values of the mean correlation coefficients for samples produced by such distributions were entirely based on closeness of the simulated sample data to the empirical sample data in the centre of the distribution. Distributions such as the Gumbel distribution at some sites, two-parameter Weibull and two-parameter log-normal were notably among such distributions that were able to model the central part of the empirical sample distribution but failed to mimic its extreme tails. It must be emphasised that in extreme value analysis, the primary aim is in modelling the behaviour of the extreme tails, and not necessarily the centre as was the case with some candidate distributions. The three parameter distributions such as

the GEV, three-parameter log-normal, log-Pearson and gamma prevailed better in these situations, that is, mimicking the characteristics of the extreme tails of the empirical distribution.

3.3.4 Assessing the goodness-of-fit (GoF) of the candidate distributions for Chokwe

A summary of the GoF analysis results for the A-D and K-S tests is presented in Table 3.3 for Chokwe. The results of the simulation procedure for Chokwe are also presented in Table 3.3 in the form of mean correlation coefficients. Regarding the GoF based on A-D and K-S tests, a rank was assigned to each distribution based on the position of the candidate distribution in the initial rank of the population of 61 distributions automatically fitted to the data using the EasyFit statistical software (Mehrannia and Pokgohar, 2014). In EasyFit statistical software the fitted distributions are ranked based on the significance of their p-values or test statistics (i.e. the distribution with the most significant (smallest) p-value is ranked first). Since ten candidate distributions were considered in the study for this chapter, the ten distributions were re-ranked based on their initial positions in the population rank of 61 distributions. The new assigned ranks ranged from 1 to 10 for both A-D and K-S tests. A rank of 1 means the best fitting distribution, while a rank of 10 means a poor fit to the data. In other words, the lower the rank, the better the fit. The A-D and K-S ranks for a particular distribution were summed together to form the total rank of the distribution. The total ranks were then used to categorise a candidate distribution as best model, 2^{nd} best, etc. as shown in Table 3.3. These categories were assigned by taking into consideration the performance of the distribution in the simulation procedure and its rank using the A-D test. In general, a candidate distribution with the least total rank would fall in the category of the best model (Table 3.3). In the event of ties, a tie-breaker would be a smaller value of the rank in the A-D test. That is, more weight was given

to the A-D test since it is more sensitive to the tails of the distribution than the K-S distribution, which in turn is more sensitive to the centre of the distribution.

Table 3.3: Ranking the candidate distributions at Chokwe and assessing the quality of GoF through a procedure of 30 simulated samples

Distribution	**K-S Test**	**A-D Test**	**Rank Total**	**Correlation Coef.**	**Remarks**
GEV	4	3	7	0.973	3^{rd} best model
GG	7	5	12	0.971	good model
Weibull 2P	3	7	10	0.968	good model
Weibull 3P	6	6	12	0.971	good model
Gumbel	8	8	16	0.969	good model
Ga2	5	4	9	0.966	good model
Ga3	2	1	3	0.973	best model
LN2	10	9	19	0.965	good model
LN3	1	2	3	0.975	2^{nd} best model
LP3	9	10	19	0.967	good model

Key: Note that for both A-D and K-S tests, H_0 was not rejected at 5% significance level (even to as high as 20% significance level) for all the best three selected distributions.

Results in Table 3.3 showed that the three best fitting distributions for Chokwe were Ga3, LN3, and GEV, in their respective order. Hypothesis testing for all the three distributions also confirmed that the observed sample data comes from these selected distributions. The mean correlation coefficients (all three > 0.97) for the data randomly generated from the fitted distributions showed a very strong positive correlation between the observed data and the simulated data. This implied that the three best ranked distributions selected for Chokwe were able to reproduce data with similar features and characteristics as the observed data, based on the estimated parameters.

3.3.5 Assessing the goodness-of-fit (GoF) of the candidate distributions for Combomune

In Table 3.4 the results for the assessment of the GoF of the ten candidate distributions for Combomune are presented. The results in Table 3.4 showed that the GEV was the best ranked distribution followed by Ga2 and a tie between LP3 and and Ga3. Using the tie-breaker of the A-D test, this would mean that the LP3 would obtain a higher rank than the Ga3 and therefore a better ranked category. Thus the final best three distribution models at Combomune were the GEV, Ga2, and LN3, in their respective order.

Table 3.4: Ranking the candidate distributions at Combomune and assessing the quality of GoF through a procedure of 30 simulated samples

Distribution	K-S Test	A-D Test	Rank Total	Correlation Coef.	Remarks
GEV	1	1	2	0.986	best model
GG	4	4	8	0.982	good model
Weibull 2P	10	10	20	0.973	good model
Weibull 3P	8	7	15	0.981	good model
Gumbel	9	9	18	0.975	good model
Ga2	3	3	6	0.984	2^{nd} best model
Ga3	2	5	7	0.985	good model
LN2	7	8	15	0.978	good model
LN3	6	6	12	0.984	good model
LP3	5	2	7	0.986	3^{rd} best model

Key: Note that for both A-D and K-S tests, H_0 was not rejected at 5% significance level (even to as high as 20% significance level) for all the best three selected distributions.

These selected distributions also passed the hypothesis test. That is, hypothesis testing for all the three distributions confirmed that the observed sample data comes from these selected distributions. The mean correlation coefficients (all three > 0.98) between the observed data and the generated random samples showed that there was a very strong positive correlation between the data

generated from the fitted distributions and the observed data. This indicated that the three best ranked distributions at Combomune were able to mimic the features and characteristics of the observed data at the site.

3.3.6 Assessing the goodness-of-fit (GoF) of the candidate distributions for Sicacate

Table 3.5 presents results for the assessment of the GoF of the ten candidate distributions for Sicacate. The results in Table 3.5 showed that the GEV was the best ranked distribution followed by a tie between Gumbel Min and GG. Since the tie-breaker is the A-D test, this would mean that the Gumbel min would obtain a higher rank than the GG and therefore a better ranked category. Thus the final best three distribution models at Sicacate were the GEV, Gumbel Min, and GG, in their respective order. The mean correlation coefficients (all three $>$ 0.95) between the observed data and the generated random samples showed that there was a very strong positive correlation between the data generated from the fitted distributions and the observed data. This is an indication that the three best ranked distributions at Sicacate were able to mimic the pattern and characteristics of the observed data at the site.

It should be noted that the 3^{rd} distribution, GG, failed to pass the hypothesis testing for the A-D test, but passed all the hypothesis testing for the K-S test up to 20% significance level. In other words, using the K-S test, H_0 was not rejected that the observed sample data comes from the GG distribution, while for the A-D test H_0 was rejected (even at as far as 20% significance level) in favour of the alternative hypothesis H_1 that the observed sample data does not come from a GG distribution. Therefore the GG distribution was simply chosen to make up the number three based only on the K-S test. This was an exceptional case in the candidate distribution choice in this chapter since a better third distribution could not be identified for the site.

Table 3.5: Ranking the candidate distributions at Sicacate and assessing the quality of GoF through a procedure of 30 simulated samples

Distribution	K-S Test	A-D Test	Rank Total	Correlation Coef.	Remarks
GEV	1	1	2	0.983	best model
GG	2	3	5	0.988	3^{rd} best model
Weibull 2P	4	4	8	0.970	poor fit
Weibull 3P	-	-	-	-	unsuitable
Gumbel	3	2	5	0.957	2^{nd} best model
Ga2	5	5	10	0.936	poor fit
Ga3	-	-	-	-	unsuitable
LN2	6	6	12	0.883	poor fit
LN3	-	-	-	-	unsuitable
LP3	-	-	-	-	unsuitable

Key: Note that for both A-D and K-S tests, H_0 was not rejected at 5% significance level (even to as high as 20% significance level) for the best two distributions GEV and Gumbel Min. The GG distribution passed the K-S hypothesis testing just like the GEV and Gumbel Min, but failed the A-D test dismally, even at 20% significance level.

3.3.7 Diagnostic plots based on the three best distributions at each site

In this subsection the diagnostic plots for the three sites are presented. These diagnostic plots are presented in the form of Q-Q plots, P-P plots, probability difference plots and probability density functions (PDFs). All the figures (Figures 3.1-3.6) for the diagnostic plots are presented at the end of this chapter in Appendix 3.1.

Diagnostic plots of Chokwe selected distributions

The diagnostic plots for the best three flood frequency distributions selected for Chokwe are presented in the following figures. Figure 3.1 in Appendix 3.1 presents graphs for the PDF and probability difference of the models of the best three fitting distributions selected for Chokwe. Figure 3.2, Appendix 3.1,

presents graphs for the Q-Q plots and the P-P plots. The graph for the PDFs in Figure 3.1 indicated that all the candidate distributions tested at Chokwe were likely to best fit the data. This also agreed with the mean correlation coefficient presented in Table 3.3. The histogram (Figure 3.1, panel A) of the annual maximum daily flood height data recorded at Chokwe revealed a unimodal and positively skewed distribution with one clear outlier to the right.

The three best fitting distributions Ga3, LN3 and GEV presented in Figure 3.1, panel A, exhibited similar probability densities and were able to model the outlier class well which the other distributions were not able to do. These three distributions were almost indistinguishable in their ability to model the data at Chokwe (Figure 3.1, panel A). This was confirmed by the P-P and Q-Q plots (Figure 3.2) and probability difference plots (Figure 3.1, panel B) where the points were very close to the line of best fit in the P-P and Q-Q plots for all the three distributions. That is, both the plots indicated reasonable linearity which implied that the models selected are the appropriate models for the site. Singo et al. (2012) states that if the maximum absolute difference (MAD) is less than 0.05 (5%) the fit can be considered good, and the model becomes very good if the MAD is less than 1% limits. The probability difference plot in Figure 3.1 showed that the MAD was almost within 0.00 limits in the upper tail, 0.02 limits in the lower tail and within 0.055 limits near the high density areas of the distributions. This was an indication of a very good fit to both the tails and the centre of the distribution of the data at Chokwe by the Ga3, LN3 and GEV distributions.

Diagnostic plots of Combomune selected distributions

Figure 3.3 presents graphs for the PDF and probability difference of the best three distributions selected for the Combomune site. The P-P plots and the Q-Q plots of the three selected distributions are presented in Figure 3.5. The graphs

for the PDFs of the selected distributions and the histogram of the empirical sample data fit quite well and the three distributions are almost indistinguishable in terms of the ability to model the empirical sample data represented by the histogram (Figure 3.3, panel A).

The probability difference (Figure 3.3, panel B) revealed that the GEV and Ga3 MAD points were within 0.05 towards the centre of the distribution and within 0.01 near the tails. The LP3 also performed well in the tails with an MAD of within 0.01 limits, but near the centre the MAD was within 0.08 limits. In general, the probability difference plot has revealed a good fit to the Combomune data for all the three selected distributions.

The Q-Q and P-P plots in Figure 3.6 both indicated reasonable linearity which implied that the models selected are the appropriate models for the site. The points in both Q-Q and P-P plots are very close to the lines of best fit. In general, all the diagnostic plots discussed above confirmed that all the three selected distributions, GEV, Ga3 and LP3 are appropriate to model flood heights at the site.

Diagnostic plots of Sicacate selected distributions

Figure 3.3 presents graphs for the PDFs and the probability difference plots of the candidate distributions selected for Sicacate. The histogram of annual maximum daily flood height data in Figure 3.3 revealed a unimodal and negatively skewed distribution. The graph of the PDFs (Figure 3.5, panel A) showed that the best fitting distributions are GEV, Gumbel Min and GG whose probability densities exhibited shapes that were closer to the distribution of the histogram of the empirical sample compared to the shapes of the other distributions such as the two-parameter gamma, Weibull and log-normal distributions.

Generally all the three distributions had probability difference (Figure 3.5, panel B) MAD of almost 0.00 limits for the upper tail and within 0.05 limits for the lower tail, but within 0.11 limits near the centre of the data which indicated poor fit in the centre of the empirical distribution, but very good fit in the upper tails and a good fit in the lower tail. It can be seen from the graph (Figure 3.5, panel B) that the GEV MAD points are closer to the zero-line of best fit than the other two selected distributions. This is an indication of good fit for the GEV distribution.

Figure 3.6 presents the P-P and Q-Q plots for the best three distributions at Sicacate. Generally all the points were reasonably close to the line of best fit in the tails, with the points of the GEV closer to the line in most regions than those of the other two distributions. Towards the centre of the distributions the points showed a tendency to deviate from the lines of best fit for both P-P and Q-Q plots. Thus the three distributions showed reasonably good results in modelling the upper tail but the GEV slightly outperformed both Gumbel Min and GG in modelling the lower tail. In most situations, the GG distribution was exposed, to a large extent, for its lack of fit. The Gumbel distribution was also exposed, to a lesser extent, in some cases. The GEV distribution prevailed as the best distribution, among the selected three, to model the extreme annual maximum flood heights at Sicacate. The Gumbel distribution coincided with literature since it is being used in the lower Limpopo River basin and it was interesting to find out that it came out in second position among the appropriate distributions for the site.

3.3.8 Expected return periods and flood height quantile estimation at the three study sites

The at-site extreme annual maximum flood heights and their corresponding return periods are presented in this subsection. The results for Chokwe are

presented first, followed by Combomune and then lastly Sicacate. These results are presented in tabular form and extensively discussed for each hydrometric station. Figure 3.7, in Appendix 3.1 presents the CDF graphs for the three sites Chokwe (panel A), Combomune (panel B) and Sicacate (panel C). The CDF graphs provide the non-exceedance probabilities, $F(x) = 1 - p$, from which the exceedance probabilities, p, can be obtained and consequently the return period, $T = \frac{1}{p}$, of a particular flood height can be obtained. These return periods and their corresponding annual maximum flood heights are presented in tabular form for each site in the next parts of the section.

Chokwe quantile estimation based on the best three distributions selected

Table 3.6 presents at-site results of the expected return periods and their corresponding probable high quantiles of annual maximum daily flood heights for Chokwe hydrometric station. All the three distributions selected for Chokwe Ga3, LN3 and GEV provided good estimates for the Chokwe hydrometric station. The flood height estimates from the three best distributions selected for Chokwe were closer to each other for smaller values of the return periods but after the 100-year flood height some of the distributions such as LN3 showed a tendency to approach higher values at a fast rate. The estimated flood heights are useful in the engineering design of hydraulic structures and major constructions in the lower Limpopo River basin. These estimated annual maximum flood heights can also be important in economic and agricultural planning since the basin is well known for its agricultural and fishing activities and it hosts the Chokwe Irrigation Scheme which stands as the main irrigation scheme in the country. Besides the agricultural and fishing activities, the basin is also very prone to extreme natural hydrological disasters that alternate between extreme floods and severe droughts (Jackson, 2013b; WMO, 2012).

Table 3.6: Results of the expected return periods and probable high quantiles of annual maximum daily flood heights at Chokwe

Expected return period: **T(years)**	**Exceedance probability:** **p**	**Non-exceedance probability or CDF: F(x)**	**Ga3 distribution: probable flood height(m)**	**LN3 distribution: probable flood height(m)**	**GEV distribution: probable flood height(m)**
2	0.5	0.5	4.94	4.93	4.94
5	0.2	0.8	6.79	6.76	6.80
10	0.1	0.9	7.87	7.86	7.91
25	0.04	0.96	9.13	9.15	9.18
50	0.02	0.98	9.99	10.06	10.03
100	0.01	0.99	10.80	10.92	10.82
200	0.005	0.995	11.58	11.76	11.54
250	0.004	0.996	11.82	12.03	11.75
500*	0.002	0.998	12.56	12.84	12.40
1000	0.001	0.999	13.27	13.64	13.00

* The 13 m flood height is over the 500-year flood level at Chokwe.

The results presented in Table 3.6 for Chokwe revealed that the 100-year flood height for Chokwe is just below 11 m based on the flood height estimates from the three distributions selected for the site and the 13 m flood height which occurred during the February-March 2000 floods in Mozambique and the neighbouring countries in Southern Africa is above the 500-year flood height at the site. These findings are a justification of the disastrous nature of the 13 m flood height in the basin that resulted in the death of more than 700 people and caused economic damages estimated at US$500 million and drowned more than 20,000 herds of cattle (Jackson, 2013a,b).

Combomune quantile estimation based on the best three distributions selected

In Table 3.7 the at-site results of the expected return periods and their corresponding probable high quantiles of annual maximum daily flood heights for

Combomune hydrometric station are presented.

Table 3.7: Results of the expected return periods and probable high quantiles of annual maximum daily flood heights at Combomune

Expected return period: **T(years)**	**Exceedance probability:** **p**	**Non-exceedance probability or CDF: F(x)**	**GEV distribution: probable flood height(m)**	**Ga2* (Ga3) distribution: probable flood height(m)**	**LP3 distribution: probable flood height(m)**
2	0.5	0.5	5.74	5.75(5.74)	5.75
5	0.2	0.8	7.47	7.47(7.45)	7.48
10	0.1	0.9	8.52	8.47(8.46)	8.50
25	0.04	0.96	9.72	9.64(9.63)	9.67
50	0.02	0.98	10.54	10.45(10.45)	10.48
100	0.01	0.99	11.30	11.21(11.22)	11.23
200	0.005	0.995	12.00	11.94(11.95)	11.94
250	0.004	0.996	12.22	12.17(12.18)	12.16
500**	0.002	0.998	12.85	12.87(12.88)	12.83
1000	0.001	0.999	13.45	13.54(13.56)	13.47

* The values in brackets are the return levels for the three-parameter gamma (Ga3) and are almost identical to the two-parameter gamma (Ga2).
** The 13 m flood height is over the 500-year flood level at Combomune. The highest flood height at the site, 10.97 m, is less than the 100-year flood height.

The results in Table 3.7 revealed that the flood height estimates of the three distributions selected for the site are very close indicating that the three distributions are almost indistinguishable in terms of their ability to model flood heights for the Combomune site. The 100-year flood height for the Combomune site is just about 11.30 m which is very close to that of Chokwe downstream. The 13 m flood height which occurred in the basin during the February-March 2000 floods is above the 500-year flood height at the site based on the three distributions selected for the site. These results are consistent with those for Chokwe where the 100-year flood height is about 11 m and the 13 m of flood height of 2000 had a return period of over 500 years.

Sicacate quantile estimation based on the best three distributions selected

The quantile estimation results for Sicacate are presented in Table 3.8. The table presents at-site results of the expected return periods and the corresponding high quantiles of annual maximum daily flood heights for Sicacate hydrometric station.

Table 3.8: Results of the expected return periods and probable high quantiles of annual maximum daily flood heights at Sicacate

Expected return period: T(years)	**Exceedance probability: p**	**Non-exceedance probability or CDF: F(x)**	**GEV distribution: probable flood height(m)**	**Gumbel Min distribution: probable flood height(m)**	**GG distribution: probable flood height(m)**
2	0.5	0.5	7.33	7.54	7.62
5	0.2	0.8	9.84	9.56	9.88
10	0.1	0.9	10.91	10.42	10.78
25	0.04	0.96	11.82	11.23	11.58
50	0.02	0.98	12.28	11.69	12.03
100	0.01	0.99	12.61	12.10	12.39
200	0.005	0.995	12.84	12.43	12.70
250*	0.004	0.996	12.90	12.53	12.78
500	0.002	0.998	13.06	12.81	13.04
1000	0.001	0.999	13.17	13.10	13.26

* The 13 m flood height is over the 250-year flood level at Sicacate.

The results in Table 3.8 revealed that the annual flood height estimates of the the three selected distributions for Sicacate were very close to each other indicating the three distributions selected for the site provide equally good estimates for the site. However, for very high quantiles the GG distribution exhibited a tendency to approach towards higher values at a faster rate and therefore lack stability. On the other hand, the Gumbel distribution showed a tendency to approach higher values at a slower rate and therefore likely to un-

derestimate extremely high flood heights. These results are in agreement with the theoretical arguments concerning the limitations of the Gumbel distributions in modelling the largest extreme hydrological events raised in literature by Koutsoyiannis (2004) and Rowinski and Strupczewski (2001).

The 100-year flood height based on the flood height estimates for the distributions selected at the site is just less than 12.65 m, and the 13 m flood height that occurred at the site during the February-March 2000 floods had a return period in excess of 250 years. These results are consistent for all the three distributions selected for Sicacate site, but sufficiently differ from the results obtained from Chokwe and Combomune. The results for Sicacate indicated that the 100-year flood height at the site is sufficiently higher than that for Chokwe and Combomune which is approximately 11 m. The higher estimates of flood heights at Sicacate seemed to be consistent with the fact that the observed flood heights at the site were generally higher than those for Chokwe and Combomune (Table 3.1).

3.4 Concluding remarks

Flood height magnitudes and return periods were developed to give the annual maximum daily series for the Limpopo River at Chokwe, Combomune and Sicacate hydrometric stations in Mozambique. The respective 60-year, 45-year and 59-year annual maximum daily flood height data series for Chokwe, Combomune and Sicacate hydrometric stations in the lower Limpopo River basin were used to develop the magnitudes and frequencies of floods at the sites. Ten candidate probability distribution models including the GEV, two-parameter Weibull, three-parameter Weibull, Gumbel, two-parameter lognormal, three-parameter lognormal, two-parameter gamma, three-parameter gamma, generalised gamma, and log-Pearson type 3 were compared for modelling annual

maximum daily flood height for lower Limpopo River basin using EasyFit Professional statistical software.

The results for both analytical and graphical goodness-of-fit analyses indicated that the three-parameter GEV is suitable for flood frequency analysis at Sicacate, while the two-parameter Gumbel Min also provided quite good fits at Sicacate as a suitable alternative model. The other alternative distribution at Sicacate is the generalised gamma distribution. The GEV, three-parameter gamma and lognormal are suitable for modelling flood frequency at Chokwe, while for Combomune the suitable distributions are GEV, two-and-three-parameter gamma and log-Pearson type 3. The GEV distribution was among the three best distributions at all the three sites. This is an indication that the GEV distribution can be considered as the overall model for modelling extreme flood heights in the basin. The consistency of the GEV as the prevailing distribution that emerged among the best fitting distributions at all the sites concur with the findings by Sukla et al. (2014) who performed flood frequency analysis of some rivers in India and found the GPD and GEV distributions as the prevailing distributions for all the five rivers studies. Therefore, this present study, with the support of the study by Sukla et al. (2014) and others mentioned in literature, has confirmed the superiority of the GEV distribution in flood frequency analysis. Based on these findings, the GEV distribution shall be considered as the main distribution for the basin throughout the next chapters.

It appears that the two-parameter Gumbel distribution which did not pass the goodness-of-fit tests at Chokwe and Combomune, but passed the tests at the downstream Sicacate appear to have difficulties in modelling the skewness of the flood height samples throughout the lower Limpopo River basin due to its fixed shape and lesser number of parameters (Singo et al., 2012). These findings are also in agreement with the theoretical arguments which criticise the

use of the Gumbel distribution in modelling hydrological extremes claiming that it underestimates the largest extreme flood heights (Koutsoyiannis, 2004; Rowinski and Strupczewski, 2001). The findings based on the Gumbel estimates for Sicacate in Table 3.8 confirm these theoretical arguments by Koutsoyiannis (2004) and Rowinski and Strupczewski (2001).

The 13 m flood height of the February-March 2000 has a return period in excess of 250 years or approximately 500 years to be precise, implying that it was indeed a very rare event. Our results are generally in agreement with the findings by Smithers et al. (2001) who studied the February 2000 floods in the Sabie River Catchment upstream of the South Africa-Mozambique border which is in the upper Limpopo River basin. Smithers et al. (2001) used the GEV distribution and obtained return periods in excess of 200 years using durations ranging from 1 to 7 days for both extreme rainfall and flood discharges. The present study used annual maximum daily flood height data series which is quite different from the data approach used by Smithers et al. (2001) and Singo et al. (2012), both in South Africa. The interesting finding in the study for this chapter is that the conclusions reached using the annual maxima flood height data for the lower Limpopo River basin of Mozambique were similar to the conclusions made by other researchers in the region, particularly in river basins and catchments of South Africa using different data sets and approaches (Alexander, 2002; Smithers et al., 2001).

Furthermore, the approaches used by Sukla et al. (2014) and Alam and Khan (2014), among others mentioned earlier in the chapter, were different from the approach used in this present study, but the findings and conclusions reached share a high degree of concurrence. For example, Sukla et al. (2014), like this study, found the GEV distribution as the prevailing distribution in modelling the daily rainfall amount in the Mahanadi Delta region of Odisha in

India, while Alam and Khan (2014) used a slightly different approach to study the water levels and discharges of five peripheral rivers around Dhaka city in Bangladesh and found the log-normal and Gumbel distributions unsuitable for modelling the flood frequency distribution of the rivers around the city of Dhaka. These conclusions were also reached in the present study for the lower Limpopo River basin of Mozambique.

3.5 Summary of the chapter

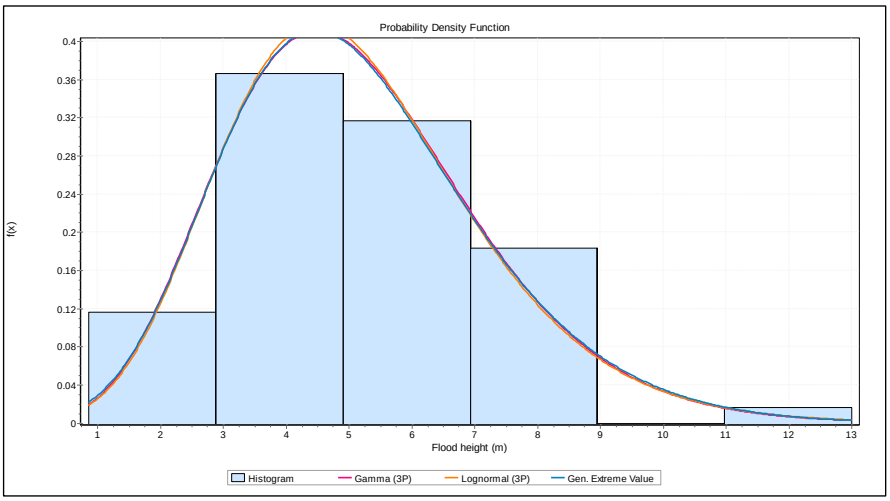
In this chapter ten candidate distributions were compared for their goodness-of-fit in modelling the annual maxima daily flood heights for the lower Limpopo River basin of Mozambique. The ten distributions compared in this chapter were the generalised extreme value, generalised gamma, two-parameter Weibull, three-parameter Weibull, two-parameter gamma, three-parameter gamma, Gumbel, two-parameter log-normal, three-parameter log-normal and three-parameter log-Pearson distribution. The three-parameter gamma, three-parameter log-normal and generalised extreme value were the best three probability distributions for Chokwe, while generalised extreme value, two-parameter gamma and three-parameter log-Pearson were the best three fitting probability distributions for Combomune. The three-parameter gamma can also be an used as an alternative distribution to model the annual maxima flood heights for the Combomune site. The generalised extreme value, Gumbel and generalised gamma distribution were the best three probability distributions for Sicacate based on their ability to model the tails of the empirical distribution. The parameter estimation methods used were the maximum likelihood estimators method, method of moments, and L-moments, among others. Goodness-of-fit was evaluated by means of Kolmogorov-Smirnov and Anderson-Darling tests, as well as probability-probability, quantile-quantile, probability difference plots and simulation studies to check whether the distribution could mimic

the observed values of the empirical distribution. Results of the probable return periods and probable high quantiles at all the three sites Chokwe, Combomune and Sicacate indicated that the 13 m flood height of the February-March 2000 was way higher than the 100-year flood height. It was also found that this 13 m flood height had a return period in excess of 250 years based on the estimates from best fitting distributions for Sicacate, and a return period in excess of 500 years based on the findings from the estimates of the best fitting distributions for Chokwe and Combomune, implying that this rare flood height has a very small likelihood of being equaled or exceeded at least once in over 250 years.

APPENDIX 3.1: DIAGNOSTIC PLOTS, PDF AND CDF PLOTS

This page is purposely made blank.

Panel A



Panel B

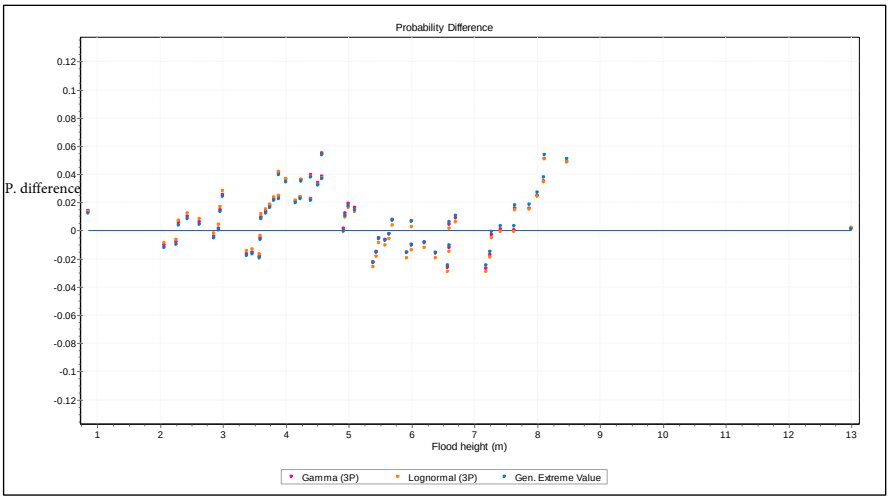
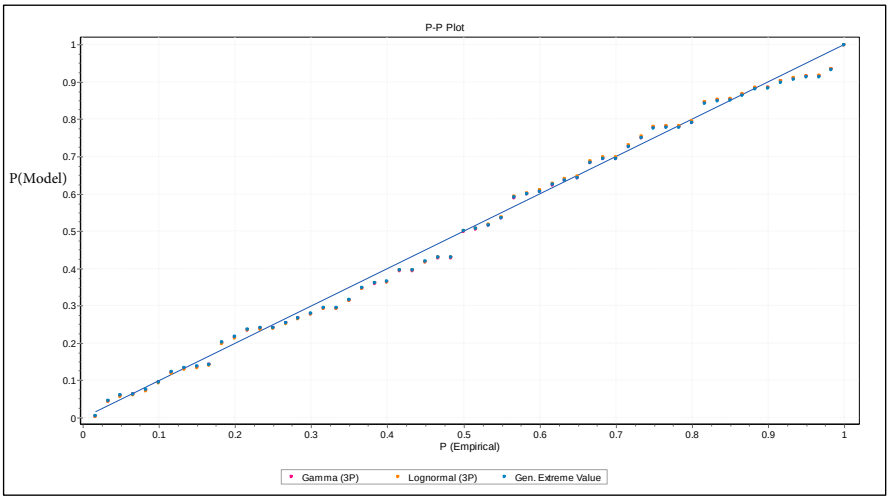


Figure 3.1: Panel A: Probability density functions (PDFs) of the best three fitting distributions and; Panel B: Probability difference of best three fitting distributions for Chokwe.

Panel A



Panel B

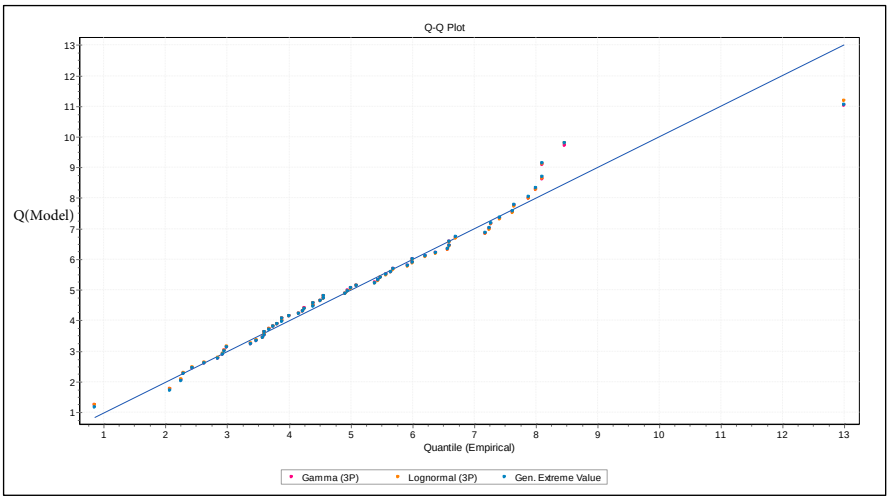
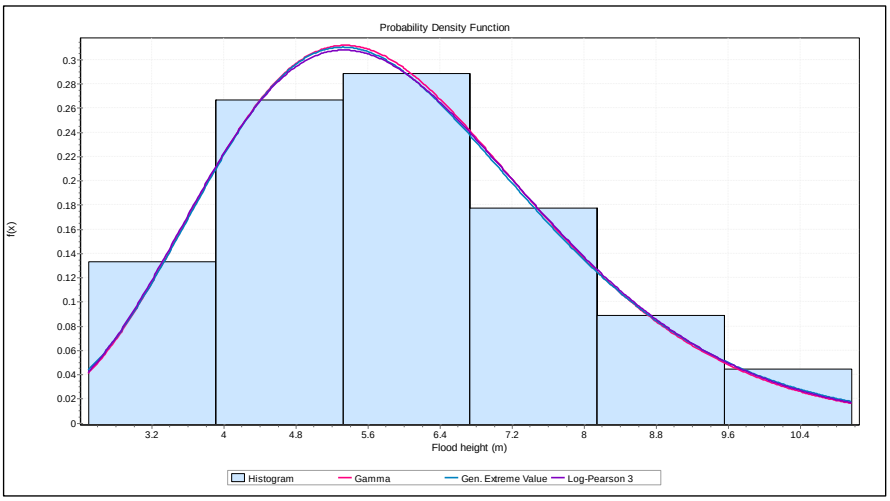


Figure 3.2: Panel A: Probability-probability (P-P) plot and; Panel B: Quantile-quantile (Q-Q) plot of the best three fitting distributions for Chokwe.

Panel A



Panel B

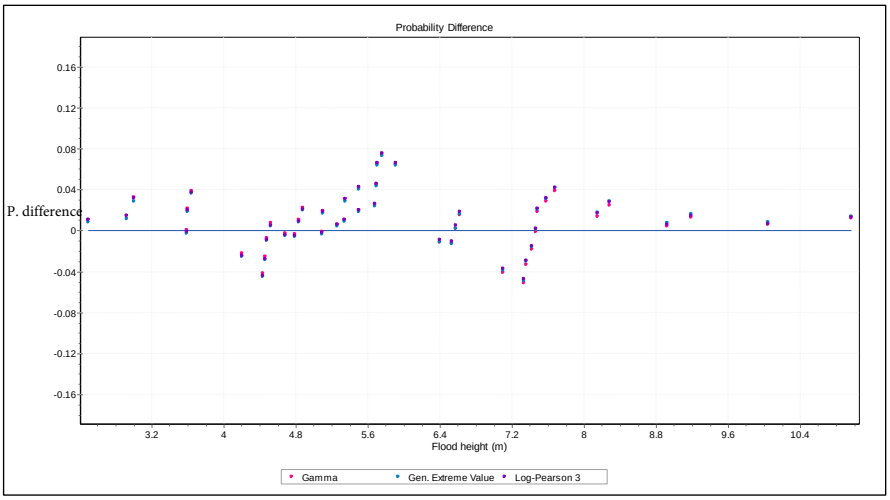
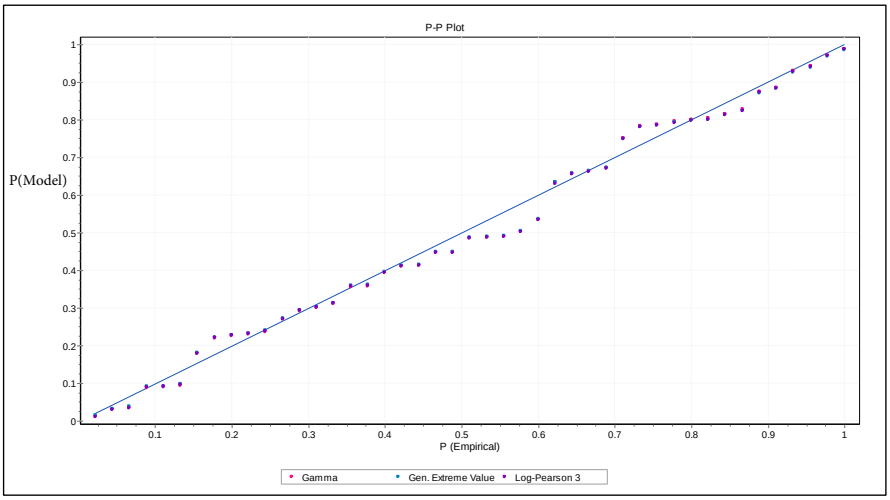


Figure 3.3: Panel A: Probability density functions (PDFs) of the best three fitting distributions and; Panel B: Probability difference of best three fitting distributions for Combomune.

Panel A



Panel B

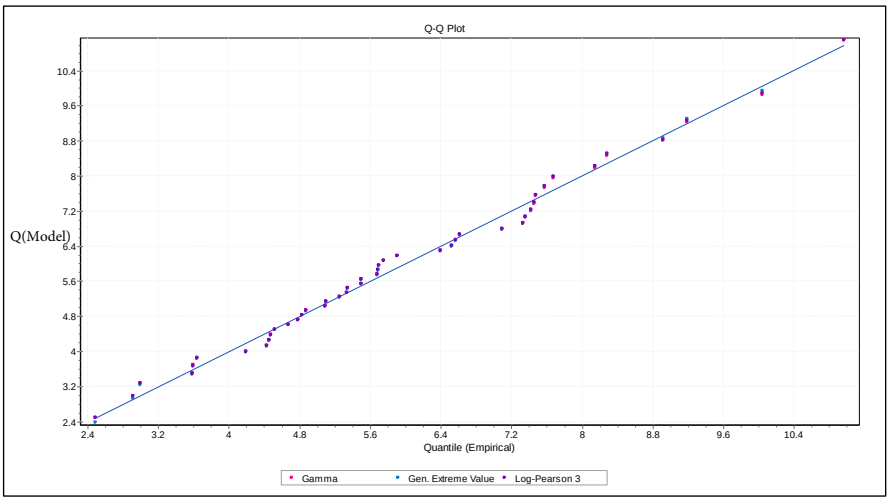
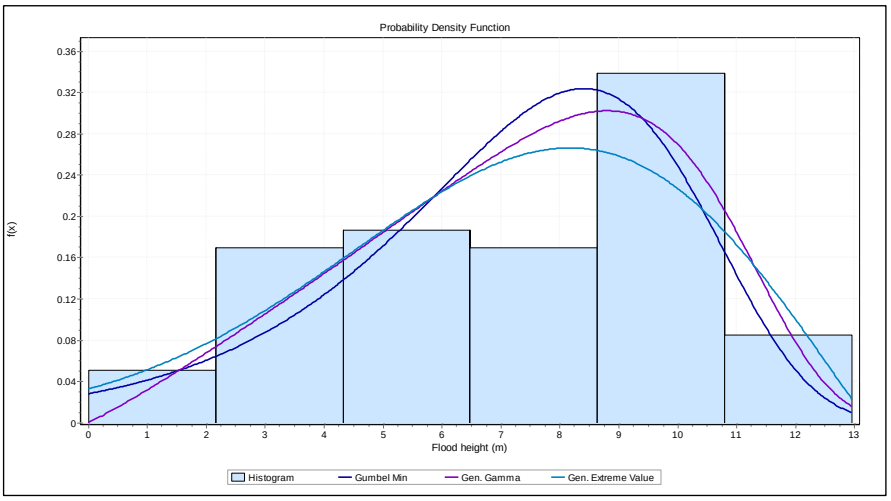


Figure 3.4: Panel A: Probability-probability (P-P) plot and; Panel B: Quantile-quantile (Q-Q) plot of the best three fitting distributions for Combomune.

Panel A



Panel B

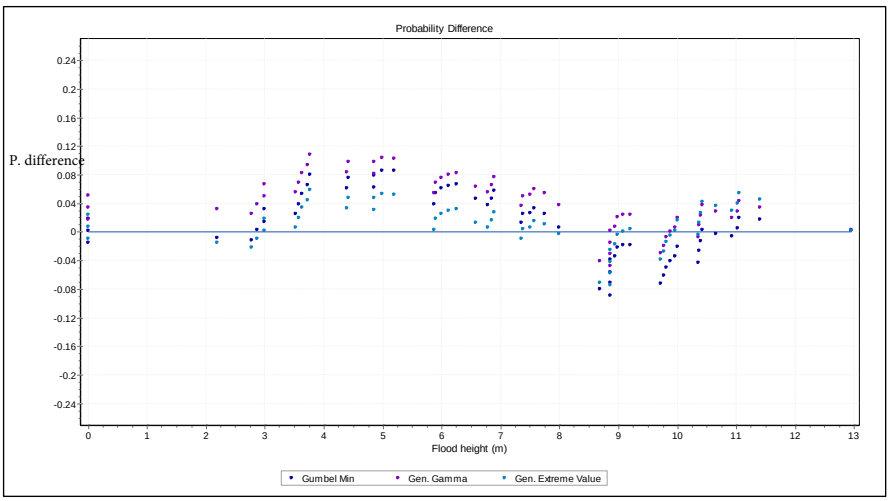
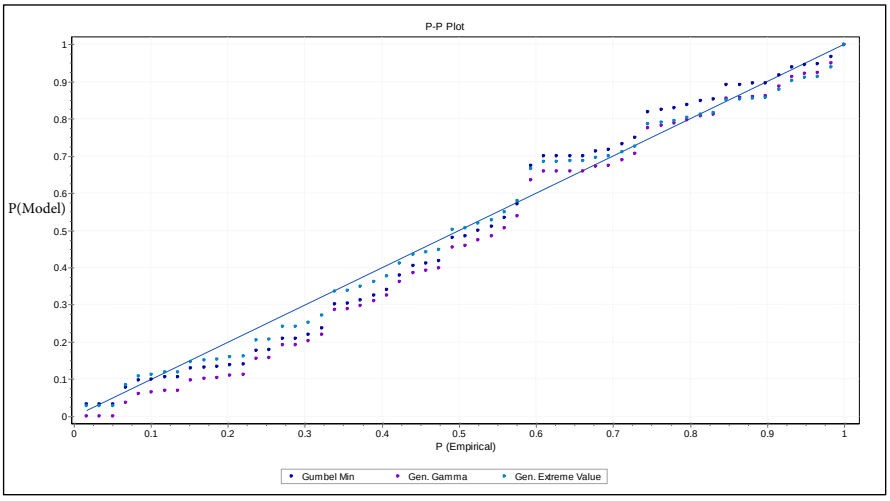


Figure 3.5: Panel A: Probability density functions (PDFs) of the best three fitting distributions and; Panel B: Probability difference of best three fitting distributions for Sicacate

Panel A



Panel B

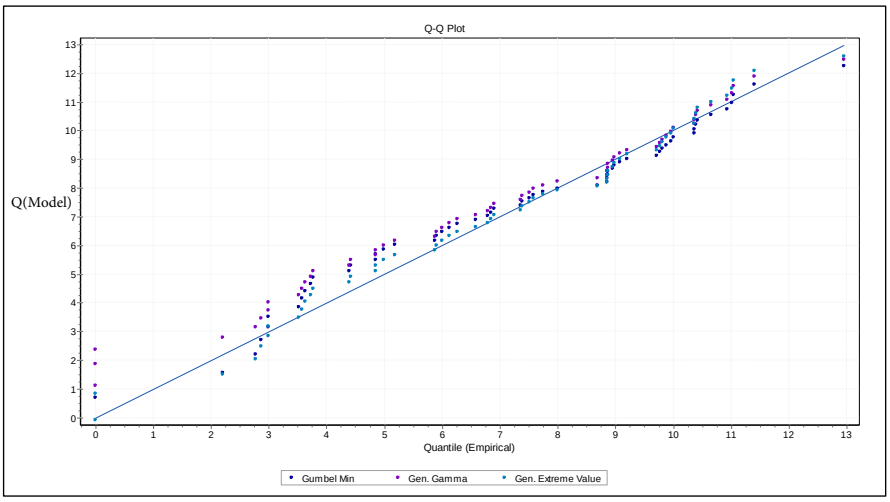
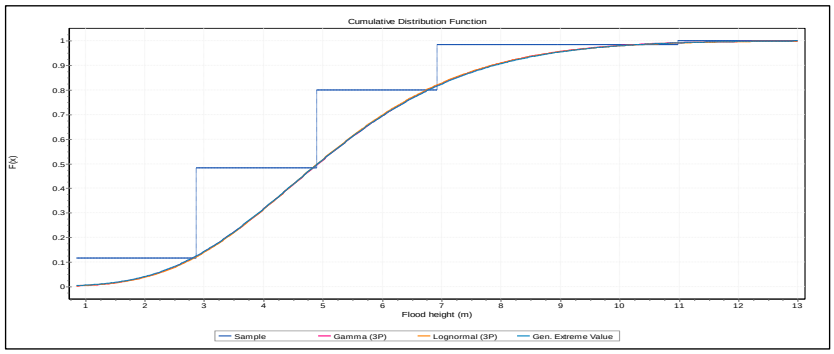
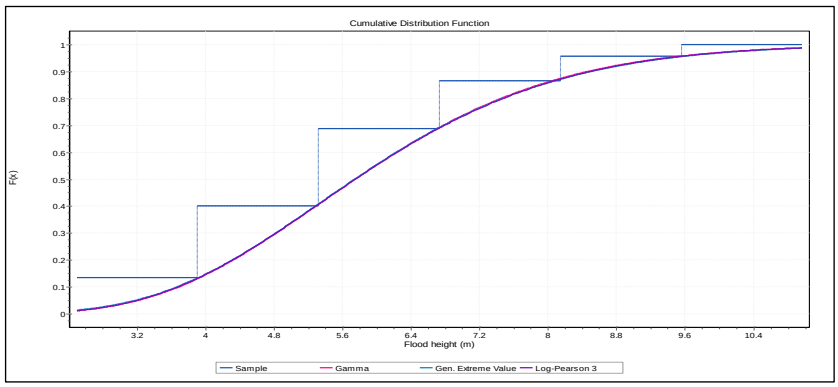


Figure 3.6: Panel A: Probability-probability (P-P) plot and; Panel B: Quantile-quantile (Q-Q) plot of the best three fitting distributions for Sicacate.

Panel A



Panel B



Panel C

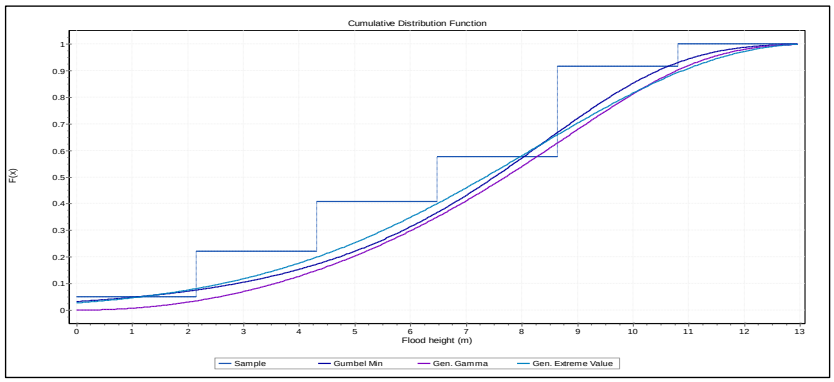


Figure 3.7: Comparison of the CDFs (or non-exceedance probabilities) of the best three performing distributions at each site; Panel A: Chokwe non-exceedance probabilities, Panel B: Combomune non-exceedance probabilities, and Panel C: Sicacate non-exceedance probabilities.

Chapter 4

A comparative analysis of annual maxima time series models along the lower Limpopo River basin of Mozambique

4.1 Introduction

The recent International Disaster and Risk Conference (IDRC) held in Davos, Switzerland, August 24–28, 2014, emphasised the need for collaborative efforts in disaster risk reduction and building of resilient communities (Stal et al., 2014). In his welcome speech, the IDRC Davos 2014 Chairman, Dr. Walter J. Ammann, pointed out that the scope, intensity, and complexity of risks of natural disasters such as floods and earthquakes were on the rise in recent years (WMO, 2013). Smakhtin (2014) presented a paper at the IDRC in which he

pointed out that floods and droughts are the major causes of destruction with regard to crop damage and loss of life. The author indicated that floods and droughts account for about 90% of all the people who are affected by natural disasters. Details of the hydrology of the LLRB were presented in Chapter 3. The Director-General of UNESCO, Ms. Irina Bokova, stated that the main goal of the 5^{th} IDRC Davos 2014 was to craft solutions to the challenges that are currently faced by societies, through risk management, disaster reduction and adaptation to climate change (Stal et al., 2014, p.3).

In another foreword speech at IDRC Davos 2014, Ms. Sally Fegan-Wyles, the UN Assistant Secretary-General, argued that the vulnerability of the world to natural disasters is increasing. She argued that despite massive development made globally many people still lose their lives to disasters, trillions of dollars continue to be lost to disasters and the cost of emergency response continues to rise (Stal et al., 2014).

The purpose of this chapter is to perform a comparative analysis of the annual maximum sums of flood heights in the lower Limpopo River basin at each of the three sites, Chokwe, Combomune and Sicacate hydrometric stations. The annual maximums considered in this study are the daily (one-day), two-day, five-day, seven-day, ten-day, and thirty-day annual maximums at each site. The assessment of the six annual maximum time series models was performed with the goal of investigating whether there are significant differences in characteristics among the six time series models at each of the three sites.

Recent advances in block maxima were derived in Ferreira and de Haan (2015), and consistency of MLEs based on block maxima was proved in Dombry (2015). McMahon et al. (2007) studied the annual stream flow characteristics of a set of 1,221 global rivers distributed worldwide, including the Zambezi River in Zim-

babwe, Southern Africa. The annual flow features examined by McMahon et al. (2007) are the mean, variability, skewness, distribution type (gamma and/or log-normal) and flow percentiles. The findings by McMahon et al. (2007) highlighted differences in annual stream flow characteristics between Australia-Southern Africa (ASA) and the rest of the world (ROW). The approach used by McMahon et al. (2007) was quite different from the approach used in this chapter in a number of ways. Firstly, the majority of the features examined in this chapter are different from those in McMahon et al. (2007) except for the skewness and coefficient of variation. Secondly, while in this chapter we compare the features among annual maximum time series models of the same river at different sites, McMahon et al. (2007) compared the features among different rivers. Baratti et al. (2012) performed flood frequency distribution analyses at seasonal and annual scales. While the approach used by Baratti et al. (2012) can be useful and applicable in other regions, it may not be relevant in Southern Africa, particularly the Limpopo River, where there are absolutely no floods during the dry season, and all annual maxima belong to the rainy season.

In a study more similar to this chapter, Machiwal and Madan (2008) performed a comparative evaluation of twenty-nine statistical tests used to detect hydrological time series characteristics by using them to analyse forty-six years of annual rainfall and forty-seven years of one-day, two-day, three-day, four-day, five-day, and six-day maximum rainfalls at Kharagpur in India. The tests revealed homogeneity in the seven rainfall series, and the time series plots revealed no evidence of trends for any of the seven rainfall series. Machiwal and Madan (2008) emphasised the evaluation of statistical tests, but at the same time gave a caution against using too many statistical tests for the same objective in time series analyses because this increases the probability of committing a Type-I Error - that is, incorrectly rejecting the null hypothesis when it is true. McMahon et al. (2007) recommended using at least two statistical

tests (but not too many) when making decisions about rejecting the null hypothesis. Instead of using rainfall data, our study uses hydrometric data to make inferences about the distribution of extreme flood heights in the LLRB of Mozambique.

The rest of the chapter is arranged as follows: Section 4.2 presents the research methodology used in the study for this chapter, Section 4.3 presents the results and discussion of the findings, Section 4.4 summarises the value added by this chapter to management and disaster risk reduction. Section 5 gives the concluding remarks and finally Section 4.6 gives a summary of the chapter.

4.2 Research methodology

In this section we present the sequential steps used to obtain the block maxima data and moving sums of block maxima, descriptive measures of variability, and the goodness-of-fit (GoF) tests for the GEV distribution.

4.2.1 Moving sums and block maxima

Similar to Chapter 3, the data used in the study for this Chapter are hydrometric daily flood heights recorded at Chokwe (1951–2010), Combomune (1966-2010), and Sicacate (1952–2010), hydrometric stations for the LLRB of Mozambique. The raw data were recorded as daily flood heights. Because our aim was to compare several annual maximum time series models, sequential steps were taken to obtain the two-day, five-day, seven-day, ten-day, and thirty-day moving sums. Further sequential steps were taken to obtain the annual maximum flood heights series for each of the moving sums, including the daily flood heights series. Finally, the following annual maximum flood heights time series models were generated: AM1, AM2, AM5, AM7, AM10, and AM30.

The approach used to determine the annual maximum time series models is known as block maxima. The block maxima (or at-site) approach in FFA is usually preferred to the POT approach when the data records have sufficiently large sample sizes and the quality of the data is adequate (Ferreira and de Haan, 2015). The data used in this study had sufficiently large annual maxima records, extending to 60 years for the Chokwe hydrometric station, 59 years for Sicacate, and 45 years for Combomune. Naturally, in hydrology, when a sample size is sufficiently large, observations are grouped (blocked) by years (Ferreira and de Haan, 2015; Beirlant et al., 2004).

The six annual maxima time series models at the three sites were assessed based on selected descriptive statistics measures namely; skewness, excess kurtosis, coefficient of variation (CV), and the GoF of the GEV distribution, particularly the Anderson-Darling (A-D) and Kolmogorov-Smirnov (K-S) statistics. The main purpose of this assessment was to investigate whether there were significant differences between the six annual maxima time series models with respect to skewness, excess kurtosis, CV, A-D, and K-S statistics.

Time series plots, probability density plots and boxplots were used in the visual assessment of the models. The GEV distribution was fitted to all the annual maxima models and its A-D and K-S statistics were recorded. An analysis of variance was performed on the models at both sites using the six annual maxima time series models as treatments (Pretorius, 2007). A correlation analysis of the six annual maxima time series models was also performed at each site to check for significant correlations among the models.

4.2.2 Overview of the theoretical models

The detailed probability framework of the block maxima is derived in Dombry (2015) and Ferreira and de Haan (2015). The theoretical set up of the block

maxima approach is presented in Chapter 2, Subsection 2.7.1.

In this chapter, the GoF of the GEV distribution in (2.6) was assessed by the A-D and K-S tests, which indicate whether or not it is reasonable to assume that a random sample data comes from a specified distribution (in this case GEV) based on the following null and alternative hypotheses (Ricci, 2005):

H_0: Sample data come from the specified distribution, versus H_1: Sample data do not come from the specified distribution.

Rejecting H_0 would imply that the specified distribution is not a good fit to the sample data. The A-D test is sensitive to the tails of the distribution, while the K-S test is sensitive to the center of the distribution.

The Pearson's skewness and kurtosis coefficients are given respectively in (4.1)

$$\gamma_1 = \frac{\sum_{i=1}^{n}(x-\mu)^3}{n\sigma^3} \text{ and } \gamma_2 = \frac{\sum_{i=1}^{n}(x-\mu)^4}{n\sigma^4} - 3. \tag{4.1}$$

Ricci (2005) argued that when the data is standardised, the distribution curves depend mainly on skewness and kurtosis measures. Excess kurtosis refers to kurtosis above or below the normal value, which is taken to be 3 (DeCarlo, 1997).

The coefficient of variation, commonly known as CV in statistics and probability theory, is a measure of relative variability or dispersion of data points relative to the mean in a data series. It is calculated using the formula in (4.2)

$$CV = \frac{s}{\bar{x}} \times 100\%. \tag{4.2}$$

The CV is a very useful statistic in making comparisons of the extent of vari-

ation from one data series to another, even if the means of the series are very different from each other, such as the case in the study for this chapter.

Detailed information on analysis of variance and correlation coefficient is given in (Pretorius, 2007). The analysis of variance serves as an extension of the independent samples t-test when more than two groups (annual maxima time series models) are compared. The null hypothesis states that the annual maximum time series models do not differ with regard to their means, while the alternative hypothesis claims that the annual maxima time series models differ (or at least one pair of annual maxima time series models differ) with regard to their means assessed on skewness, excess kurtosis, CV, A-D and K-S statistics as in (4.3)

$$H_0 : \mu_1 = \mu_2 = ... = \mu_6 \text{ versus } H_1 : \mu_i \neq \mu_j, \text{ for } i \neq j \text{ and} i, j = 1, 2, ... \quad (4.3)$$

The correlation coefficient between two quantitative variables or Pearson's product-moment correlation, denoted by, r, is given in Pretorius (2007) as in (4.4)

$$r = \frac{n\left(\sum xy\right) - \left(\sum x\right)\left(\sum y\right)}{\sqrt{\left[n\left(\sum x^2\right) - \left(\sum x\right)^2\right]\left[n\left(\sum y^2\right) - \left(\sum y\right)^2\right]}}, \text{ for } -1 \leq r \leq 1. \quad (4.4)$$

A value of r close to 1 indicates a very strong positive correlation, while a value of r close to -1 indicates a very strong negative correlation, and a value of r close to 0 indicates lack of correlation between the two variables (Pretorius, 2007).

4.3 Results and discussion

In this section we present, interpret, and discuss the results of our analysis. The R programming software was used to analyse the results presented in this chapter (Southworth and Heffernan, 2013; Ribatet, 2006).

4.3.1 Chokwe comparative analysis of characteristics of the annual maxima flood heights moving sums

A comparative analysis that includes visual and analytic techniques is presented in this subsection for the site of Chokwe. The results are presented in the form of graphs.

Figures 4.1–4.3 present the time series plots, probability density plots and boxplots, respectively, for Chokwe hydrometric station for the period 1951–2010. The annual maxima time series plots in Figure 4.1 (overleaf) exhibited similar variability with respect to trend, cyclic, and random variations, with a few exceptions to AM30 towards higher values at the right-end of the series. However, these minor differences between the six annual maxima time series plots appear to be marginal. All the annual maxima time series models showed that the peak flood height occurred in the year 2000.

The probability density plots in Figure 4.2 and the boxplots in Figure 4.3 exhibited positive skewness, in general, for all the six models. The probability density plots in Figure 4.2 for models AM1, AM2, AM7, AM10, and AM30 exhibited bi-modality, while the AM5 probability density plot indicated a unimodal distribution. All the boxplots in Figure 4.3 exhibited one outlier, except for AM30 which showed two outliers.

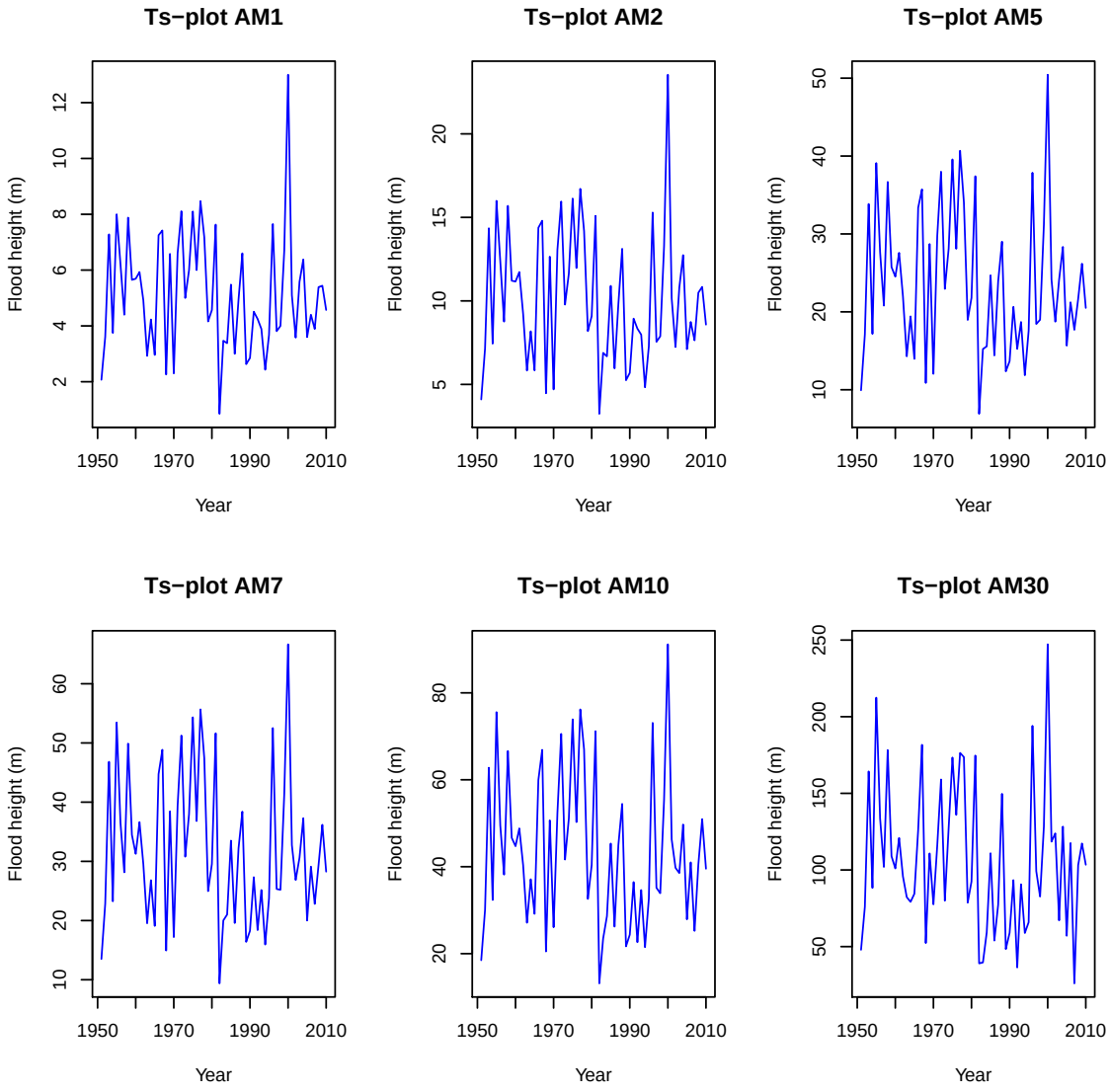


Figure 4.1: Comparison of time series plots of the annual maxima time series models for the moving sums of Chokwe

4.3.2 Combomune comparative analysis of characteristics of the annual maxima flood heights moving sums

The comparison of annual maxima time series models for Combomune that includes visual and analytic techniques are presented in this subsection. The results are presented in graphical forms.

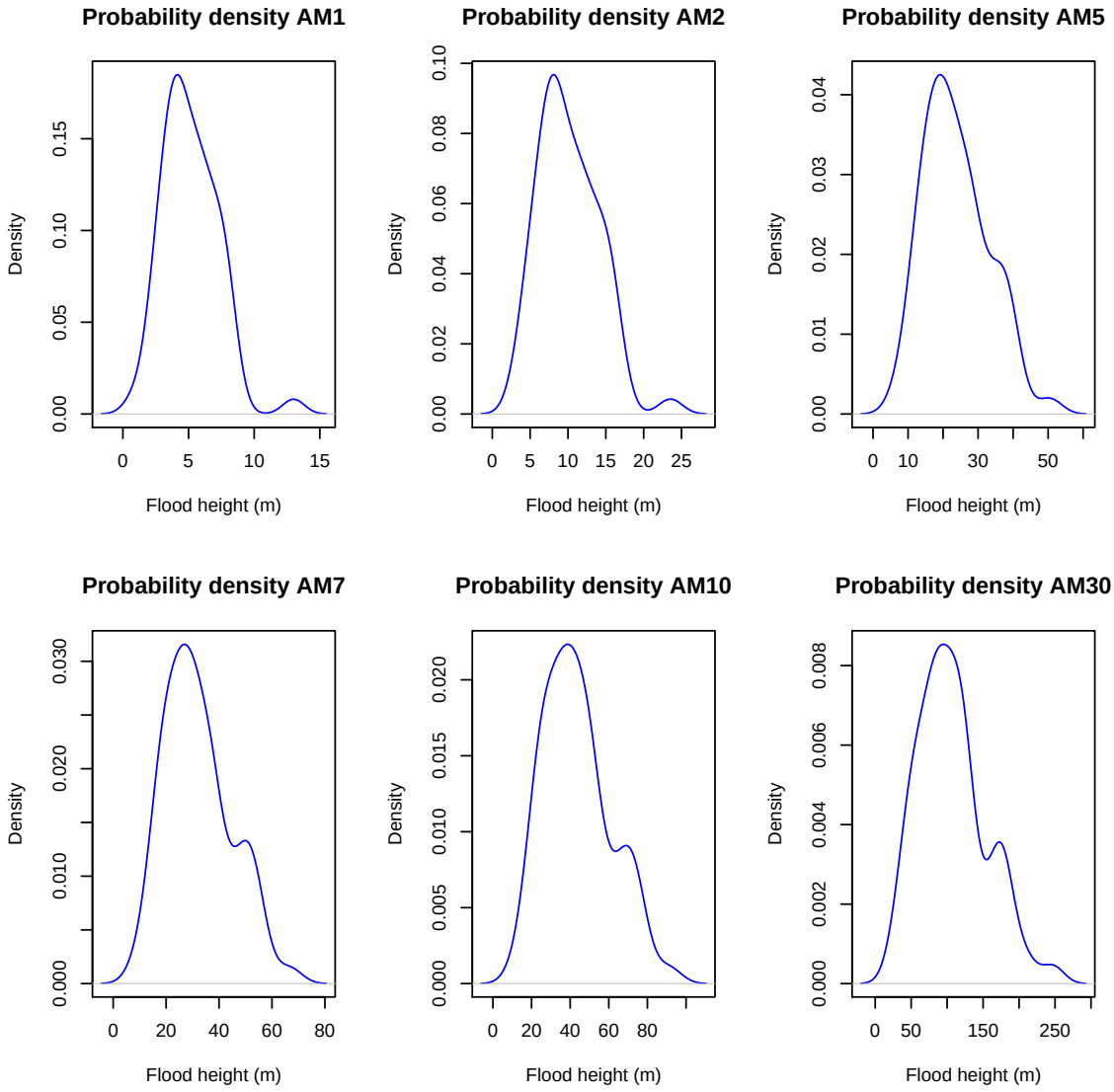


Figure 4.2: Comparison of probability density plots of the annual maxima time series models for the moving sums of Chokwe

Figures 4.4–4.6 present the time series plots, probability density plots and box-plots, respectively, for Combomune hydrometric station for the period 1966–2010. The time series plots in Figure 4.4 exhibited very similar variability with regard to trend, cyclic and random variations. All the annual maxima time series models for Combomune showed that the peak annual maximum

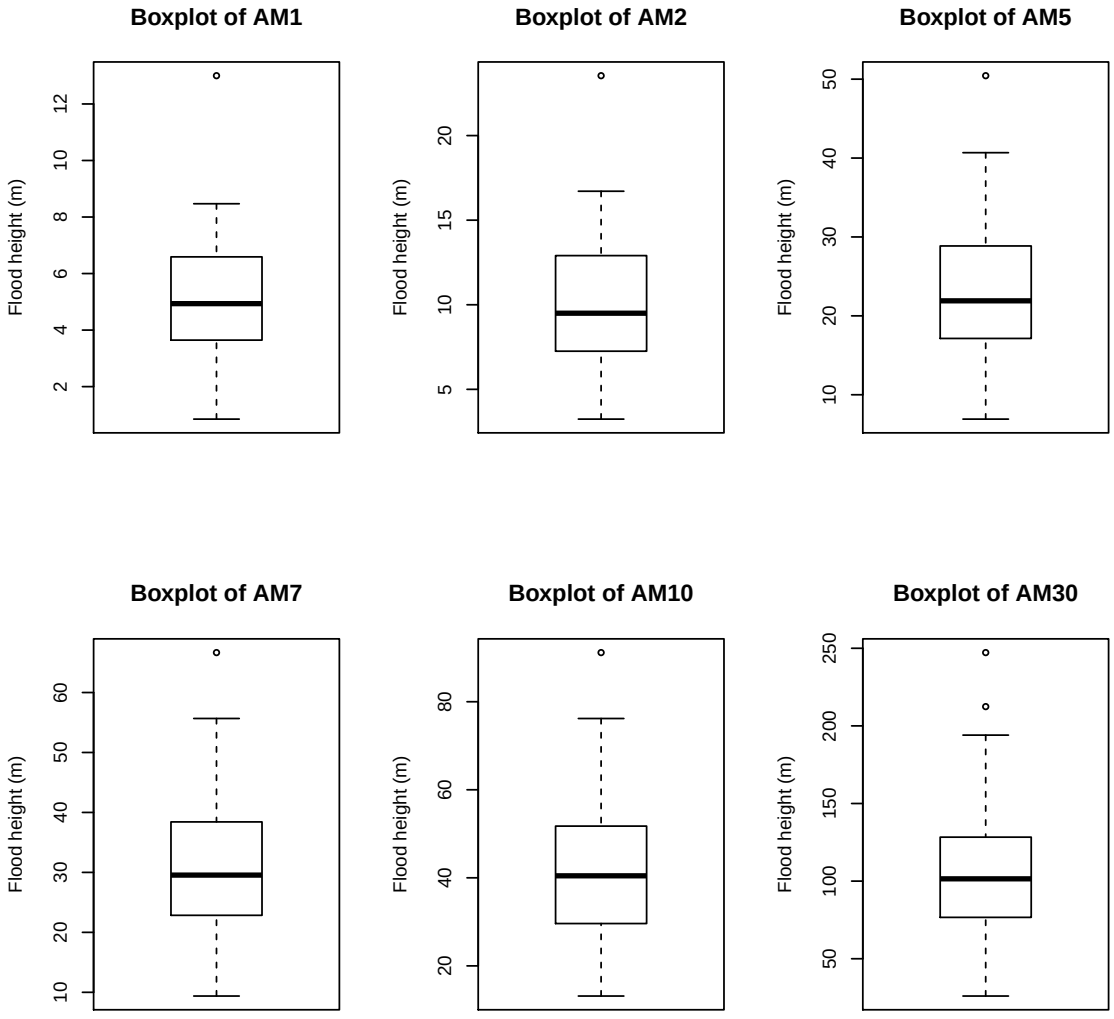


Figure 4.3: Comparison of boxplots of the annual maxima time series models for the moving sums of Chokwe

flood height occurred in the year 2000 at the site. This is consistent with the results at Chokwe, which is downstream along the same river.

The probability density plots in Figure 4.5 revealed that the shape of the distribution is consistent from AM1 to AM5 but thereafter it changed quite substantially from AM7 to AM30. The distribution of annual maxima time series

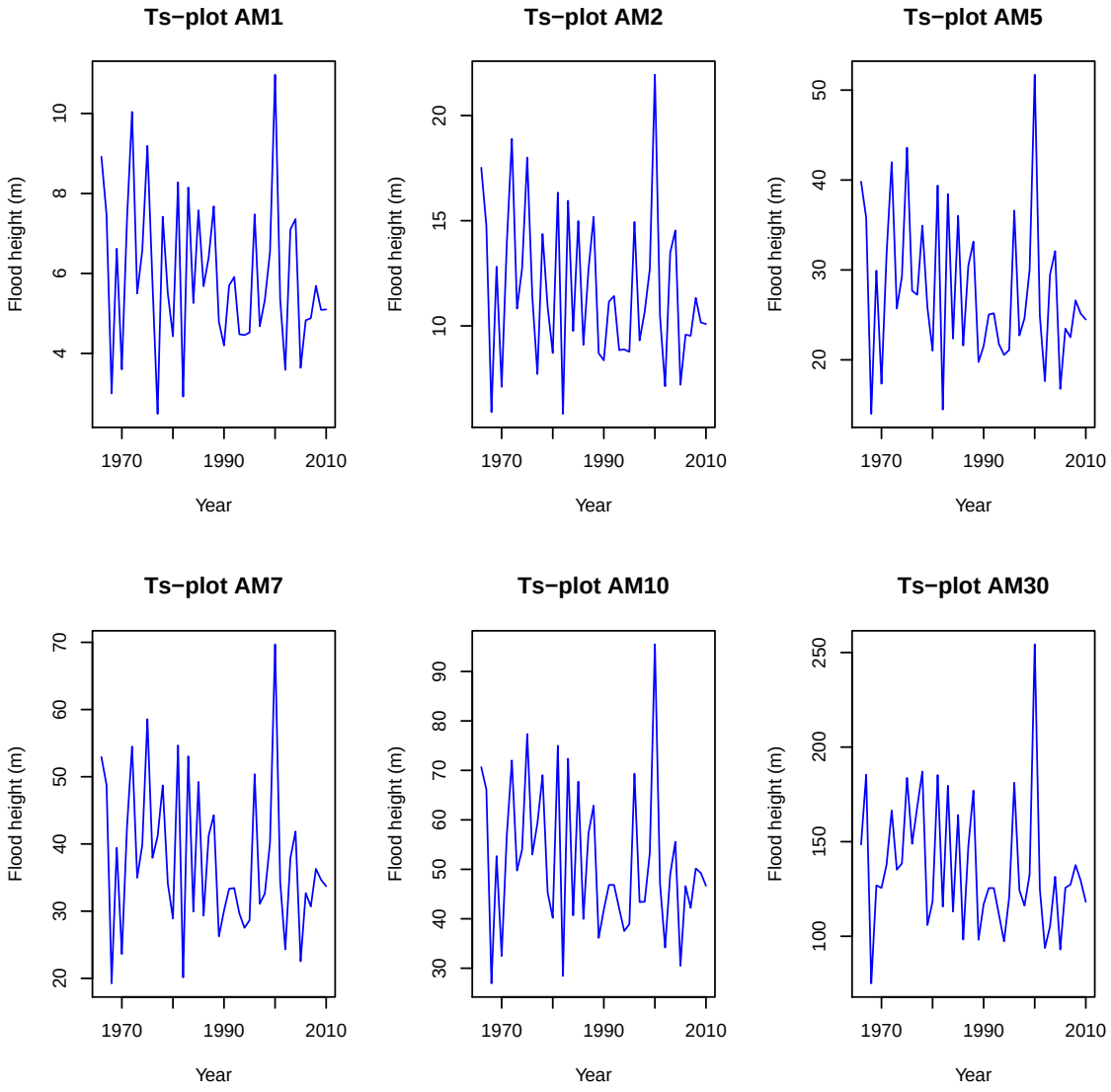


Figure 4.4: Comparison of time series plots of the annual maxima time series models for the moving sums of Combomune

models changed from unimodal (AM1-AM5) to multimodal (AM10-AM30) as the number of days of the moving sums increase. In other words, it can be concluded that based on the visual probability density plots the distribution of AM1 is sufficiently different from that of AM7, AM10, and AM30 which are more like trimodal distributions or tend towards trimodality. In addition to that, AM10 had a much similar shape to that of AM30.

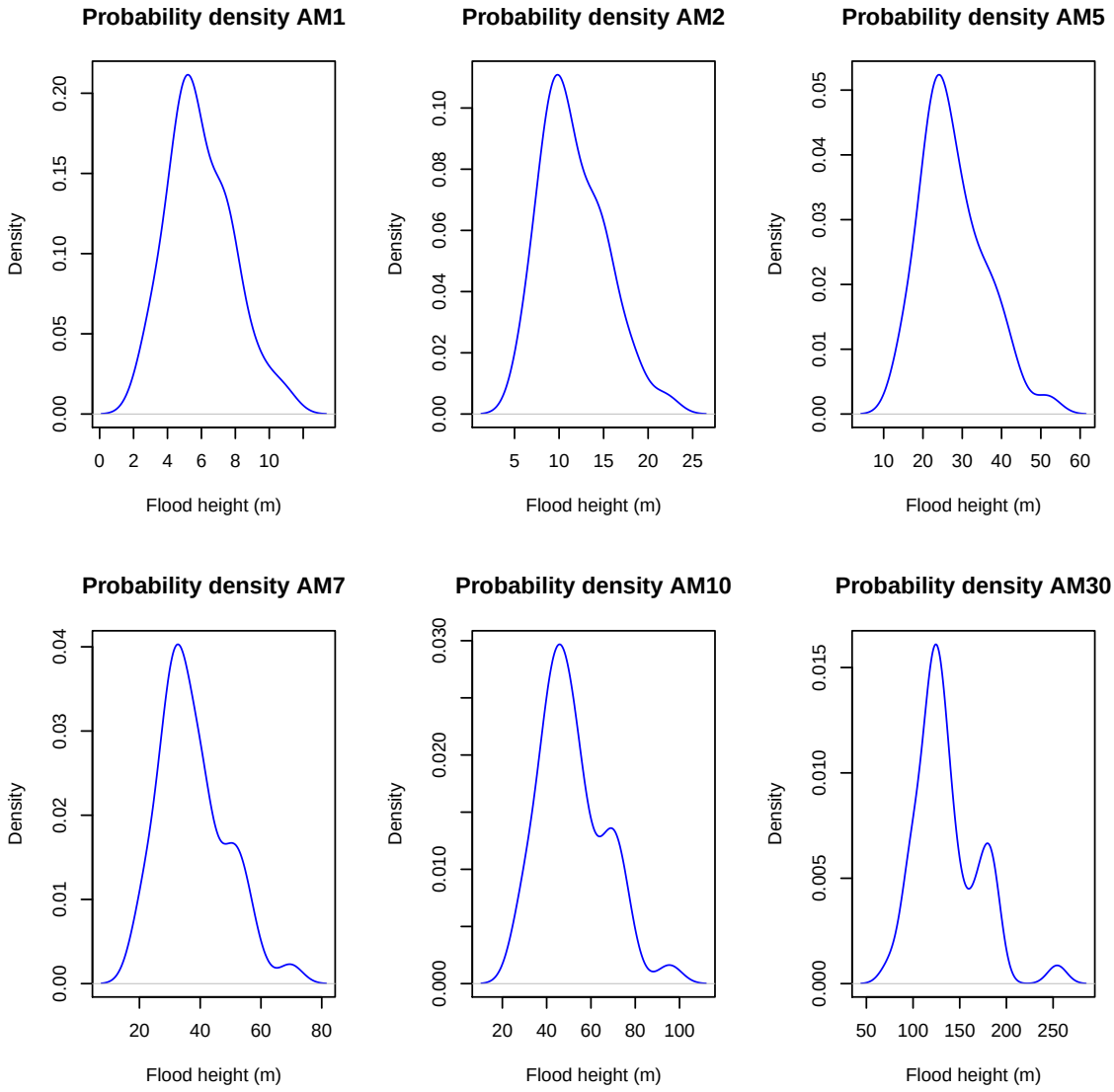


Figure 4.5: Comparison of probability density plots of the annual maxima time series models for the moving sums of Combomune

The results from Figure 4.5 revealed that the distribution of annual maxima flood heights moving sums is consistent during the first 5 days and changes rapidly thereafter at Combomune. The shape of the distribution is clearly right-skewed for the first 5 days (i.e. from A1 to A5) and from day 7 to day 30 the shape of the distribution began to exhibit multimodality. Similar re-

sults were suggested at Chokwe for the AM7, AM10, and AM30 annual maxima models.

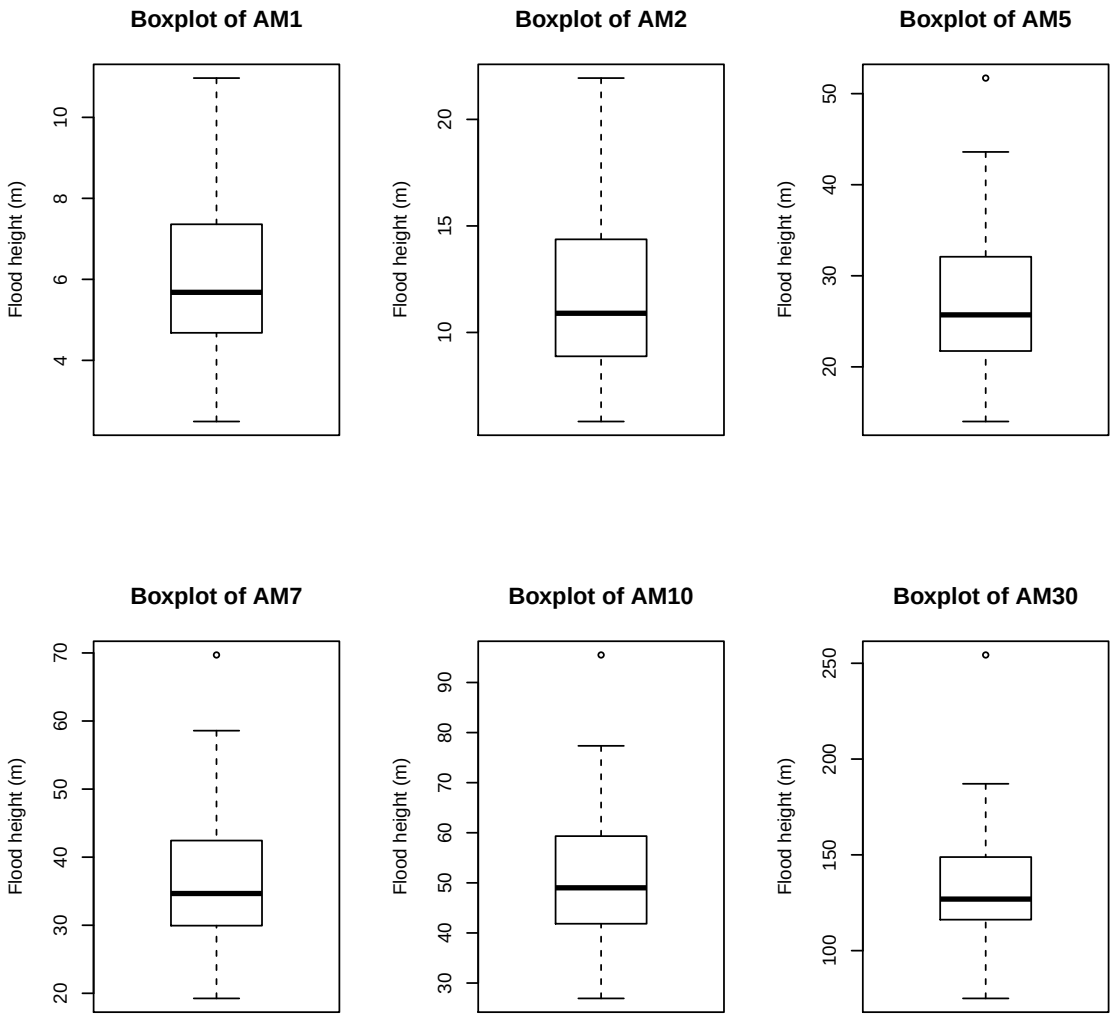


Figure 4.6: Comparison of boxplots of the annual maxima time series models for the moving sums of Combomune

The boxplots in Figure 4.6 exhibited positive skewness, in general, from AM1 through to AM7. The models AM10 and AM30 once again had very similar boxplots. At Combomune, boxplots for AM1 and AM2 showed no outliers in the

data series, but from AM5 to AM30 they showed the presence of outliers. Based on the boxplots, it can also be concluded that the distribution of annual maxima time series models changed significantly with the increase in the number days of the moving sums for the Combomune site.

4.3.3 Sicacate comparative analysis of characteristics of the annual maxima flood heights moving sums

In this subsection comparative analysis of the Sicacate annual maxima time series models is presented through the use of visual and analytic techniques. The results are presented in the form graphs.

Figures 4.7–4.9 present the time series plots, probability density plots and boxplots, respectively, for the Sicacate hydrometric station for the period 1952–2010. The time series plots in Figure 4.7 exhibited very similar variability with regard to trend, cyclic and random variations. All the annual maxima time series models showed that the peak annual maximum flood height occurred in the year 2000 at the site. This is consistent with the results at Chokwe and Combomune, which are upstream along the same river.

The probability density plots in Figure 4.8 exhibited unimodality and left-skewness for all the models except for AM30, which showed some symmetric characteristics (or slight positive skewness). The differences for all the models at Sicacate appear to be marginal.

The boxplots in Figure 4.9 indicated negative skewness for all the models except AM30, which appears positively skewed (or nearly symmetric), coinciding with the results of the probability density plots.

It is interesting to note that the shapes of the distributions of the annual max-

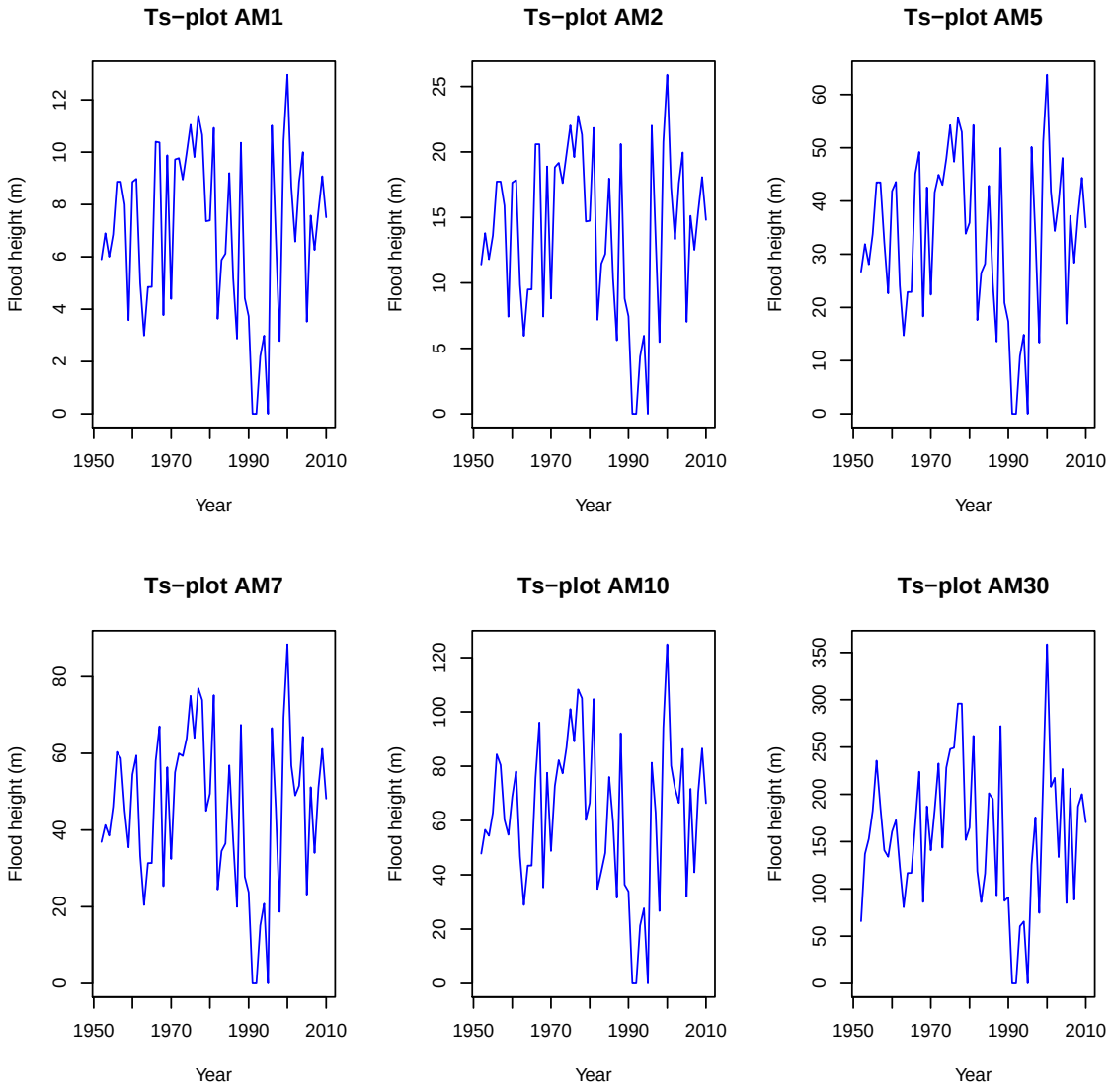


Figure 4.7: Comparison of time series plots of the annual maxima time series models for the moving sums of Sicacate

ima time series models for Chokwe and Combomune, both upstream, were positively skewed. On the contrary, the distributions of the annual maxima time series models for Sicacate were negatively skewed. Recall that in Chapter 3 the distribution of annual maxima flood heights at Sicacate was found to be negatively skewed, with a high 95^{th} percentile flood height at the site. These results are thus consistent with those in Chapter 3. The skewness towards lower val-

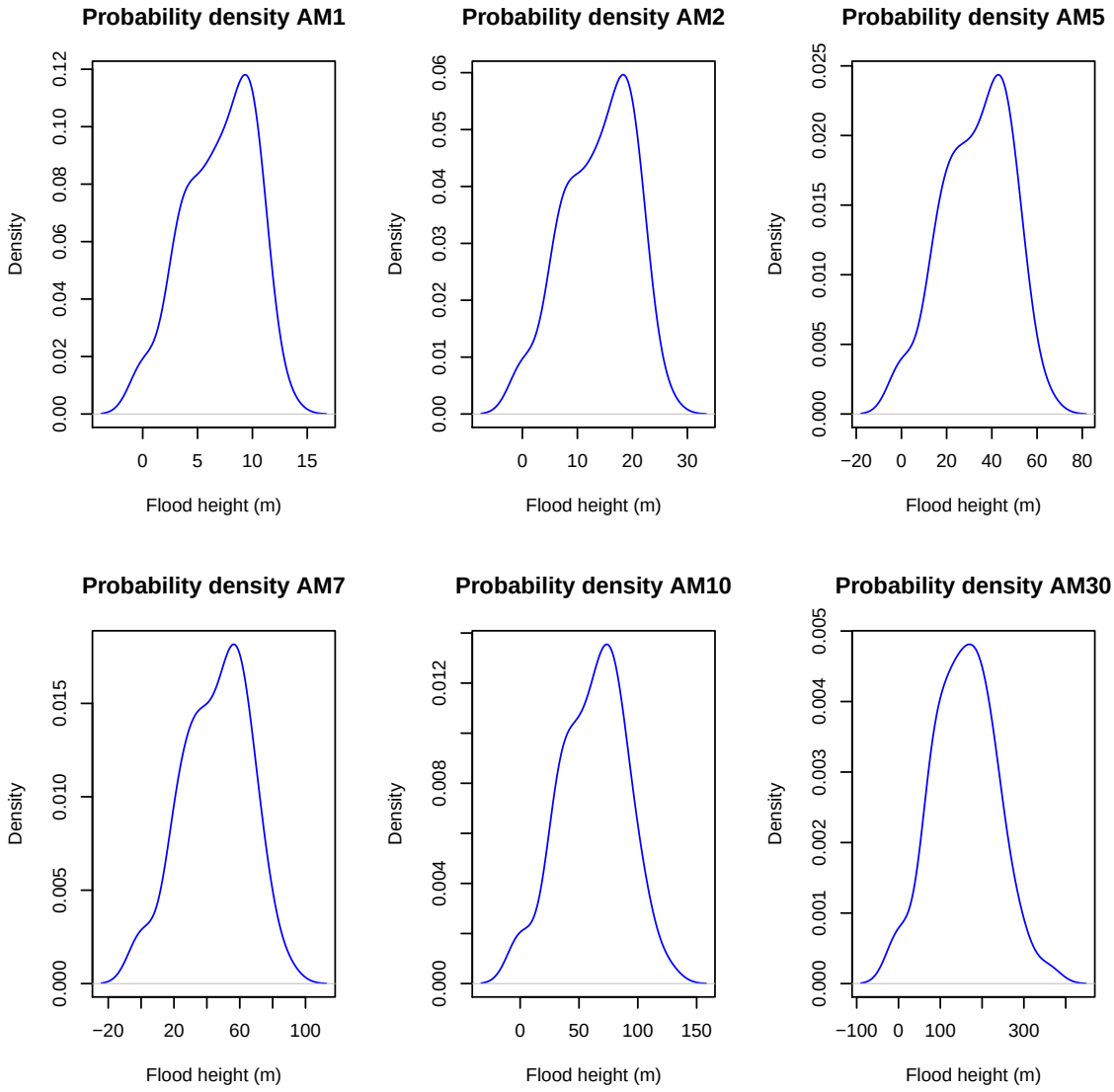


Figure 4.8: Comparison of probability density plots of the annual maxima time series models for the moving sums of Sicacate

ues (or negative skewness) is an indication that Sicacate, being downstream, has higher water levels more often in a given year compared to the other two sites in this study.

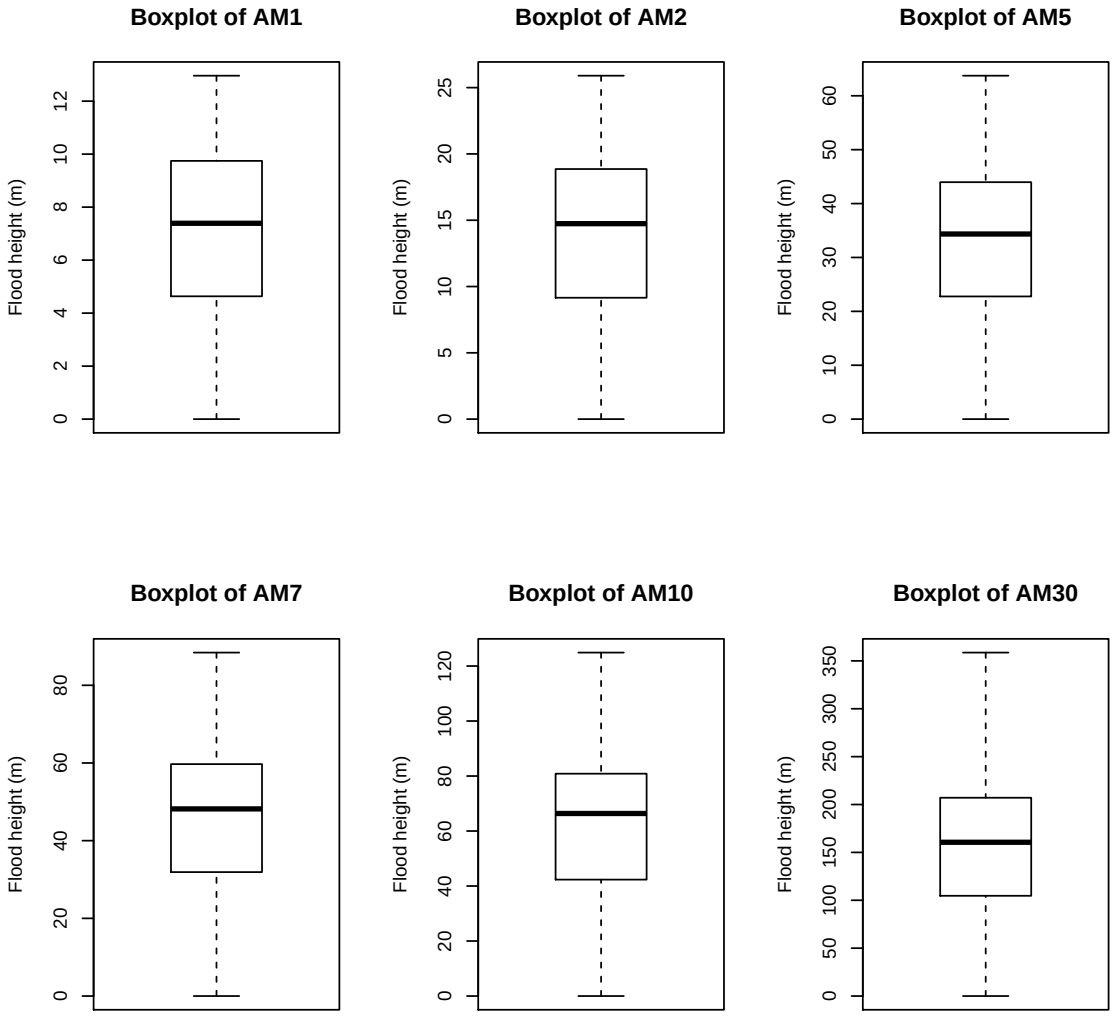


Figure 4.9: Comparison of boxplots of the annual maxima time series models for the moving sums of Sicacate

4.3.4 Analysis of variance results for the three sites

Tables 4.1, 4.2 and 4.3 present the results of the descriptive statistics for each annual maxima model and the ANOVA p-values for Chokwe, Combomune and Sicacate, respectively. The results in Table 4.1 showed an ANOVA p-value of $0.384\ (> 0.05)$, which is an indication of a lack of sufficient evidence to support the existence of significant differences among the six annual maxima time se-

ries models for Chokwe at a 5% level of significance. In other words, the six annual maxima time series models at Chokwe hydrometric station do not differ significantly with respect to skewness, excess kurtosis, CV, A-D and K-S statistics.

The results in Table 4.2 for Combomune showed that the six annual max-

Table 4.1: Summary descriptive statistics of the characteristics of the annual maxima moving sums for Chokwe

Statistic	**AM1**	**AM2**	**AM5**	**AM7**	**AM10**	**AM30**
A-D	0.0500	0.0500	0.0500	0.0600	0.0600	0.0600
K-S	0.3938	0.2300	0.1800	0.2100	0.2200	0.1800
CV	0.2400	0.3890	0.3870	0.3920	0.3970	0.4450
Skewness	0.4063	0.6880	0.5830	0.5970	0. 6030	0.6790
Excess K	1.9940	0.7580	-0.0680	-0.1960	-0.1950	0.2130

Key: The p-value $= 0.384\ (> 0.05)$ for the one-way ANOVA. A-D and K-S represent Anderson-Darling and Kolmogorov-Smirnov statistics, (respectively), based on the GEV distribution. CV is coefficient of variation and Excess K represents excess kurtosis.

ima time series models are not significantly different from each other (p-value $= 0.327 > 0.05$). These findings are consistent with those at the other two sites. This implies that for each particular site, the annual maxima time series models do not significantly differ with respect to skewness, excess kurtosis, CV, A-D and K-S tests.

Similarly, the results in Table 4.3 showed an ANOVA p-value of $0.958\ (> 0.05)$, which suggests a lack of sufficient evidence at 5% level of significance to support the existence of significant differences among the six annual maxima time series models at Sicacate hydrometric station. The findings at Sicacate, further downstream the LLR of Mozambique, are consistent with those at Chokwe and Combomune in that the six annual maxima time series models at each site exhibited characteristics that were not significantly different (p-value > 0.05) with regard to the dispersion measures of skewness, excess kurtosis and coeffi-

Table 4.2: Summary descriptive statistics of the characteristics of the annual maxima moving sums for Combomune

Statistic	AM1	AM2	AM5	AM7	AM10	AM30
A-D	0.2002	0.1852	0.1740	0.1942	0.2062	0.4358
K-S	0.0730	0.0666	0.0668	0.0731	0.0683	0.0842
CV	0.3181	0.3106	0.2957	0.2923	0.2859	0.2483
Skewness	0.5273	0.6890	0.7400	0.7300	0.7183	1.1006
Excess K	0.1084	0.2207	0.4544	0.4000	0.4500	2.0302

Key: The p-value $= 0.327\ (> 0.05)$ for the one-way ANOVA. A-D and K-S represent Anderson-Darling and Kolmogorov-Smirnov statistics, (respectively), based on the GEV distribution. CV is coefficient of variation and Excess K represents excess kurtosis.

cient of variation, and the goodness-of-fit of the GEV distribution with respect to Anderson-Darling and Kolmogorov-Smirnov statistics. These findings are also consistent with the findings of Chapter 3 that the GEV distribution fits very well at all the three sites in the basin.

Table 4.3: Summary descriptive statistics of the characteristics of the annual maxima moving sums for Sicacate

Statistic	AM1	AM2	AM5	AM7	AM10	AM30
A-D	0.0913	0.0937	0.0850	0.0574	0.0475	0.0552
K-S	0.3938	0.3697	0.2898	0.2440	0.2162	0.2202
CV	0.4504	0.4508	0.4491	0.4461	0.4459	0.4737
Skewness	-0.4374	-0.4271	-0.3653	-0.3251	-0.2548	0.0820
Excess K	-0.5832	-0.5700	-0.4640	-0.3364	-0.1958	0.0670

Key: The p-value $= 0.958\ (> 0.05)$ for the one-way ANOVA. A-D and K-S represent Anderson-Darling and Kolmogorov-Smirnov statistics, (respectively), based on the GEV distribution. CV is coefficient of variation and Excess K represents excess kurtosis.

4.3.5 Correlation coefficient results

Tables 4.4, 4.5 and 4.6 present results for the correlation matrices of Chokwe, Combomune and Sicacate hydrometric stations, respectively. The p-values ($<$

0.001) at all the three sites Chokwe, Combomune and Sicacate indicated that the correlations between all the models (variables) at the three sites are highly significant. The results in all Tables 4.4, 4.5 and 4.6 revealed strong positive correlations among the annual maxima time series models at the sites.

Table 4.4: Correlation matrix of the annual maxima moving sums for Chokwe

Statistic	**AM1**	**AM2**	**AM5**	**AM7**	**AM10**	**AM30**
AM1	1					
AM2	0.9967^{***}	1				
AM5	0.9731^{***}	0.9906^{***}	1			
AM7	0.9732^{***}	0.9821^{***}	0.9980^{***}	1		
AM10	0.9617^{***}	0.9713^{***}	0.9910^{***}	0.9959^{***}	1	
AM30	0.8958^{***}	0.9058^{***}	0.9288^{***}	0.9401^{***}	0.9584^{***}	1

Key: *** represents correlations that are highly significant (p-value < 0.001). The correlation coefficients in the table are quite high, exceeding 0.89.

The strong correlations among the annual maxima time series models suggest that one annual maximum time series model at each particular site can be used to represent (or in place of) the rest of the annual maxima time series models in FFA at that site. In these circumstances, it is common to use the annual daily maximum flood height because it is convenient and easy to obtain. The researcher warns against using the 30-day annual maximum (AM30) series since graphical techniques in this study revealed that the 30-day annual maxima series often depicts slightly different distribution shapes. In the event of data with outliers, the researcher would recommend the use of annual maxima time series moving sums of up to 5 days.

4.3.6 Empirical cumulative distribution functions

The empirical CDF for the three sites are presented in Figures 4.10, 4.11 and 4.12. In this case the assumption is that the empirical CDF, $F(x)$, is not known

Table 4.5: Correlation matrix of the annual maxima moving sums for Combomune

Statistic	AM1	AM2	AM5	AM7	AM10	AM30
AM1	1					
AM2	0.9879^{***}	1				
AM5	0.9452^{***}	0.9794^{***}	1			
AM7	0.9160^{***}	0.9586^{***}	0.9957^{***}	1		
AM10	0.8920^{***}	0.9404^{***}	0.9860^{***}	0.9961^{***}	1	
AM30	0.7485^{**}	0.8128^{**}	0.8817^{***}	0.9075^{***}	0.9296^{***}	1

Key: ** represents correlations that are significant for p-value < 0.01
*** represents correlations that are highly significant (p-value < 0.001). The correlation coefficients in the table are sufficiently high, exceeding 0.74

Table 4.6: Correlation matrix of the annual maxima moving sums for Sicacate

Statistic	AM1	AM2	AM5	AM7	AM10	AM30
AM1	1					
AM2	0.9997^{***}	1				
AM5	0.9925^{***}	0.9942^{***}	1			
AM7	0.9838^{***}	0.9862^{***}	0.9971^{***}	1		
AM10	0.9641^{***}	0.9672^{***}	0.9845^{***}	0.9938^{***}	1	
AM30	0.8727^{***}	0.8771^{***}	0.9051^{***}	0.9221^{***}	0.9521^{***}	1

Key: *** represents correlations that are highly significant (p-value < 0.001). The correlation coefficients in the table are quite high, exceeding 0.87.

but can be used to derive the return levels and their corresponding return periods. Figure 4.10 presents the empirical CDF for Chokwe. The non-exceedance probabilities given by the CDF graph can be used to derive the return periods for the corresponding return levels provided on the graph. Accordingly, the results of the return periods and their corresponding return levels (flood heights) can be presented in tabular form in a similar way to the tables in Chapter 3.

Figure 4.11 present the empirical CDF graph for Combomune. The unknown empirical distribution, $F(x)$, can be used to derive the distribution of return periods and their corresponding return levels for the Combomune site. These

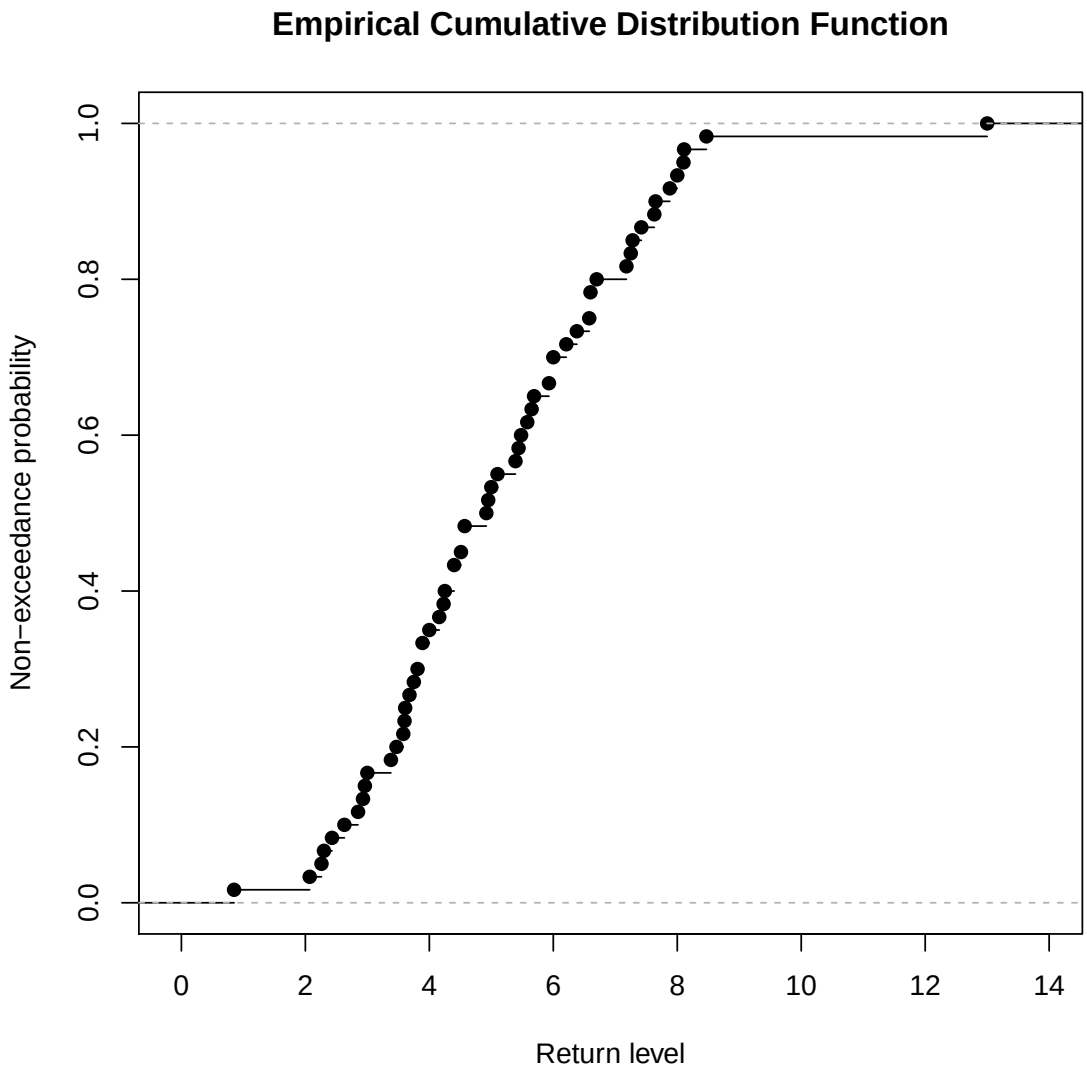


Figure 4.10: Empirical CDF of the AM1 model for Chokwe

return periods and their corresponding return levels are often presented in tabular form in the same way as those tables presented in Chapter 3. The return periods shown on the graph appeared to be consistent with those in Chapter 3 which were based on the best fitted distributions at the sites such as GEV and others.

Figure 4.12 presents the empirical CDF graph for Sicacate. The empirical CDF,

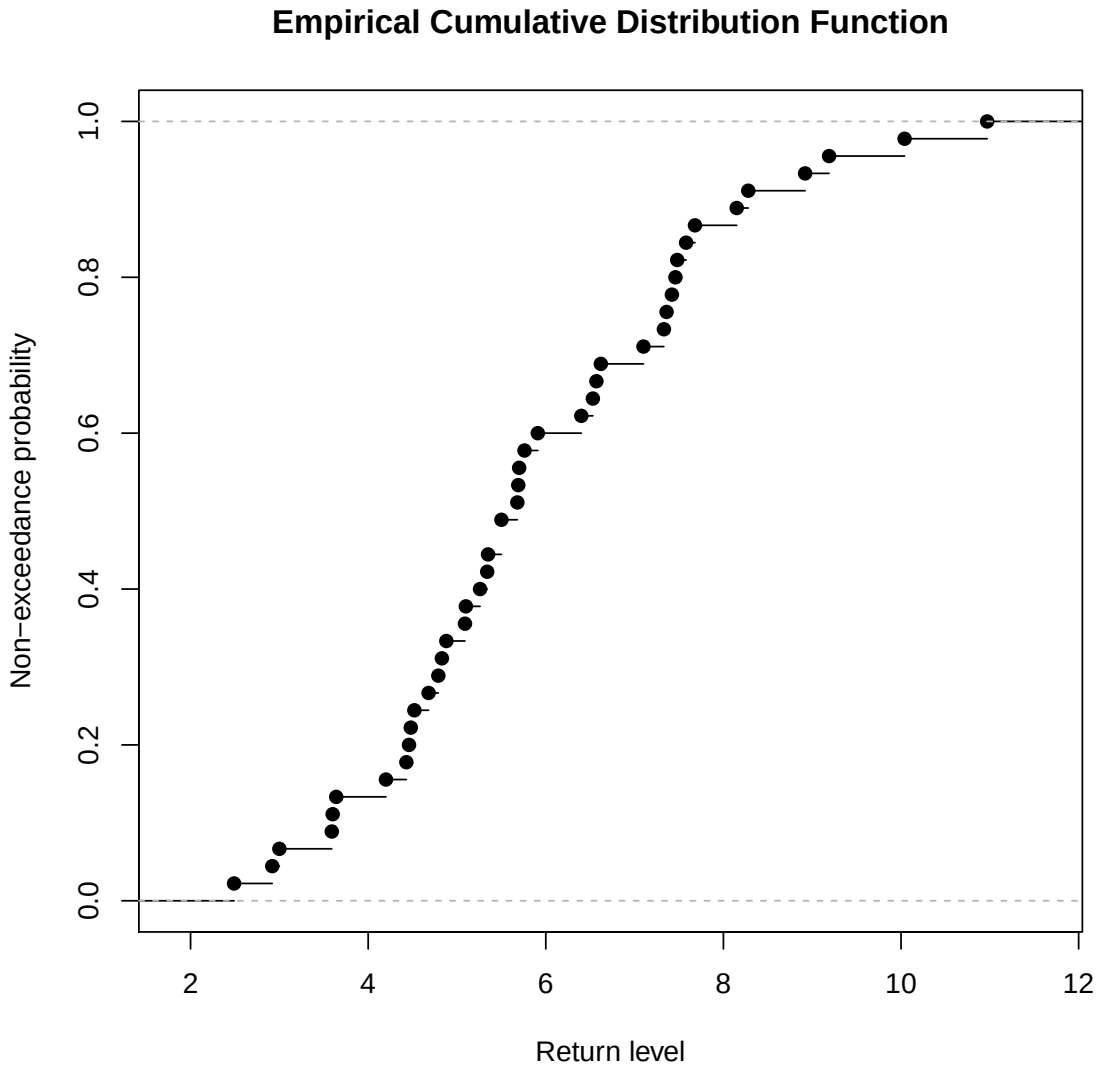


Figure 4.11: Empirical CDF of the AM1 model for Combomune

$F(x)$, is usually unknown but assumed to be in the domain of attraction of an extreme value distribution. It can easily be seen that the return levels and their corresponding return periods presented on the graph in Figure 4.12 are consistent with the results in Chapter 3 that were developed using the best fitted distributions at a particular site.

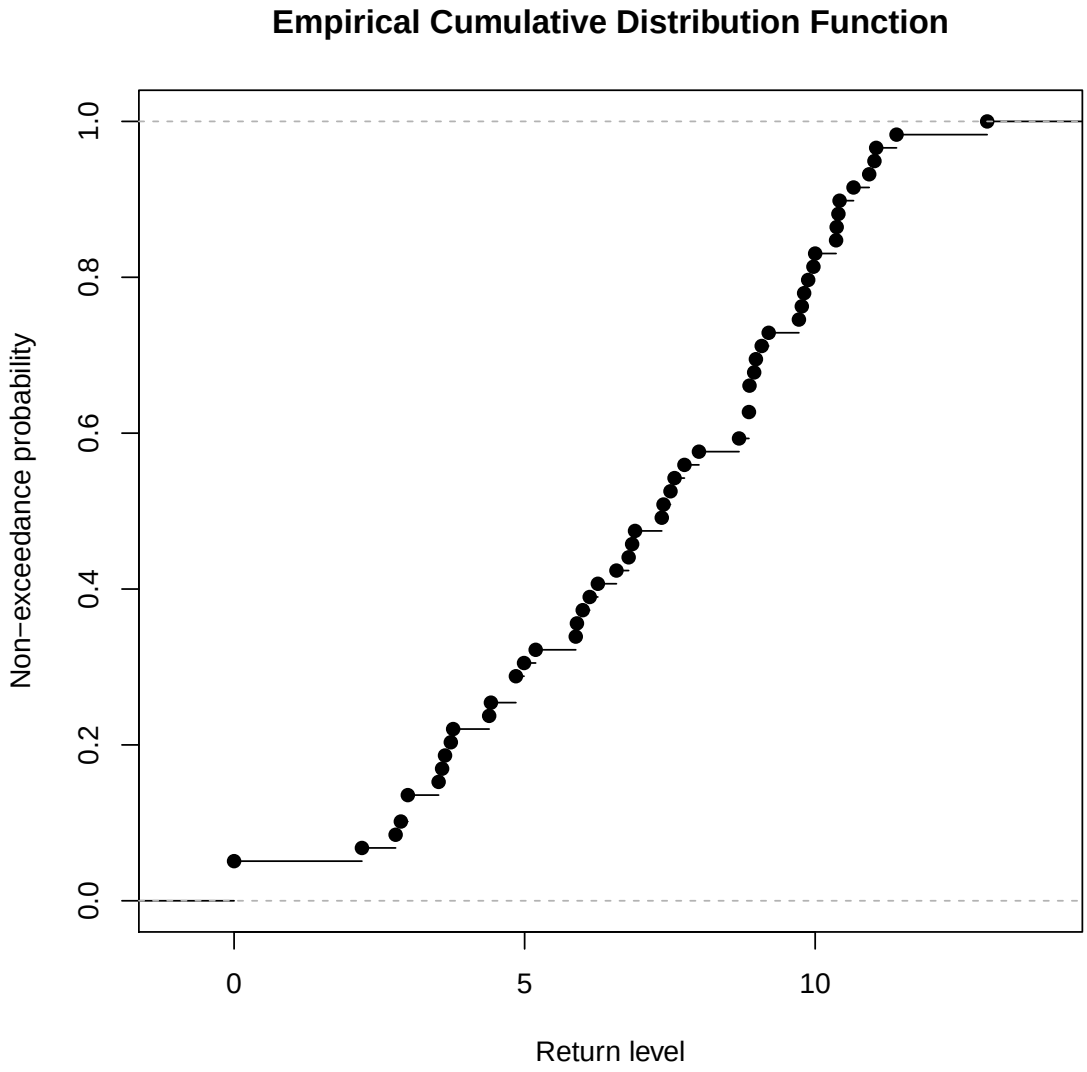


Figure 4.12: Empirical CDF of the AM1 model for Chokwe

4.4 Added value for disaster management and disaster risk reduction

The work in this chapter contributes to the decision-making process in disaster management and disaster risk reduction through addressing the gaps and challenges in reducing underlying risk factors. This work can also prepare vul-

nerable communities for effective response.

While most studies of flood frequency analyses use rainfall (precipitation) and river discharges (flood volumes) data series, this study used flood height data series. Quality data is not readily available for most countries, particularly in Southern Africa, where the gauging instruments may either be non-functional due to lack of service or simply unavailable due to budget constraints. This study has revealed that in a country where quality rainfall and river discharges data series records are scarce, flood heights can be used to make important decisions in disaster management and disaster risk reduction. In Chapter 3 it was shown that the findings deduced from flood frequency analyses of flood heights data series in Mozambique are closely comparable to those deduced using rainfall or river discharges data series in the neighbouring South Africa.

Knowledge of the characteristics of annual maximum time series moving sums is crucial in flood frequency analysis. Accurate identification of suitable annual maxima time series moving sums to use in flood frequency analysis helps improve the accuracy of the forecasts of these rare events. For instance, accurate forecasts in the return levels and their corresponding return periods would make great improvements in the design of engineering and hydraulic structures and consequently saving the amount of aid money spent on disaster relief operations.

Mozambique is a developing, flood-prone, and economically challenged country situated in Southern Africa and therefore natural disaster-related studies in this country will indeed help in substantial reduction of flood events-related losses in human lives and damage to infrastructure such as bridges and other mega constructions in the country.

Improved knowledge of the estimates (forecasts) of the return periods of rare events such as extremely high flood heights will help make the vulnerable communities better prepared and therefore respond effectively to the flood-related disasters.

This Chapter has demonstrated the worth and significance of forecasting, which can be applied everywhere in the world using the data available at a particular site. In summary, community preparedness and effective response to natural disasters such as floods will help reduce the associated risks and mitigate the undesirable impacts of these disasters on humans and property. Consequently, this will contribute to a substantial reduction in the amount of aid money spent on post-disaster relief operations.

4.5 Concluding remarks

In this chapter, the block maxima approach of EVT was considered for the at-site data series. The study was conducted at three sites; Chokwe, Combomune and Sicacate hydrometric stations in the lower Limpopo River basin of Mozambique. At each site, six annual maximum time series models were considered and compared with respect to skewness, excess kurtosis, coefficient of variation, and the GoF of the GEV distribution as measured by Anderson-Darling and Kolmogorov-Smirnov test statistics. The six annual maxima time series models considered at each site were AM1, AM2, AM5, AM7, AM10, and AM30; that is, annual daily, two-day, five-day, seven-day, ten-day, and thirty-day maximum flood heights, respectively.

The findings in this study revealed no sufficient evidence of significant differences among the six annual maxima time series models at all the three sites with respect to skewness, excess kurtosis, coefficient of variation, and GoF of

the GEV distribution assessed by Anderson-Darling and Kolmogorov-Smirnov statistics. Moreover, the study revealed overwhelming evidence of very strong positive correlations among the six annual maximum time series models at all the three sites in the basin.

The study also revealed notable differences between the three sites in overall skewness of the annual maxima time series models, with Chokwe and Combomune (upstream) dominated by positive skewness and Sicacate (downstream) dominated by negative skewness. This is reasonable in the real world since it is expected that there is usually more water downstream than upstream and consequently the flood heights become higher downstream than upstream.

Based on the findings in this study, it can be concluded that using the annual daily maximum flood heights model or any of the other five annual maxima time series models in flood frequency analysis has no significant effect on the forecasts of extreme return levels and their corresponding return periods. Without loss of generality, the researcher recommends the use of the annual daily maximum flood heights model in flood frequency analysis to construct the flood frequency curves mainly due to its simplicity and the relative ease of obtaining it.

4.6 Summary of the chapter

This chapter explored a comparative analysis of six annual maximum (AM) flood heights time series models at three sites: Chokwe, Combomune and Sicacate in the lower Limpopo River basin of Mozambique. The six AM time series models considered were the annual daily maximum (AM1), annual two-day maximum (AM2), annual five-day maximum (AM5), annual seven-day maximum (AM7), annual ten-day maximum (AM10), and annual thirty-day maxi-

mum (AM30). A GEV distribution was fitted to each of the six AM time series models. The goodness-of-fit of the GEV distribution at each site was assessed using Anderson-Darling (A-D) and Kolmogorov-Smirnov (K-S) statistics. An analysis of variance was performed to check for significant differences between the six AM models at each site with reference to skewness, coefficient of variation , excess kurtosis, A-D statistics, and K-S statistics. The results revealed no evidence of significant differences at the 5% level of significance among the six AM time series models in terms of skewness, CV, excess kurtosis, and A-D and K-S statistics. A correlation analysis was also performed to check for significant correlations among the time series models: the results revealed, in general, high correlations among all six time series models at each site. These findings suggest that any one of the models at each site can be used in place of the other five annual maximum time series models in flood frequency analyses of the lower Limpopo River. Without losing generality, the annual daily maximum, AM1, flood height time series model will be used for further analysis to obtain flood frequency curves in the succeeding chapters mainly because of its simplicity, relative ease of use and interpretation.

Chapter 5

Estimating high quantiles of extreme flood heights in the lower Limpopo River basin of Mozambique using model based Bayesian approach

5.1 Introduction

The increased frequency and intensity of floods is yet to be fully explained in modern day science and statistics, particularly in Mozambique where the use of EVT is still limited. The use of Bayesian parameter estimates of the GEV model in explaining the increased frequency of floods in the LLRB of Mozambique is discussed in this chapter. Bayesian analysis of extreme flood heights

has a provision of taking into account the uncertainty encountered in parameter estimation. This uncertainty is usually incorporated into the Bayesian analysis through the incorporation of either informative or non-informative prior distributions. It is argued in literature that Bayesian modelling approach is debateably more informative than the frequentist statistical modelling approach (Bayarri and Berger, 2004).

5.2 Brief background and review of related literature

Mozambique has nine transboundary rivers from neighbouring countries such as Malawi, South Africa, Tanzania and Zimbabwe. Among the transboundary rivers in Mozambique, Zambezi River is the largest river in the territory followed by Limpopo River which is the second largest African river that drains into the Indian Ocean. Unlike the Zambezi River which is characterised by very large dams such as Kariba and Cohora Bassa, the Limpopo River has no large dams implying that the flow is not highly regulated. The hydrology of the LRB is characterised by one cycle of rainfall that extends from October of the previous year to April with peak monthly totals in February, while the dry season runs from May to September (Spaliveiro et al., 2014; WMO, 2012). The Limpopo River is well pronounced by extreme natural hazards; alternating between extreme floods and severe droughts (WMO, 2012).

Gohil and Chowdhary (2013) defined flood as a rare high event of a river usually as a result of extremely high rainfall triggered by unusual meteorological conditions. Several authors have provided results that portend that the frequency, magnitude and intensity of extreme weather events such as floods and temperature are on the rise (Yilmaz et al., 2014; MunichRe, 2013; WMO, 2013). For instance, WMO (2013) showed that floods were the most frequently ex-

perienced extreme events over the course of the decade 2001-2010 worldwide including Africa, while MunichRe (2013) affirmed that natural catastrophic statistics for the year 2013 was dominated by floods that caused billions of American dollars in losses.

Mujere (2011) argued that despite the accurate short-term flood forecasts provided by meteorological forecasts, the shortage of time allowed for disaster preparedness and the incidence of false alarms lead people not to take the short-term forecasts seriously. These shortcomings in flood forecasting justify the need to use statistical methods which provide long-term forecasts. Mabaso and Manyena (2013) advocated for contingency planning in Southern Africa to be considered as an event rather than a process in disaster preparedness and response planning in an effort to reduce disaster risk. Most recently Spaliveiro et al. (2014) gave a detailed account of flood risk analysis in the LRB from a geoscientific point of view based on the river's past evolution and geomorphological characteristics.

Mondlane et al. (2013) performed a comparative analysis of extreme flood frequency distribution models based on 20 years of rainfall data recorded at Xai-Xai precipitation station in LLRB of Mozambique using the Gumbel, Fréchet, Pareto and Weibull distributions. The histogram of the collected data in Mondlane et al. (2013) showed a multi-modal distribution, and the Gumbel Max distribution appeared to approximate its skewness better compared to other distributions in the paper while the two-parameter Weibull and Gumbel Min fitted the negatively skewed and unimodal distribution of the randomly simulated data.

Smithers (2012) detailed a comprehensive literature of advances that have been made to model extreme floods. Nevertheless, Smithers (2012) stressed

that the demand for reliable and improved estimates of flood frequency in terms of flood peaks and return periods have not been met and still poses a challenge in hydrology despite the improved understanding of the fundamental hydrological processes.

The purpose of the study for this chapter is to perform a comparative analysis of maximum likelihood and Bayesian estimates of the GEV distribution. The researcher uses Bayesian Markov Chain Monte Carlo (MCMC) inference, which has the advantage of not requiring to satisfy the regularity conditions, to estimate the parameters of a GEV distribution and further make predictions of the return levels and their corresponding return periods. These estimates and predictions are compared with those from frequentist approach based on maximum likelihood estimates of the GEV distribution in a block maxima framework. Recently Ferreira and de Haan (2015) demonstrated conditions under which the block maxima method may prevail over the POT method and Dombry (2015) showed the existence and consistency of MLEs in a block maxima framework.

Gaioni et al. (2010) alluded that the GEV distribution arises naturally when modelling the maxima over a sequence of observations. These authors proposed a model based Bayesian approach for direct quantile elicitation which translates into prior distribution assessment. They argued that although the proposed approach was quite general, it was deemed particularly useful in river data cases in which direct assessments on the prior distribution are extremely difficult. Most recently Vidal (2014) emphasised the novelty of model based Bayesian inference approach when he used Bayesian estimates of the Gumbel distribution to analyse extreme rainfall data in Chile. Vidal (2014) left the use of GEV in Bayesian analysis to further research. This provides some evidence that the application of this method in hydrology is still relatively new and thus

supports its application in least developed countries such as Mozambique. In this study we propose to use a Bayesian MCMC approach with the GEV distribution as the likelihood function in order to use the prior to develop the predictive distribution which provides the basis for future expectations regarding the behavior of the lower Limpopo River. The advantages of using Bayesian models are discussed in Gaioni et al. (2010), and for further reading on Bayesian inference models the reader is referred to Reiss and Thomas (2007); Reis and Stedinger (2005) and Beirlant et al. (2004). Some of the advantages of using Bayesian parameter estimation methods are the use of prior knowledge or information and that the modeler is able to capture uncertainty of the parameter estimates.

The rest of the chapter is such that Section 5.2 provides a brief overview of both the frequentist and Bayesian approaches, the data, the types of priors used in the analysis, while Section 5.3 presents and discusses the results. Section 5.4 outlines the value added by the chapter, with reference in particular, to disaster risk reduction. Section 5.5 gives the general remarks on the results, while Section 5.6 concludes the chapter and finally Section 5.7 provides a summary of the chapter. Additionally, Appendix 5.1 provides the flood frequency curves for the three sites, while Appendix 5.2 presents R program code for the Bayesian MCMC used to produce the results for the Chapter.

5.3 Research methodology

This section presents the methods used in the study for the analysis of the data in this chapter. The methods range from algorithms, prior distribution methods, probability of framework of block maxima in the Bayesian and frequentist paradigms.

5.3.1 The data

Like in the preceding chapters, the data used in this chapter was block maxima data for the three sites; Chokwe, Combomune and Sicacate hydrometric stations which are in the LLRB of Mozambique.

5.3.2 The frequentist flood frequency analysis probability framework

The probability framework of the frequentist flood frequency analysis based on the block maxima approach has been discussed in the previous chapters, in particular Chapter 2, Subsection 2.7.1.

The log-likelihood function for the GEV parameters in (2.6), for $\xi \neq 0$, is given by

$$\begin{aligned} L_k = \ell(\mu, \sigma, \xi) &= -k \ln \sigma - \left(\frac{1}{\xi} + 1\right) \sum_{i=1}^{k} \ln \left[1 + \xi \left(\frac{x_i - \mu}{\sigma}\right)\right]_+ \\ &\quad - \sum_{i=1}^{k} \left[1 + \xi \left(\frac{x_i - \mu}{\sigma}\right)\right]_+^{\frac{-1}{\xi}}, \end{aligned} \tag{5.1}$$

The log-likelihood for case $\xi = 0$, using the Gumbel limit of the GEV (Coles, 2001), is

$$L_k = \ell(\mu, \sigma) = -k \ln \sigma - \sum_{i=1}^{k} \left(\frac{x_i - \mu}{\sigma}\right) - \sum_{i=1}^{k} \exp\left\{-\left(\frac{x_i - \mu}{\sigma}\right)\right\} \tag{5.2}$$

where k is the number of years (blocks) and x_i is the flood height.

The MLEs of the parameters μ, σ and ξ are obtained from the solution of the likelihood equations obtained from the partial derivatives of the log-likelihood

L_k,

$$\nabla L_k = 0 \quad \text{with } \nabla L_k = \left(\frac{\partial L_k}{\partial \mu}, \frac{\partial L_k}{\partial \sigma}, \frac{\partial L_k}{\partial \xi} \right). \tag{5.3}$$

The MLE is any solution of (5.3) with a negative definite Hessian matrix (Dombry, 2015). Most recent literature on the existence of consistent MLEs in block maxima is found in Dombry (2015). More details on the regularity conditions of the MLEs are found in literature (Rajaram, 2006; Beirlant et al., 2004; Coles, 2001; Smith, 1987, with references therein).

The quantile estimates of the GEV are obtained from the quantile function, X_p

$$X_p = G^{-1}(1-p) \;\; = \;\; \begin{cases} \hat{\mu} + \frac{\hat{\sigma}}{\hat{\xi}} \left[(-\ln(1-p))^{-\hat{\xi}} - 1 \right], & \hat{\xi} \neq 0, \\ \hat{\mu} - \hat{\sigma} \ln(-\ln(1-p)), & \hat{\xi} = 0, \end{cases} \tag{5.4}$$

where p is the exceedance probability and $T = \frac{1}{p}$ is the return period of an extreme flood height, X_p. The extreme quantiles in this chapter will be estimated using both the MLE and Bayesian parameter estimates.

5.3.3 Bayesian MCMC flood frequency modelling framework

The subsection presents a theoretical overview of the Bayesian MCMC flood frequency modelling. Suppose the observation vector $\boldsymbol{x} = \{x_1, ..., x_k\}$ consists of iid realisations of annual maximum flood heights and parameter vector $\boldsymbol{\theta} = \{\mu, \sigma, \xi\}$. The posterior distribution is computed using Bayes' Theorem

$$\pi(\boldsymbol{\theta}|\boldsymbol{x}) \;\; = \;\; \frac{\pi(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int_{\boldsymbol{\Theta}} \pi(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) d\boldsymbol{\theta}} \tag{5.5}$$

which is usually written as

$$\pi(\boldsymbol{\theta}|\boldsymbol{x}) \;\; \propto \;\; \pi(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) \tag{5.6}$$

where $\boldsymbol{x}$ is a vector of observations, $\boldsymbol{\theta}$ is a parameter vector, $\pi(\boldsymbol{\theta})$ is the prior distribution, $\pi(\boldsymbol{\theta}|\boldsymbol{x})$ is the posterior distribution, $\pi(\boldsymbol{x}|\boldsymbol{\theta})$ is the likelihood function, $\pi(\boldsymbol{x})$ is the normalisation constant and $\boldsymbol{\Theta}$ is the space parameter.

The $100(1-\alpha)\%$ Bayesian credible set C (or in particular credible interval) is a subset of the space parameter Θ such that

$$\int_C \pi(\boldsymbol{\theta}|\boldsymbol{x})d\boldsymbol{\theta} = 1-\alpha, \tag{5.7}$$

where the sum replaces the integral if the space parameter Θ is discrete. The quantile-based credible intervals are such that if $\boldsymbol{\theta}_L^*$ is the $\alpha/2$ posterior quantile for $\boldsymbol{\theta}$, and $\boldsymbol{\theta}_U^*$ is the $1-\alpha/2$ posterior quantile for $\boldsymbol{\theta}$, then $(\boldsymbol{\theta}_L^*, \boldsymbol{\theta}_U^*)$ is the $100(1-\alpha)\%$ credible interval for $\boldsymbol{\theta}$.

Markov chains methods

Numerous Markov chains are available in literature for sampling from a posterior distribution. The two main examples of Markov chain algorithms are the Gibbs sampler and Metropolis-Hastings algorithms (Tierney, 1994). The Gibbs sampler algorithm is briefly discussed in this section. Details on the Metropolis-Hastings algorithm and other Monte Carlo methods are found in Tierney (1994).

The Gibbs sampler introduced by Geman and Geman (1984) is a special case of the Metropolis-Hastings algorithm in Hastings (1970). The Gibbs algorithm is as follows: Suppose $\boldsymbol{\theta} = (\theta_1, ..., \theta_p)$. Given all the remaining components of $\boldsymbol{\theta}$, every $\theta_j^{(i)}$ is sampled from the conditional distribution. Suppose there exists a

set of values $\left\{\theta_1^{(i)}, \theta_2^{(i)}, ..., \theta_p^{(i)}\right\}$ then the algorithm follows:

$$\begin{aligned}
\text{Select } \theta_1^{(i+1)} &\sim \pi(\theta_1|\theta_2^{(i)}, \theta_3^{(i)}, ..., \theta_p^{(i)}, x) \\
\text{Select } \theta_2^{(i+1)} &\sim \pi(\theta_2|\theta_1^{(i+1)}, \theta_3^{(i)}, ..., \theta_p^{(i)}, x) \\
&\vdots \\
\text{Select } \theta_p^{(i+1)} &\sim \pi(\theta_p|\theta_1^{(i+1)}, ..., \theta_{p-1}^{(i+1)}, x).
\end{aligned}$$

5.3.4 Prior distributions

In Bayesian analysis framework, prior distribution refers to the distribution that is based on historical information or expert knowledge about a particular extreme event such as extreme floods and extreme temperatures. Prior distributions are generally classified into informative and noninformative priors.

Noninformative priors, also known as objective priors are sometimes referred to as vague or flat priors (Sigauke, 2014; Gelman et al., 2004). The name flat prior is derived from the fact that the likelihood on all the possible values of a parameter is flat or equal (Gelman et al., 2004). Details on the development of noninformative priors, the rules and techniques for deriving these noninformative priors are provided in Kass et al. (1996).

Informative priors (or nonobjective priors) are used when prior or historical information concerning the parameter is available (Koch, 2007). Prior information generally comes in the form of previous studies, previous knowledge or expert opinions and this is usually difficult in practice to incorporate in the Bayesian analysis. Several techniques are proposed in literature to overcome this challenge (Reis and Stedinger, 2005; Beirlant et al., 2004; Ibrahim and

Chen, 2000). Among the types of informative priors, conjugate priors are a special type of informative priors. According to Gelman et al. (2004) conjugate priors are very special in that they provide computational convenience in such a way that the posterior and prior distributions belong to the same distribution family.

Two examples of noninformative priors are the Jeffreys' prior and the maximal data information (MDI) prior (Beirlant et al., 2004). The MDI prior provides maximal average data information and the prior is not invariant under reparameterisation. Constraints on the parameter are built into the prior (Sigauke, 2014, with references therein). The Jeffreys' prior is briefly discussed as follows, although both will not be used for this thesis. Instead the trivariate normal prior, which is conjugate prior, is used in this chapter.

Jeffreys' prior

The Jeffreys' prior is a noninformative or objective prior and it is invariant under reparameterisation. Nevertheless, Jeffreys' prior violates the likelihood principle (Beirlant et al., 2004, with references therein). According to Beirlant et al. (2004), the Jeffreys' prior is defined as

$$\pi(\boldsymbol{\theta}) \quad \propto \quad \sqrt{|I(\boldsymbol{\theta})|} \tag{5.8}$$

where $\boldsymbol{\theta} = (\theta_1, ..., \theta_p)$ and $I(\boldsymbol{\theta})$ is Fisher's information matrix with $(i, j)^{th}$ element.

$$I_{i,j}(\boldsymbol{\theta}) \quad = \quad E\left\{-\frac{\partial^2 \log f(\boldsymbol{X}|\boldsymbol{\theta})}{\partial\theta_i \partial\theta_j}\right\}, i, j = 1, ..., p \tag{5.9}$$

Conjugate (Trivariate normal) prior distribution

The likelihood function used in this chapter is the GEV distribution. According to Beirlant et al. (2004), Jeffreys' prior is complicated and it only exists for $\xi > -0.5$. Therefore, the researcher uses the trivariate normal (conjugate) prior which is discussed in Stephenson and Ribatet (2006), and Coles and Powell (1996). Suppose the prior distribution is denoted by $\pi(\boldsymbol{\theta})$, where $\boldsymbol{\theta} = \{\mu, \sigma, \xi\}$. Now suppose

$$\boldsymbol{\theta}' = \{\mu, \ln \sigma, \xi\}$$

then according to Stephenson and Ribatet (2006) the prior distribution on $\boldsymbol{\theta}$ is defined as follows:

$$\pi(\boldsymbol{\theta}') \propto \frac{1}{\sigma} \exp\left\{-\frac{1}{2}\left(\boldsymbol{\theta}' - \boldsymbol{\vartheta}\right)^T \Sigma^{-1} \left(\boldsymbol{\theta}' - \boldsymbol{\vartheta}\right)\right\} \tag{5.10}$$

where $\boldsymbol{\vartheta}$ is the mean vector and Σ is the symmetric (3×3) positive definite covariance matrix (Stephenson and Ribatet, 2006).

The likelihood function

The likelihood function is

$$\pi(x|\boldsymbol{\theta}) = \prod_{i=1}^{k} \frac{1}{\sigma}\left[1 + \xi\left(\frac{x_i - \mu}{\sigma}\right)\right]^{-\frac{1}{\xi}-1} \exp\left\{-\left[1 + \xi\left(\frac{x_i - \mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\} \tag{5.11}$$

The joint posterior distribution

The joint posterior density is

$$\begin{aligned}
\pi(\boldsymbol{\theta}|x) &\propto \pi(\boldsymbol{\theta})\pi(x|\boldsymbol{\theta}) \\
\pi(\boldsymbol{\theta}|x) &\propto \frac{1}{\sigma}\exp\left\{-\frac{1}{2}\left(\boldsymbol{\theta}'-\boldsymbol{\vartheta}\right)^T\Sigma^{-1}\left(\boldsymbol{\theta}'-\boldsymbol{\vartheta}\right)\right\} \\
&\times \prod_{i=1}^{k}\frac{1}{\sigma}\left[1+\xi\left(\frac{x_i-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}-1}\exp\left\{-\left[1+\xi\left(\frac{x_i-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\} \\
\pi(\boldsymbol{\theta}|x) &\propto \frac{1}{\sigma^{k+1}}\exp\left\{-\frac{1}{2}\left(\boldsymbol{\theta}'-\boldsymbol{\vartheta}\right)^T\Sigma^{-1}\left(\boldsymbol{\theta}'-\boldsymbol{\vartheta}\right)\right\} \\
&\times \exp\left\{-\Sigma_{i=1}^{k}\left[1+\xi\left(\frac{x_i-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\}
\end{aligned}$$

$$\begin{aligned}
&\times \prod_{i=1}^{k}\frac{1}{\sigma}\left[1+\xi\left(\frac{x_i-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}-1} \\
\pi(\boldsymbol{\theta}|x) &\propto \frac{1}{\sigma^{k+1}}\exp\left\{-\frac{1}{2}\left(\boldsymbol{\theta}'-\boldsymbol{\vartheta}\right)^T\Sigma^{-1}\left(\boldsymbol{\theta}'-\boldsymbol{\vartheta}\right)-\sum_{i=1}^{k}\left[1+\xi\left(\frac{x_i-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\} \\
&\times \prod_{i=1}^{k}\left[1+\xi\left(\frac{x_i-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}-1} \qquad (5.12)
\end{aligned}$$

Suppose that $\boldsymbol{\theta_0} = \{\mu_0, \sigma_0, \xi_0\}$ and $s = \{s_\mu, s_\sigma, s_\xi\}$ denote the initial values of the MCMC. The researcher uses the maximum likelihood estimates for these initial values. Suppose the chain start at state $\theta_t = \{\mu_t, \sigma_t, \xi_t\}$, then the subsequent states θ_{t+1} are generated as illustrated in the following procedure

(Stephenson and Ribatet, 2006).

If the log-normal distribution is denoted by LN then the algorithm that generates θ_{t+1} follows: Suppose $\mu^* \sim N(\mu_t, s_\mu^2)$, then let

$$\Delta = \frac{\pi(\mu^*, \sigma_t, \xi_t|\boldsymbol{x})}{\pi(\mu_t, \sigma_t, \xi_t|\boldsymbol{x})}. \tag{5.13}$$

Let $\mu_{t+1} = \mu^*$ with probability min $\{1, \Delta\}$, else set $\mu_{t+1} = \mu_t$.
Suppose $\sigma^* \sim LN(\ln \sigma_t, s_\sigma^2)$, then let

$$\Delta = \frac{\pi(\mu_{t+1}, \sigma^*, \xi_t|\boldsymbol{x})}{\pi(\mu_{t+1}, \sigma_t, \xi_t|\boldsymbol{x})} \frac{\sigma^*}{\xi_t}. \tag{5.14}$$

Let $\sigma_{t+1} = \sigma^*$ with probability min $\{1, \Delta\}$, else set $\sigma_{t+1} = \sigma_t$.

Let $\xi^* \sim N(\xi_t, s_\xi^2)$, then set

$$\Delta = \frac{\pi(\mu_{t+1}, \sigma_{t+1}, \xi^*|\boldsymbol{x})}{\pi(\mu_{t+1}, \sigma_{t+1}, \xi_t|\boldsymbol{x})}. \tag{5.15}$$

Let $\xi_{t+1} = \xi^*$ with probability min $\{1, \Delta\}$, else set $\xi_{t+1} = \xi_t$.

The posterior predictive density

The posterior predictive density can be used to predict the future posterior predictive tail probabilities of a future observation, X_0 as follows:

$$\begin{aligned} P(X_0 > x_0|x_1, ..., x_k) &= \int_{\boldsymbol{\theta}} P(X_0 > x_0|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|x)d\boldsymbol{\theta} \\ P(X_0 > x_0|x_1, ..., x_k) &\approx \frac{1}{k-a+1}\sum_{i=a}^{k} P(X_0 > x_0|\boldsymbol{\theta_i}) \end{aligned} \tag{5.16}$$

where $P(X_0 > x_0)$ is the GEV distribution evaluated at x_0 and a is the burn-in period (Stephenson and Ribatet, 2006).

Now suppose X_N is the annual daily maximum flood height over some future period of N years, then according to Stephenson and Ribatet (2006) the posterior predictive distribution is given by

$$P(X_N > x_0|x_1, ..., x_k) \approx \frac{1}{k-a+1}\sum_{i=a}^{k} P(X_0 > x_0|\boldsymbol{\theta_i})^N. \quad (5.17)$$

The following section presents the results and discussion of the results.

5.4 Results and discussion

This section presents results for both the Bayesian and frequentist realisations. These results are presented in the form of tables and figures. Tables 5.1-5.6 are presented in the main chapter, while Figures 5.1-5.3 are presented at the end of the chapter.

5.4.1 Chokwe site

Table 5.1: Parameter estimates for Chokwe

	ML estimates		
Parameter	**Estimate**	**SE**	**95% CI**
μ	4.264	0.253	$(3.757, 4.772)$
σ	1.789	0.177	$(1.434, 2.144)$
ξ	-0.084	0.073	$(-0.229, 0.062)$
	Bayesian estimates		
Parameter	**Estimate**	**Naive SE**	**95% CI**
μ	4.272	0.006	$(3.764, 4.787)$
σ	1.901	0.005	$(1.557, 2.400)$
ξ	-0.068	0.002	$(-0.203, 0.105)$

Table 5.1 presents results for the parameter estimates of both the maximum likelihood (ML) estimates and Bayesian MCMC estimates for Chokwe. Results

in Table 5.1 showed that the ML estimates of the parameters are generally lower than the Bayesian estimates. Moreover, the Bayesian credible intervals are generally wider than their corresponding confidence intervals for the ML approach. Note that the naive standard errors for the Bayesian estimates are given by dividing the actual standard deviation by the number of iterations (Stephenson and Ribatet, 2006). This is due to the fact that the Bayesian approach allows for an additional source of variation, which implies that the parameters now have probability distributions with hyperparameters. Both the 95% confidence intervals for the ML estimates and 95% credible intervals for the Bayesian estimates revealed that the shape parameter of the GEV distribution at Chokwe is not significantly different from zero, indicating that annual maximum flood heights at the site can be modelled by a light-tailed Gumbel family of distributions. The parameter estimates presented in Table 5.1 were used to produce Table 5.2 using the quantile function in (5.5).

Table 5.2: Tail quantile estimation and prediction of extreme flood heights for Chokwe

1-p	**p**	**T**	**ML estimate** (Exceedances)*	**Bayesian estimate** (Exceedances)*
95^{th}	0.05	20 years	8.9 m (1)	9.38 m (1)
98^{th}	0.02	50 years	10.22 m (1)	10.79 m (1)
99^{th}	0.01	100 years	11.10 m (1)	11.78 m (1)
99.5^{th}	0.005	200 years	11.92 m (1)	12.72 m (1)
99.6^{th}	0.004	250 years	12.18 m (1)	13.02 m (0)
99.8^{th}	0.002	500 years	12.94 m (1)	13.90 m (0)
99.9^{th}	0.001	1,000 years	13.65 m (0)	14.74 m (0)
99.99^{th}	0.0001	10,000 years	15.76 m (0)	17.27 m (0)

Key: $1-p$ represents non-exceedance probability.
p represents exceedance Probability.
T represents return period.
(Exceedance)* represents the number of empirical observations above the flood level at the site.

Table 5.2 presents results for the tail quantile estimates and predicted val-

ues of extreme flood heights at Chokwe. The results in Table 5.2 showed that Bayesian estimates of the extreme flood heights are generally higher than their corresponding ML estimates, which is consistent with results from Table 5.1. The empirically observed maximum flood heights were evaluated over the predicted flood heights and it was found that only the 13 m flood height which occurred during the year 2000 disastrous floods is higher than the 200-year flood level at Chokwe based on both the Bayesian and frequentist approaches. It was further found that based on the ML estimates of the flood heights the 13 m flood height has a return period in excess of 500 years, while the Bayesian approach revealed that the same flood level has a return period of about 250 years. The results for the ML approach are also consistent with the findings in Chapter 3 based on the GEV L-moments estimates.

Figure 5.1 in Appendix 5.1 presents the return level plot of the posterior distribution with 95% credible intervals that are shown by dashed lines. In other words, the graph represents the flood frequency curves for Chokwe based on the Bayesian MCMC approach. The flood frequency curves can be used to obtain any flood heights of choice together with their corresponding return periods.

5.4.2 Combomune site

Table 5.3 presents results for the parameter estimates of both the ML and Bayesian MCMC estimates for Combomune. The results in Table 5.3 showed that the credible intervals based on Bayesian approach are wider than their corresponding ML approach confidence intervals. The 95% confidence intervals based on the ML estimates revealed that the shape parameter of the GEV distribution is significantly negative (interval does not include zero) indicating that the flood heights at the site based on ML approach can be modelled by a short-tailed Weibull family of distributions. On the contrary, the 95% Bayesian

Table 5.3: Parameter estimates for Combomune

	ML estimates		
Parameter	**Estimate**	**SE**	**95% CI**
μ	5.162	0.278	$(4.884, 5.440)$
σ	1.661	0.198	$(1.463, 1.859)$
ξ	-0.123	0.109	$(-0.232, -0.014)$
	Bayesian estimates		
Parameter	**Estimate**	**Naive SE**	**95% CI**
μ	5.146	0.007	$(4.580, 5.741)$
σ	1.740	0.005	$(1.378, 2.162)$
ξ	-0.098	0.002	$(-0.027, 0.132)$

credible intervals showed that the shape parameter of the GEV distribution at Combomune is not significantly different from zero, implying that the flood heights at the site can be modelled by a Gumbel family of distributions. The parameter estimates in Table 5.3 were used to produce the results in Table 5.4 through the use of the quantile function in (5.5).

Table 5.4 presents results for the tail quantile estimates and predicted extreme flood heights. Generally the predicted flood heights based on the ML parameter estimates are lower than their corresponding Bayesian predicted flood heights. The results in Table 5.4 revealed that the highest flood height which occurred at the site is less than the 100-year flood height based on both the Bayesian and ML approach. However, it must be noted the maximum flood height at the site, 10.97 m, which occurred during the year 2000 disastrous floods is just about the 100-year flood height (11.00 m) based on the ML approach. The 13 m flood height which occurred at downstream Chokwe and Sicacate during the year 2000 floods has a return period in excess of 1,000 years based on the ML approach and in excess of 250 years based on the Bayesian approach. There is indication of spatial variability in the year 2000 floods.

Table 5.4: Tail quantile estimation and prediction of extreme flood heights for Combomune

1-p	p	T	**ML estimate** (Exceedances)*	**Bayesian estimate** (Exceedances)*
95^{th}	0.05	20 years	9.29 m (1)	9.63 m (1)
98^{th}	0.02	50 years	10.31 m (1)	10.78 m (1)
99^{th}	0.01	100 years	11.00 m (0)	11.59 m (0)
99.5^{th}	0.005	200 years	11.63 m (0)	12.33 m (0)
99.6^{th}	0.004	250 years	11.82 m (0)	12.56 m (0)
99.8^{th}	0.002	500 years	12.38 m (0)	13.24 m (0)
99.9^{th}	0.001	1,000 years	12.89 m (0)	13.88 m (0)
99.99^{th}	0.0001	10,000 years	14.32 m (0)	15.70 m (0)

Key: $1-p$ represents non-exceedance probability.
p represents exceedance Probability.
T represents return period.
(Exceedance)* represents the number of empirical observations above the flood level at the site.

Figure 5.2 in Appendix 5.1 presents the return level plot of the posterior distribution for Combomune. The graph also gives the credible intervals of the flood height estimates. The flood frequency curves in Figure 5.2 are based on the Bayesian MCMC inference and can be used to obtain any flood heights of interest and their corresponding return periods.

5.4.3 Sicacate site

Table 5.5 presents results for the parameter estimates of both the ML and Bayesian MCMC estimates for Sicacate. The results in Table 5.5 showed that credible intervals based on the Bayesian approach estimates are much wider than the confidence intervals based on the ML estimates. The shape parameter of the GEV distribution at Sicacate is significantly different from zero based on both the Bayesian and ML approaches. This is an indication that the annual maximum flood heights at Sicacate can be modelled by a short-tailed Weibull family of distributions.

Table 5.5: Parameter estimates for Sicacate

	ML estimates		
Parameter	**Estimate**	**SE**	**95% CI**
μ	6.151	0.467	$(5.684, 6.618)$
σ	3.328	0.347	$(2.981, 3.675)$
ξ	-0.454	0.071	$(-0.525, -0.383)$
	Bayesian estimates		
Parameter	**Estimate**	**Naive SE**	**95% CI**
μ	6.060	0.012	$(5.004, 7.139)$
σ	3.370	0.008	$(2.818, 4.347)$
ξ	-0.410	0.002	$(-0.572, -0.201)$

Table 5.6: Tail quantile estimation and prediction of extreme flood heights for Sicacate

1-p	**p**	**T**	**ML estimate** (Exceedances)*	**Bayesian estimate** (Exceedances)*
95^{th}	0.05	20 years	11.58 m (1)	11.85 m (1)
98^{th}	0.02	50 years	12.23 m (1)	12.62 m (1)
99^{th}	0.01	100 years	12.57 m (1)	13.03 m (0)
99.5^{th}	0.005	200 years	12.82 m (1)	13.34 m (0)
99.6^{th}	0.004	250 years	12.88 m (1)	13.42 m (0)
99.8^{th}	0.002	500 years	13.04 m (0)	13.64 m (0)
99.9^{th}	0.001	1,000 years	13.16 m (0)	13.80 m (0)
99.99^{th}	0.0001	10,000 years	13.37 m (0)	14.09 m (0)

Key: $1 - p$ represents non-exceedance probability.
p represents exceedance Probability.
T represents return period.
(Exceedance)* represents the number of empirical observations above the flood level at the site.

Table 5.6 presents results for the tail quantile estimates and the predicted extreme flood heights for Sicacate. The results in Table 5.6 showed that the predicted flood heights based on the ML parameter estimates are generally lower than their corresponding Bayesian MCMC based predicted flood heights. The empirically observed highest flood height of roughly 13 m that occurred

at the Sicacate during the year 2000 disastrous floods has a return period in excess of 250 (or nearly 500) years based on the ML approach and just about the 100-year flood height based on the Bayesian approach. These conflicting results would imply that based on the Bayesian approach the 13 m flood height at Sicacate, though extremely high to reach the 100-year return level, was not as devastating as would be portrayed by the ML approach.

Figure 5.3 in Appendix 5.1 presents the return level plot of the posterior distribution for Sicacate. The graph also gives the credible intervals for the estimates of the flood heights. The flood frequency curves in Figure 5.3 can be used to obtain any flood heights of interest and their associated return periods for Sicacate. The provision of credible intervals on the flood frequency curves makes the frequency curves more valid and usable since a range of return levels can be obtained for a particular return period or vice-versa.

5.5 General remarks on the results

Interesting findings were revealed in this chapter. It was generally found that the Bayesian MCMC estimates and predicted extreme flood heights were generally higher than their corresponding ML estimates and predicted extreme flood heights across all the three sites in the study. Moreover, for a particular extreme flood height (return level), the Bayesian MCMC based return period is substantially lower than that of the corresponding ML based return period. Although it is debateable to say which of the approaches give realistic estimates, it can be argued that the inclusion of uncertainties through the prior distribution in Bayesian MCMC based approach has greatly improved the results of the estimates and predicted flood heights. The consistency of the findings across all the sites strengthens the validity of these findings.

5.6 Added value and significance of the study in this chapter

The work in this chapter has been analysed with Bayesian MCMC using conjugate priors. The findings based on the Bayesian MCMC approach when compared to those based on ML approach (a frequentist approach) showed the importance of Bayesian inference approach in the basin, which is an approach that takes into account of the uncertainties associated with flood heights in the basin. Thus Bayesian MCMC can add value in complementing the other statistical approaches currently used, not only in the lower Limpopo River basin of Mozambique, but also in other various rivers and basins in Southern Africa and the world over. The use of Bayesian MCMC approach in flood frequency analysis is still limited, particularly in Southern Africa.

This work also supports the implementation of the Hyogo Framework for Action (UN, 2005) through addressing the gaps and challenges in reducing underlying risk factors and enhancing preparedness for effective floods disaster response and recovery. Accurate forecasts of the return periods of extreme floods help reduce the uncertainties associated with these natural hazards and thereby reduce the underlying risk factors and make the people well prepared to respond to these rare events. Mozambique is one of the developing (or least developed) countries in Southern Africa which is flood-prone and therefore warrants particular attention because of its vulnerability and risk levels which exceed its capacity to respond to and recover from flood disasters. Knowledge of the distribution of maximum flood heights and their corresponding return periods helps in substantial reduction of flood disaster-related losses in lives, in the social, economic and environmental assets of the country.

This work also contributes in the sharing of research findings, lessons learned

and best practices which are some of the aspects needed to enhance international and regional cooperation and assistance in disaster risk reduction.

5.7 Concluding remarks

This chapter has considered the block maxima approach of extreme value theory in an at-site flood frequency approach. The generalised extreme value distribution was used as the likelihood function and its parameters were estimated using two approaches; the Bayesian MCMC approach and maximum likelihood method (for the frequentist approach). The choice of the GEV distribution over other alternative distributions for the lower Limpopo River at the three sites Chokwe, Combomune and Sicacate hydrometric stations was inspired by the findings in the preceding chapters. The importance of the GEV distribution whose parameters are estimated by the Bayesian MCMC method was revealed in modelling the lower Limpopo River maximum flood heights.

The findings in this study have revealed one of the main merits of Bayesian MCMC approach of including uncertainties in the analysis through a prior distribution. It can be concluded that the inclusion of a prior distribution has substantially improved the precision of the tail quantile estimates and at-site predicted extreme flood heights. It was also found that the Bayesian 95% credible intervals were consistently wider that their corresponding maximum likelihood confidence intervals across all the three sites. Based on the findings of this study it can also be concluded that the estimates of the return periods based on Bayesian MCMC inference approach are substantially lower than those of their corresponding maximum likelihood approach. Additionally, the return levels based on Bayesian MCMC approach are also substantially higher than their corresponding maximum likelihood based return levels. These Bayesian estimates of maximum flood heights and their associated return periods appear

to be closer to reality than those based on ML approach. Modelling the upper tail efficiently is very important in flood frequency analysis as it results in reducing the impact of a flood event through disaster preparedness and management. These findings are in agreement with those in Reis and Stedinger (2005).

It has been found in this study that the Bayesian MCMC approach offers an extra mile in modelling the upper tails of annual maximum flood heights for the lower Limpopo River. Based on the findings in this chapter, it can be concluded that the Bayesian approach offer an improvement to the estimation of the parameters of the GEV distribution due to its ability to take into account the uncertainties involved in the hydrological processes of flood heights. The maximum likelihood approach (or frequentist in general) is not able to take into account such uncertainties.

5.8 Summary of the chapter

This chapter has considered the block maxima approach of EVT in an at-site realisation. The GEV distribution was used in this study as the likelihood function and its parameters were estimated by two approaches; Bayesian MCMC and ML (frequentist) approaches. The results in this study revealed that the 95% Bayesian credible intervals were wider than their corresponding ML 95% confidence intervals across all sites. It was also found that the estimates of the flood heights based on Bayesian MCMC approach were sufficiently higher than their associated ML based flood heights estimates across all sites. Modelling the upper tail efficiently is very important in flood frequency analysis as it results in reducing the flood event impact on humans and properties through disaster preparedness and management. The study has demonstrated the superiority of the Bayesian MCMC parameter estimation approach which does not depend on regularity conditions in a block maxima realisation. The ability

of the Bayesian MCMC method to take into account of uncertainties involved in the hydrological processes of flood heights offers an extra mile in modelling the upper tail behaviour of the annual maximum flood heights for the lower Limpopo River basin of Mozambique.

APPENDIX 5.1: FLOOD FREQUENCY CURVES OF POSTERIOR DISTRIBUTIONS FOR ALL THE THREE SITES

Return level plot for each site

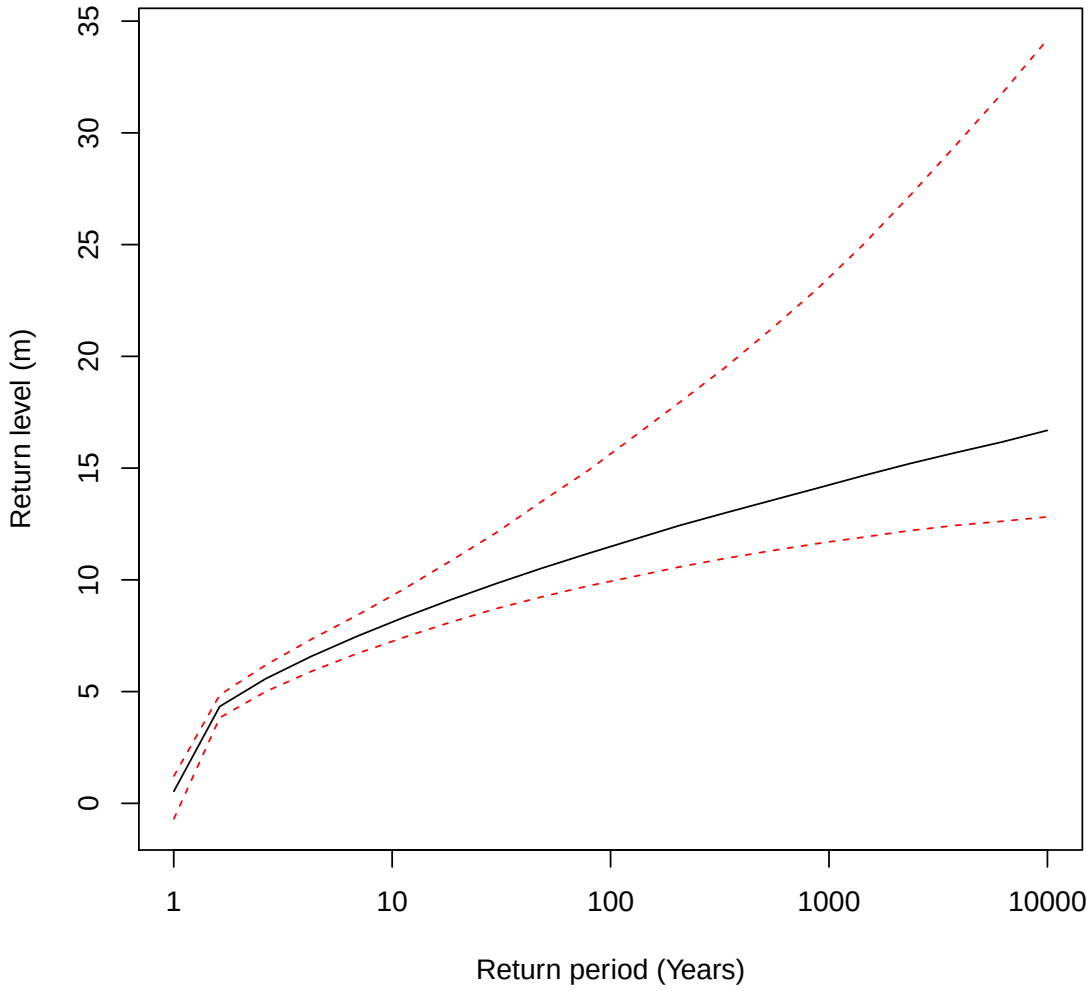


Figure 5.1: Return level plot of posterior distribution with 95% Bayesian credible intervals (dashed lines) at Chokwe hydrometric station

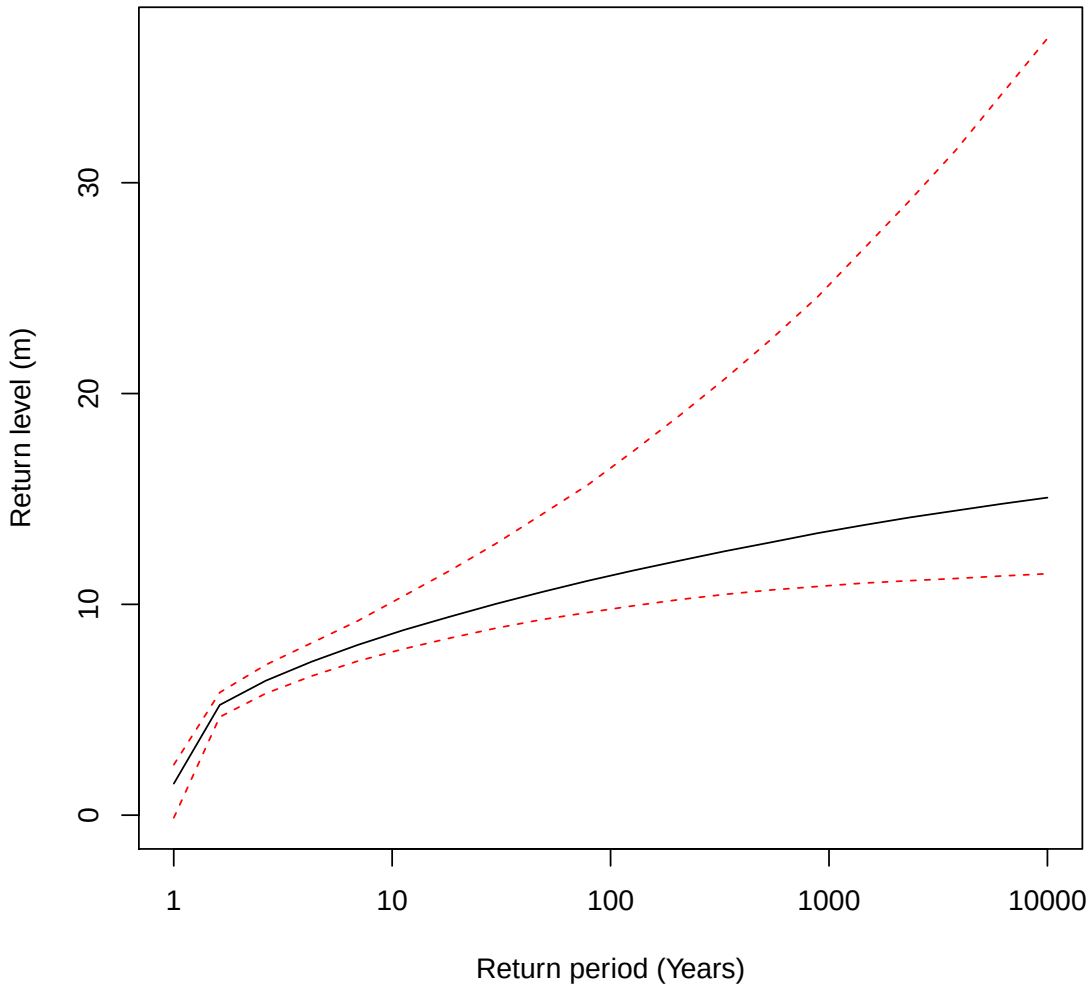


Figure 5.2: Return level plot of posterior distribution with 95% Bayesian credible intervals (dashed lines) at Combomune hydrometric station

APPENDIX 5.2: R PROGRAM CODE FOR BAYESIAN MCMC (SELECTED)

Bayesian MCMC R program code for Combomune site

attach(Combomune)

head(Combomune)

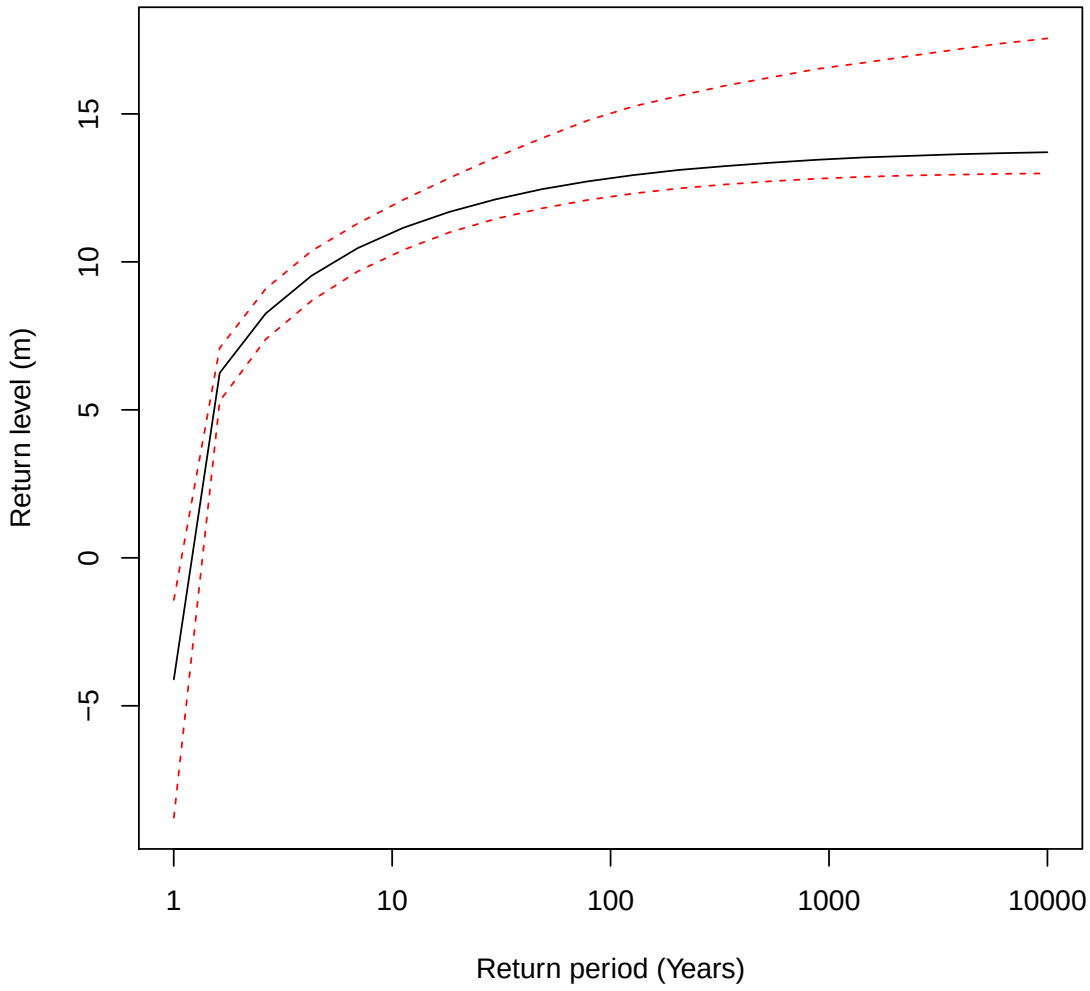


Figure 5.3: Return level plot of posterior distribution with 95% Bayesian credible intervals (dashed lines) at Sicacate hydrometric station

```
tail(Combomune)
summary(Combomune)


#create a new window for graphs
win.graph()
win.graph() par(mfrow = c(2,3)) # graph window with 2 rows and 3 columns
```

```
#fitting GEVD using R package ismev (Frequentist approach) MLE
library(ismev)
freq=gev.fit(AM1)
gev.diag(freq)

#Bayesian MCMC approach install.packages("evdbayes") library(evdbayes)
mat = diag(c(10000, 10000, 100))
pn = prior.norm(mean = c(0,0,0), cov = mat)

#Using best t0 and s
n = 2000 ; t0 = c(5.1623177,1.6611994,-0.1230409);s = c(0.2783756,0.1978411,0.1091684)
ptpmc = posterior(n, t0, prior = pn, lh = "gev", data = AM1, psd = s)
attributes(ptpmc)$ar

#MCMC works with coda
install.packages("coda")
library(coda)
ptp.mcmc = mcmc(ptpmc, start = 0, end = 2000)
plot(ptp.mcmc, den = FALSE, sm = FALSE, main="")

#best initial values of to
maxpst = mposterior(t0, prior = pn, lh = "gev", method = "SANN",data = AM1)
round(maxpst$par, 2)

#best initial values of s
t0 = c(5.16, 1.66, -0.12); psd = rep(0.01, 3)
psd = ar.choice(init = t0, prior = pn, lh = "gev", data = ADMS, psd = psd, tol =
rep(0.02, 3))$psd
round(psd, 2)
```

# TRACE OF THE PARAMETERS
t0 = c(5.16, 1.66, -0.12) ; s ¡- c(0.28,0.20,0.11)
ptpmc = posterior(n, t0, prior = pn, lh = ”gev”, data = ADMS, psd = s)
ptp.mcmc = mcmc(ptpmc, start = 0, end = 2000)
plot(ptp.mcmc, den = FALSE, sm = FALSE,col=”blue”, main=””)

#diagnostic checks Geweke
ptp.mcmc = window(ptp.mcmc, start = 100)
geweke.diag(ptp.mcmc)
geweke.diag(ptp.mcmc, 0.2, 0.4)

#Marginal posterior densities and summary statistics
bwf = function(x) sd(x)/2
plot(ptp.mcmc, trace = T, bwf = bwf,col=”blue”, main=””)
summary(ptp.mcmc)

#Return level plot of posterior distribution with 95% Bayesian credible intervals
rl.pst(ptp.mcmc, npy, lh = ”gev”, ci = 0.95, lty = c(2,1), col = c(2,1),
xlab = ”Return period (Years)”, ylab = ”Return level (m)”)

# Quantile estimation using the quantile function
u = 5.162
sig = 1.661
xi = -0.123
p = 0.0001
$X_p = u + (sig/xi)((-log(1-p))^{-\xi} - 1)$
Xp

Chapter 6

Modelling nonstationary extremes using a GEV distribution in the lower Limpopo River basin of Mozambique

6.1 Introduction

There is a general notion that the occurrence of extreme events has changed over these recent years and is anticipated to continue to change in terms of intensity, frequency and complexity of the risks (Stal et al., 2014). These recent changes are mainly attributed to global warming and natural modes of interannual and interdecadal variability such as the El Niño phenomenon (Katz, 2010;

Towler et al., 2010). These climate changes which often result in extreme natural hazards such as floods, droughts, tsunamis and earthquakes, can also cause negative societal impacts and disruptions, for instance; destruction of schools, children dropping out of schools leading to early marriages particularly for girls and therefore creating a vicious poverty circle in the community (Mudavanhu, 2014). According to Katz (2010) previous studies in EVT have shown that the frequency of all forms of extreme events, whether in the form of a single value or a sequence of annual maxima, is more sensitive to variations in the scale parameter (or in particular, the standard deviation) than to the location parameter (or mean) of a distribution. Cooley (2009) wrote a commentary on the potential application of statistics of extremes to climate change based on the previous work of Wigley.

The block maxima, also known as annual maximum series (AMS), has long been employed to estimate the distribution of extreme events such as flood flows, precipitation and wind speeds (see Chapters 2 & 3). The time-homogeneous GEV distribution which uses standard properties of the likelihood function has traditionally been used in design flood estimation (Towler et al., 2010; Coles, 2001). The use of a stationary GEV distribution assumes that climate changes and all other variables that may affect the validity of the estimation of design floods remain constant over time (Gumbel, 1941). More than six decades ago, Gumbel (1941) realised that the then newly developed EVT techniques which could, apparently, only be applied under the assumption of stationarity at the time could not survive the test of time, particularly if there were changes in climate and other important changes in the basins.

According to Coles (2001), stationarity refers to a physical process whose random variables may be mutually dependent, but have time-homogeneous stochastic properties. Stationarity is a more realistic assumption to a large number

of physical processes. One of the main characteristics of nonstationarity is to change with time (Yilmaz et al., 2014; Coles, 2001). In environmental processes such as floods, nonstationarity is often a result of seasonality and trends. The former is usually attributed to climate changing with months and the latter is possibly due to long-term climate changes (Coles, 2001). When nonstationarity arises in block maxima or POT, standard models are usually modified to accommodate the changes in the parameters.

There has been a gradual increase in the frequency of floods. For instance, a unique survey of 139 national meteorological and hydrological services carried out by the WMO in 2013 revealed that floods were the most frequently experienced extreme events worldwide over the course of the decade 2001-2010 (WMO, 2013). Floods and droughts account for 90% of all the people that are affected by natural disasters (Smakhtin, 2014; MunichRe, 2013). According to MunichRe (2013) the natural catastrophic statistics for the year 2013 was dominated by floods that caused billions of American dollars in losses.

Evidence of nonstationarity is exhibited in Figures 1.4, 1.5 & 1.6 in Chapter 1 at all the three sites. However, a visual inspection of the plots in the figures shows no apparent trend, but reveals that the year 2000 flood height was a very rare extreme event at all the three sites.

The present study considers a nonstationary time-dependent GEV distribution model whose location and scale parameters are expected to vary linearly or nonlinearly with time, while the shape parameter remains constant over time (Vasiliades et al., 2015; Towler et al., 2010). Detailed studies on the GoF of the time-homogeneous GEV distribution in the basin are found in Chapter 3. In this present study, the researcher advocates for a statistical modelling approach based on MLE technique in the possible presence of covariates. The co-

variates of particular interest in the present study are the trend and a physical variable known as seasonal oscillation index (SOI) (Coles, 2001). The inclusion of SOI in the model is motivated by the frequency of cyclone-induced floods in the LLRB of Mozambique.

SOI is an index that is used in the quantification of El Niño southern oscillation (ENSO) (Reid, 2000). In a study to analyse the signal of ENSO in tropospheric and stratospheric temperatures Fernández et al. (2004) used data for the period 1979-2000 from Microwave Sounding Unit (MSU) in Madrid, Spain. These authors used regressions of the data from the MSU on the principal components of the tropical empirical orthogonal functions to estimate the global signal of ENSO. The results of the study showed that over two-thirds of the variability of temperature in the tropical troposphere (atmosphere) is explained by ENSO. Reid (2000) described ENSO as an abnormally large scale ocean-atmosphere system that is usually associated with strong variability in ocean currents and atmospheric surface temperatures. El Niño is a South American term that means “The Little Boy ” and it refers to anomalous warm ocean temperatures, while another common phenomenon La Niña is a Spanish term that means “The Little Girl” and it refers to unusually cold surface temperatures in the equatorial part of the Pacific (Reid, 2000).

According to Reid (2000) El Niño events usually bring drier conditions, while La Niña events generally bring wet conditions resulting in unusually high rainfall in some regions. In the past El Niño events were relatively known to be rare events, however, in recent decades, particularly in the 1980s and 1990s these events occurred more frequently than before and lasted longer (Fernández et al., 2004; Reid, 2000). In the 20^{th} century the longest El Niño persisted from 1991 to 1995. According to WMO (1999) the five well-known El Niño events that also affected the region of Southern Africa since 1970 were in 1972-1973,

1982-1983, 1986-1988, 1991-1995 and 1997-1998. These events are also reflected in flood heights data series used in this thesis (Figures 1.1-1.6, Chapter 1). In recent years since the beginning of the 21^{st} century there have not been severe incidence of El Niño events, particularly in Southern Africa. However, La Niña effects have been frequently felt in the current century evidenced by an increased frequency of flooding events and cyclone activities particularly in Mozambique and most countries in Southern Africa (Jackson, 2013a,b).

Negative values of SOI represent El Niño events which are characterised by warmer than average conditions, while large positive values of SOI represent La Niña conditions, which are usually characterised by wetter than usual conditions (Reid, 2000).

According to Katz et al. (2002), the most popular parameter estimation method in the application of hydrological extremes are the L-moments and PWM compared to MLE method mainly due to their computational simplicity and better performance for small samples where MLE is often inconsistent. However, the PWM technique has the drawback of being unable to readily incorporate covariates (Katz et al., 2002). On the other hand, the application of MLE technique in the presence of covariates is straightforward in both block maxima and POT approaches (Yilmaz et al., 2014; Katz et al., 2002; Coles, 2001).

An extensive literature review on nonstationary extremes is covered in Chapter 2 of this thesis. In Australia Verdon-Kidd and Kiem (2015) argued that the assumption of climate stationarity has been put to question over the past 15 years. The author recommended that in order to address the challenges associated with impact of climate change on extreme events increased collaboration is required between climate scientists and statisticians.

Yilmaz et al. (2014) asserted that the increased frequency and magnitude of floods and other extreme events due to climate change, put to question the assumption of stationarity. Yilmaz et al. (2014) used nonstationary GPD models to investigate trends and other nonstationarity characteristics in extreme flood events and also investigated the potential climate change impacts and variability on intensity-frequency-duration (IFD) relationships at an observation station in Melbourne, Australia. These authors found trend to be statistically present in the storm durations of half-an hour, 3 hours and 48 hours. On the other hand, they did not find evidence of nonstationarity for all storm durations and concluded that stationary GPD models could be used to model the rainfall data for all durations at the site. Using IFD curves analysis, Yilmaz et al. (2014) concluded that urban flash floods that produce hourly rainfall had increased over time for Melbourne.

In Africa, Aich et al. (2014) studied climate variability in four large river basins including the LLRB of Mozambique. These authors used a geoscientific approach to study the impacts of climate change on the streamflow of the Limpopo, Niger, Oubangui and Upper Blue Nile. An eco-hydrological model named Soil and Water Integrated Model (SWIM) was set up to each of the four basins and the validation of the models showed that the results were adequately good. These authors assessed the impact of climate change through a comparison of the trends in seasonality, mean discharges and hydrological extremes in the 21^{st} century. From the study, Aich et al. (2014) found the uncertainty of projections to be lowest in the Upper Blue Nile which is also most likely to experience an increased streamflow. The Limpopo and Niger basins were found to experience high magnitudes of trends accompanied by a wide range of uncertainty. The least impact of climate change was found to be at Oubangui among the other basins. These results by Aich et al. (2014) are of paramount importance to the researcher, particularly for this chapter and the next one, since the

Limpopo River is being studied in this thesis. The existence of high magnitude of trends in the mean and hydrological extremes, and wide range of uncertainty will be investigated in the nonstationary context using statistics of extremes in this thesis. It will be of interest to find out if the two distinct approaches can lead to the same conclusions. To the best of our knowledge no similar work in the previous studies relating to statistics of extremes in a changing climate has been done for the LLRB of Mozambique.

The outline of the rest of the chapter is such that Section 6.2 presents the research methodology, Section 6.3 presents the results and discussion of the findings, Section 6.4 briefly highlights the value added and significance of this chapter, while Section 6.5 concludes the chapter and Section 6.6 presents a brief summary of the chapter. Finally Appendix 6.1 presents selected R code programs used in modelling the data for the chapter, while Appendix 6.2 presents some figures and diagnostic plots for the chapter.

6.2 Research methodology

This section presents the methods used to analyse the data. It also discusses, in brief, the probability framework of block maxima including the extension of the time-homogeneous GEV model to linear and quadratic models that include a trend and a physical meteorological variable SOI. A method for the choice of a worthwhile model in nested GEV models is also given in this section.

6.2.1 Block maxima and moving sums

Like in the previous chapters, the data used in the study for this chapter was provided by the Mozambique National Directorate of Water (DNA), the authority responsible for water management in Mozambique. The data are hydrometric daily flood heights (in metres) recorded at the sites Chokwe (1951-2010),

Combomune (1966-2010) and Sicacate (1952-2010) hydrometric stations for the lower Limpopo River of Mozambique. The three sites are such that Combomune is located in the upper part of the basin about 162 km from the border with South Africa and Zimbabwe, Chokwe is located in the middle of the basin about 130 km downstream of Combomune and Sicacate is further downstream of Chokwe in the lower part of the basin on way to the sea.

In order to obtain AMS sequential steps were taken to obtain the highest flood height in each hydrological year (or block). Further steps were taken to obtain annual maximum (AM) flood heights of the moving sums of 2-days, 5-days, 7-days, 10-days, and 30-days. Finally the following AM time series models were obtained: annual daily maximum (AM1), annual 2-day maximum (AM2), annual 5-day maximum (AM5), annual 7-day maximum (AM7), annual 10-day maximum (AM10), and annual 30-day maximum (AM30). The procedure to obtain these cumulative AM time series models was necessitated by the need to investigate whether the cumulative annual floods have any significant effect on the long-term linear or quadratic trend in either location, scale or both.

6.2.2 Nonstationary extreme value models

The reader is now familiar with the two fundamental approaches of flood frequency analysis (FFA), at-site and regional (Smithers, 2012) and the two fundamental approaches of EVT, block maxima and POT (Ferreira and de Haan, 2015). The approach used in this study is block maxima at a particular site. In hydrological studies, when sample sizes are large it is natural to block observations by years (Ferreira and de Haan, 2015; Smithers, 2012). Comprehensive details of probability framework of block maxima and the practical reasons for using block maxima over POT are given in Ferreira and de Haan (2015) and also in Chapter 2.

The time-homogeneous GEV model in (2.6) shall be called M_0 and it shall be used as a reference point such that all other extended models are compared to it for their significance and worthiness.

The log-likelihood function for the GEV parameters in (2.6), for $\xi \neq 0$, is given by

$$\begin{aligned} L_k = \ell(\mu, \sigma, \xi) &= -k \ln \sigma - \left(\frac{1}{\xi} + 1\right) \sum_{i=1}^{k} \ln \left[1 + \xi \left(\frac{x_i - \mu}{\sigma}\right)\right]_+ \\ &\quad - \sum_{i=1}^{k} \left[1 + \xi \left(\frac{x_i - \mu}{\sigma}\right)\right]_+^{\frac{-1}{\xi}}. \end{aligned} \tag{6.1}$$

The log-likelihood for case $\xi = 0$, using the Gumbel limit of the GEV (Coles, 2001), is

$$L_k = \ell(\mu, \sigma) = -k \ln \sigma - \sum_{i=1}^{k} \left(\frac{x_i - \mu}{\sigma}\right) - \sum_{i=1}^{k} \exp\left\{-\left(\frac{x_i - \mu}{\sigma}\right)\right\} \tag{6.2}$$

where k is the number of years (blocks) and x_i is the flood height.

The MLEs of the parameters μ, σ and ξ are obtained from the solution of the likelihood equations obtained from the partial derivatives of the log-likelihood L_k,

$$\nabla L_k = 0 \quad \text{with } \nabla L_k = \left(\frac{\partial L_k}{\partial \mu}, \frac{\partial L_k}{\partial \sigma}, \frac{\partial L_k}{\partial \xi}\right). \tag{6.3}$$

The MLE is any solution of (6.3) with a negative definite Hessian matrix (Dombry, 2015). Most recent literature on the existence of consistent MLEs is abundant in Dombry (2015). More details on the regularity conditions of the MLEs are found in literature (Rajaram, 2006; Beirlant et al., 2004; Coles, 2001; Smith, 1987, with references therein).

The extreme quantile estimates of the GEV are provided by the quantile func-

tion, x_p

$$X_{p_i} = G^{-1}(1-p_i) \;=\; \begin{cases} \mu + \frac{\sigma}{\xi}\left[(-\ln(1-p_i))^{-\xi}\right], & \xi \neq 0, \\ \mu - \sigma \ln\left(-\ln(1-p_i)\right), & \xi = 0, \end{cases} \tag{6.4}$$

where p_i is the exceedance probability and $T_i = \frac{1}{p_i}$ is the return period of an extreme flood height, x_{p_i}.

Now consider the time-heterogeneous GEV model, call it M_1, with a linear trend in the location and scale parameters such that $\mu(t) = \mu_0 + \mu_1 t$, $\log\sigma(t) = \sigma_0 + \sigma_1(t)$, and $\xi(t) = \xi$, where t is time in years, then the general model is given in (6.5)

$$G(\mu(t), \sigma(t), \xi(t); x_i, t) \;=\; \begin{cases} \exp\left(-\left[1+\xi\left(\frac{x_i-(\mu_0+\mu_1 t)}{\exp(\sigma_0+\sigma_1 t)}\right)\right]_+^{-\frac{1}{\xi}}\right), & \xi \neq 0, \\ \exp\left(-\exp\left(-\frac{x_i-(\mu_0+\mu_1 t)}{\exp(\sigma_0+\sigma_1 t)}\right)\right), & x_i \in \mathbb{R},\ \ \xi = 0, \end{cases} \tag{6.5}$$

The log-likelihood function of model M_1 in (6.5) is

$$\begin{aligned} L_k(t) = \ell(\mu(t), \sigma(t), \xi(t); x_i, t) \;=\; & -k\ln(\sigma_0+\sigma_1 t) \\ & -\left(\frac{1}{\xi}+1\right)\sum_{i=1}^{k}\ln\left[1+\xi\left(\frac{x_i-(\mu_0+\mu_1 t)}{\exp(\sigma_0+\sigma_1 t)}\right)\right]_+ \\ & -\sum_{i=1}^{k}\left[1+\xi\left(\frac{x_i-(\mu_0+\mu_1 t)}{\exp(\sigma_0+\sigma_1 t)}\right)\right]_+^{\frac{-1}{\xi}}, \end{aligned} \tag{6.6}$$

with the usual replacement when $\xi = 0$. The MLEs of the parameters $\mu_0, \mu_1, \sigma_0, \sigma_1$ and ξ are obtained by writing a computer programme in R software and running the programme.

In the present study three more models are proposed for the linear trend: M_2, M_3 and M_4. Model M_2 has a linear trend in the location parameter such that $\mu(t) = \mu_0 + \mu_1 t$, $\sigma(t) = \sigma$ and $\xi(t) = \xi$, and hence model M_2 and its log-likelihood are of the form $G(\mu(t), \sigma, \xi; x_i, t)$and $l\left(\mu_0, \mu_1, \sigma, \xi; x_i, t\right)$. As for the other two models, M_3 has a linear trend in the scale parameter and M_4 has a nonlinear quadratic trend in the location parameter such that $\mu(t) = \mu$, $\log \sigma(t) = \sigma_0 + \sigma_1 t$, $\xi(t) = \xi$ and $\mu(t) = \mu_0 + \mu_1 t + \mu_2 t^2$, $\sigma(t) = \sigma$, $\xi(t) = \xi$, for models M_3 and M_4, respectively. The model for M_3 and its log-likelihood are of the form $G(\mu, \sigma(t), \xi; x_i, t)$ and $l(\mu, \sigma_0, \sigma_1, \xi; x_i, t)$, respectively, while the model for M_4 and its log-likelihood are of the form $G(\mu(t), \sigma, \xi; x_i, t)$ and $l(\mu_0, \mu_1, \mu_2, \sigma, \xi; x_i, t)$, respectively.

Additionally, two more models are proposed for the climate change covariate as measured by the SOI variable. More details on the SOI are given in the results section. Concerning the inclusion of SOI in the model consider two more models M_5 and M_6. Let model M_5 contain SOI only in the GEV model such that the location parameter of the GEV model becomes $\mu(t) = \mu_0 + \mu_1 SOI(t)$ while model M_6 includes both a trend and SOI such that the location parameter of the GEV model is of the form $\mu(t) = \mu_0 + \mu_1 SOI(t) + \mu_2 t$. The model for M_5 and its log-likelihood are of the form $G(\mu(t), \sigma, \xi; x_i, t)$ and $l(\mu_0, \mu_1, \sigma, \xi; x_i, t)$, in respective order, while the model for M_6 and its log-likelihood are of the form $G(\mu(t), \sigma, \xi; x_i, t)$ and $l(\mu_0, \mu_1, \mu_2, \sigma, \xi; x_i, t)$, respectively.

6.2.3 Model choice

One important question to answer in extreme value analysis is whether the nonstationary model provides an improvement in fit over the time-homogeneous simpler model M_0, i.e., is it worthwhile to have the nonstationary model? The MLE of nested models uses a simple procedure called the deviance (D) statistic to compare one model against the other (Yilmaz et al., 2014; Coles, 2001;

Smith, 1987). In this study the time-homogeneous GEV model, M_0, is a special case of the time-dependent models M_1, M_2, M_3 and M_4. In general, we consider $M_0 \subset M_{i,i=1,2,3,4}$, then we define deviance statistic, D, as

$$D = 2\left[l_i(M_i) - l_0(M_0)\right], \tag{6.7}$$

where $l_i(M_i)$ and $l_0(M_0)$ are the maximum negative log-likelihood (NLLH) for models $M_{i,i=1,2,3,4}$, and M_0, respectively. D has a Chi-square, $\chi^2_{k,\alpha}$, asymptotic distribution, with k degrees of freedom tested at α $(= 0.05$ or $5\%)$ level of significance, where k is the difference in dimensionality (or difference in number of parameters) between M_i and M_0. Thus, D is compared to critical values of $\chi^2_{k,\alpha}$ where $D > \chi^2_{k,\alpha}$ suggests that model M_i explains more of the variability in the data than model M_0 (Coles, 2001).

6.3 Results and discussion

This section presents the results of the study for this chapter. In order to avoid presenting too many tables for this chapter, only the tables and figures for the AM1 time series data are presented for each of the three sites: Chokwe, Combomune and Sicacate in Tables 6.1, 6.2 and 6.3, respectively. The results for the AMS moving sums for each site are only discussed in detail if there is discrepancy with the AM1 results, else they are simply mentioned if there is consistency. The order of the models is maintained for the AMS moving sums, e.g. for AM2 time series data, model M_1 still refers to a time-heterogeneous GEV model with a linear trend in both the location and scale parameters as in AM1. The diagnostic plots results for this chapter are presented in Appendix 6.2 at the end of the chapter. Appendix 6.1 presents some selected R programs written for this chapter.

6.3.1 Chokwe models

Consider the pair of models (M_0, M_1) from Table 6.1 where M_0 is taken as the reference model: $\chi^2_{2,0.05} = 5.991$, $D = 2(-125.802 - (-126.313)) = 1.022$ and the likelihood ratio test for $\mu = 0$ has p-value $= 0.4928$ and for $\sigma_1 = 0$ has p-value $= 0.1676$. Since D is too small compared to the critical value (5.991) and the likelihood ratio test is not significant at 5% level of significance (p-value > 0.05) for both the location and scale parameters, it clearly shows that the nonstationary model is not important and does not give any improvement in fit over the time-homogeneous GEV model. Similar insignificant results were obtained for the AMS moving sums AM2, AM5, AM7, AM10, and AM30 for model M_1.

Table 6.1: AM1 time-heterogeneous GEV models for Chokwe for the period 1951-2010.

Model	$\hat{\mu_0}$	$\hat{\mu_1}$	$\hat{\mu_2}$	$\hat{\sigma_0}$	$\hat{\sigma_1}$	$\hat{\xi}$	**NLLH**
M_0	4.248	0	0	1.785	0	-0.081	126.313
M_1	4.222	-0.0003	0	2.204	-0.015	-0.041	125.802
M_2	4.294	-0.0015	0	1.787	0	-0.081	126.308
M_3	4.212	0	0	2.205	-0.015	-0.040	125.802
M_4	4.237	-0.0002	0.0000	1.784	0	-0.080	126.310

The other pairs from Table 6.1 (M_0, M_2) and (M_0, M_3) have $D = 0.01$ and 1.022, respectively, with a critical value of $\chi^2_{1,0.05} = 3.841$, for both pairs. The likelihood ratio test for $\mu_1 = 0$ has p-value $= 0.4591$ and $\sigma_1 = 0$ has p-value $= 0.1653$ for M_2 and M_3, respectively, which is insignificant for both models at the 5% level of significance. The D statistic is again too small (< 3.841) for both models implying that both models do not provide any improvement in fit over the time-homogeneous GEV model. Similar insignificant results were obtained for the AMS moving sums AM2, AM5, AM7, AM10, and AM30 for both models M_2 and M_3.

The quadratic model pair (M_0, M_4) in Table 6.1 has a D statistic value of 0.006 with a critical value of $\chi^2_{2,0.05} = 5.991$, implying that model M_4 does not provide any improvement in fit to justify its importance over the time-homogeneous model. The likelihood ratio tests for $\mu_1 = 0$ and $\mu_2 = 0$ are also not significant at 5% level of significance (p-value > 0.05). Again similar results were obtained for the AMS moving sums AM2, AM5, AM7, AM10, and AM30.

In general, the results for Chokwe showed that the prevailing model for the site is the time-homogeneous GEV model given by

$$G(\mu, \sigma, \xi; x_i) = \exp\left\{-\left(1 + \frac{-0.081\,(x_i - 4.248)}{1.785}\right)_+^{\frac{1}{0.081}}\right\}, \tag{6.8}$$

where $x_{i,\forall i=1,2,\ldots,k}$ is the annual maximum flood height at the site.

The shape parameter (-0.081) for the prevailing model in (6.8) is not significantly different from zero (p-value $= 0.136 > 0.05$) implying that it is not significantly negative. This suggests that the extreme floods at Chokwe hydrometric station follow a Gumbel distribution which is light-tailed. The diagnostic plots for the time-homogeneous model in (6.8) are presented in Figure 6.4 in Appendix 6.2. The diagnostic plots in Figure 6.3 show that the model is of good fit, with the exception of the year 2000 flood height which falls slightly outside the confidence limits.

6.3.2 Combomune models

Consider the pair (M_0, M_1) from Table 6.2 with $\chi^2_{2,0.05} = 5.991$, $D = 5.36$, and likelihood ratio test for $\mu_1 = 0$ with p-value $= 0.2570$ and $\sigma_1 = 0$ with p-value $= 0.0117$. These results show that the linear trend in location parameter is not significant at 5% significance level (p-value > 0.05), while the linear trend in

scale parameter is significant (p-value < 0.05) in the model. In other words, the scale parameter is time-dependent while the location parameter is time-homogeneous. However, the D statistic (5.36) is less than the critical value of 5.991 at 2 degrees of freedom which implies that the nonstationary model M_1 is not worthwhile compared to the time-homogeneous GEV model, M_0. The same conclusions were reached for the moving sums of AM2, AM5, AM7, AM10 and AM30.

Table 6.2: AM1 time-heterogeneous GEV models for Combomune for the period 1966-2010.

Model	$\hat{\mu_0}$	$\hat{\mu_1}$	$\hat{\mu_2}$	$\hat{\sigma_0}$	$\hat{\sigma_1}$	$\hat{\xi}$	**NLLH**
M_0	5.163	0	0	1.660	0	-0.124	90.740
M_1	5.394	-0.012	0	2.321	-0.033	-0.045	88.060
M_2	5.445	-0.011	0	1.685	0	-0.150	90.614
M_3	5.034	0	0	2.268	-0.031	-0.043	88.276
M_4	5.338	-0.000	-0.000	1.681	0	-0.146	90.627

Now consider the pairs (M_0, M_2) and (M_0, M_3) from Table 6.2. The critical value for both pairs is $\chi^2_{1,0.05} = 3.841$ with respective D statistic values of 0.252 and 4.928 for the two pairs. The likelihood ratio test for $\mu_1 = 0$ has p-value $= 0.3092$ and $\sigma_1 = 0$ has p-value $= 0.0141$ for models M_2 and M_3, respectively. These results show that model M_2 is not significant at 5% significance level (p-value > 0.05). On the other hand, model M_3 is significant at 5% significance level (p-value < 0.05) and provides an improvement in fit over the time-homogeneous GEV model since the D statistic value of 4.928 (> 3.841) is significantly large. Similar findings were obtained for all the other AMS moving sums for the site.

The nonlinear quadratic model pair (M_0, M_4) in Table 6.2 has a D statistic of 0.226 which is too small compared to the critical value of 5.991 with 2 degrees of freedom. The likelihood ratio tests for $\mu_1 = 0$ and $\mu_2 = 0$ are not signifi-

cant at 5% significance level (p-value > 0.05). Thus the nonlinear quadratic model M_4 is neither significant nor worthwhile over the time-homogeneous GEV model. Likewise the same conclusions were reached for all the other AMS moving sums.

Overall, the final model for Combomune is the nonstationary model, M_3, with a linear trend in the scale parameter of the GEV. The general model for Combomune is given by

$$G\left(\mu, \sigma(t), \xi; x_i, t\right) = \exp\left\{-\left(1 + \frac{-0.043\left(x_i - 5.034\right)}{\exp\left(2.268 - 0.031 t_i\right)}\right)_{+}^{\frac{1}{0.043}}\right\}, \tag{6.9}$$

where $t_i = \tau_i - 1965$, $\tau_i = 1966, 1967, ...$ and $t_i = 1, 2, 3, ...$ is the time in years and $x_{i,\forall i=1,2,...,k}$ is the annual maximum flood height at the site.

The shape parameter (-0.043) for the prevailing model in (6.9) is not significantly different from zero (p-value $= 0.334 > 0.05$). This implies that the shape parameter (-0.043) is not significantly negative, suggesting that the extreme floods at Combomune hydrometric station follow a Gumbel distribution which is light-tailed. The diagnostic plots for the time-heterogeneous model in (6.9) are presented in Figure 6.5. The residual probability plot suggests a good fit to the data.

6.3.3 Sicacate models

The model pair (M_0, M_1) from Table 6.3 has $\chi^2_{2,0.05} = 5.991$ and a D statistic value of 8.482. The likelihood ratio test for $\mu_1 = 0$ has p-value $= 0.3217$ and $\sigma_1 = 0$ has p-value $= 0.0045$ for the model M_1 which indicates that the linear trend in location parameter is not significant at 5% significance level (p-value > 0.05) whereas the linear trend in scale parameter is highly significant (p-value < 0.05) in the model. Since the D statistic value (8.482) is greater than the critical

value of 5.991 we conclude that model M_1 provides an improvement in fit over the time-homogeneous GEV model, that is, model M_1 is worthwhile. These findings are consistent with findings from AMS moving sums AM2, AM5, AM7, AM10 and AM30.

Table 6.3: AM1 time-heterogeneous GEV models for Sicacate for the period 1952-2010.

Model	$\hat{\mu_0}$	$\hat{\mu_1}$	$\hat{\mu_2}$	$\hat{\sigma_0}$	$\hat{\sigma_1}$	$\hat{\xi}$	**NLLH**
M_0	6.151	0	0	3.328	0	-0.454	148.547
M_1	6.901	-0.012	0	1.813	0.061	-0.682	144.306
M_2	5.499	-0.025	0	3.443	0	-0.526	148.279
M_3	6.675	0	0	1.966	0.055	-0.693	144.413
M_4	5.887	-0.000	-0.0003	3.396	0	-0.499	148.373

The model pairs (M_0, M_2) and (M_0, M_3) in Table 6.3 share a critical value of $\chi^2_{1,0.05} = 3.841$ with a D statistic value of 0.536 and 8.268 for M_2 and M_3, respectively. The likelihood ratio test for $\mu_1 = 0$ has p-value $= 0.2638$ and $\sigma_1 = 0$ has p-value $= 0.0013$ for M_2 and M_3, respectively. This indicates that model M_2 is insignificant at 5% level of significance (p-value > 0.05) and not worthwhile ($D < 3.841$), while model M_3 is highly significant at 5% significance level (p-value < 0.05) and provides an improvement in fit over the time-homogeneous GEV model, with a large D statistic value of 8.268 (> 3.841). These findings are also consistent with findings from all the AMS moving sums.

The nonlinear quadratic model pair (M_0, M_4) in Table 6.3 has a D statistic value of 0.348 with a critical value of $\chi^2_{2,0.05} = 5.991$. The likelihood ratio test for $\mu_1 = 0$ has p-value $= 0.4991$ and $\mu_2 = 0$ has p-value $= 0.0001$. This implies that the quadratic trend term in location parameter is highly significant (p-value < 0.0001). However, the overall nonlinear quadratic model is not worthwhile since the D statistic value of 0.348 is too small compared to the critical value of 5.991. Again, these findings are consistent with findings from

all the AMS moving sums.

The two best competing 'best' nonstationary linear time-homogeneous models for Sicacate ranked based on their standard errors and p-values as main and alternative model, are respectively

$$G(\mu, \sigma(t), \xi; x_i, t) = \exp\left\{-\left(1 + \frac{-0.693\,(x_i - 6.675)}{\exp\,(1.966 + 0.055 t_i)}\right)_+^{\frac{1}{0.693}}\right\}, \tag{6.10}$$

where $t_i = \tau_i - 1951$, $\tau_i = 1952, 1953, ...$ and $t_i = 1, 2, 3, ...$ is the time in years and $x_{i,\forall i=1,2,...,k}$ is the annual maximum flood height at the site. The alternative nonstationary linear trend in location and scale GEV model is

$$G(\mu(t), \sigma(t), \xi; x_i, t) = \exp\left\{-\left(1 + \frac{-0.682\,(x_i - (6.901 - 0.012 t_i))}{\exp\,(1.813 + 0.061 t_i)}\right)_+^{\frac{1}{0.682}}\right\}, \tag{6.11}$$

where $t_i = \tau_i - 1951$, $\tau_i = 1952, 1953, ...$ and $t_i = 1, 2, 3, ...$ is the time in years and $x_{i,\forall i=1,2,...,k}$ is the annual maximum flood height at the site.

The shape parameters in (6.10) and (6.11), that is, -0.693 and -0.682, are significantly different from zero (p-value < 0.0001) implying that the shape parameters for the two models at Sicacate are significantly negative. This suggests that extreme floods at Sicacate hydrometric station can be modelled by a Weibull distribution which is short-tailed. The diagnostic plots for the time-heterogeneous models in (6.10) and (6.11) are presented in Figures 6.6 and 6.7, respectively. The residual probability plots for both models suggest a good fit to the data.

6.3.4 The southern oscillation index (SOI) effect on flood heights at the three sites

This subsection presents results for the inclusion of the southern oscillation index (SOI) term in the GEV model in order to investigate the effect of climate change at the three sites in the LLRB of Mozambique.

The SOI is usually used as a proxy for abnormal meteorological activities such as El Niño effect and La Niña effect (Coles, 2001; Reid, 2000). A link between annual maximum flood heights and SOI would be an indication of meteorological volatility due to the El Niño effect. The LLRB is a basin known to be characterised by the El Niño effect that often results in tropical cyclones such as Cyclone Eline, Cyclone Gloria, etc. that have affected the basin in the past causing very heavy rainfall leading to destruction of hydraulic and engineering structures, and loss of human lives in the basin (Jackson, 2013a,b; Musiya, 2013).

The SOI is considered in this chapter as a climate variable, other than the long-term time trend, that can also influence flood heights. Scatter plots for the annual maximum flood heights against the October SOI were plotted (Figures 6.1-6.3 overleaf) and the results of the plots in Figures 6.5 & 6.6 indicate a positive linear relationship between SOI and annual maximum flood heights at Chokwe and Combomune. There is no visual clear indication of a linear relationship between SOI and annual maximum flood heights at Sicacate.

Chokwe site

The scatter plot in Figure 6.1 showed that there is probabilistic linear relationship between annual maximum flood heights and the October SOI values. This

implies that the climate change may influence annual daily maximum flood heights at Chokwe. The relationship is further investigated using Table 6.4.

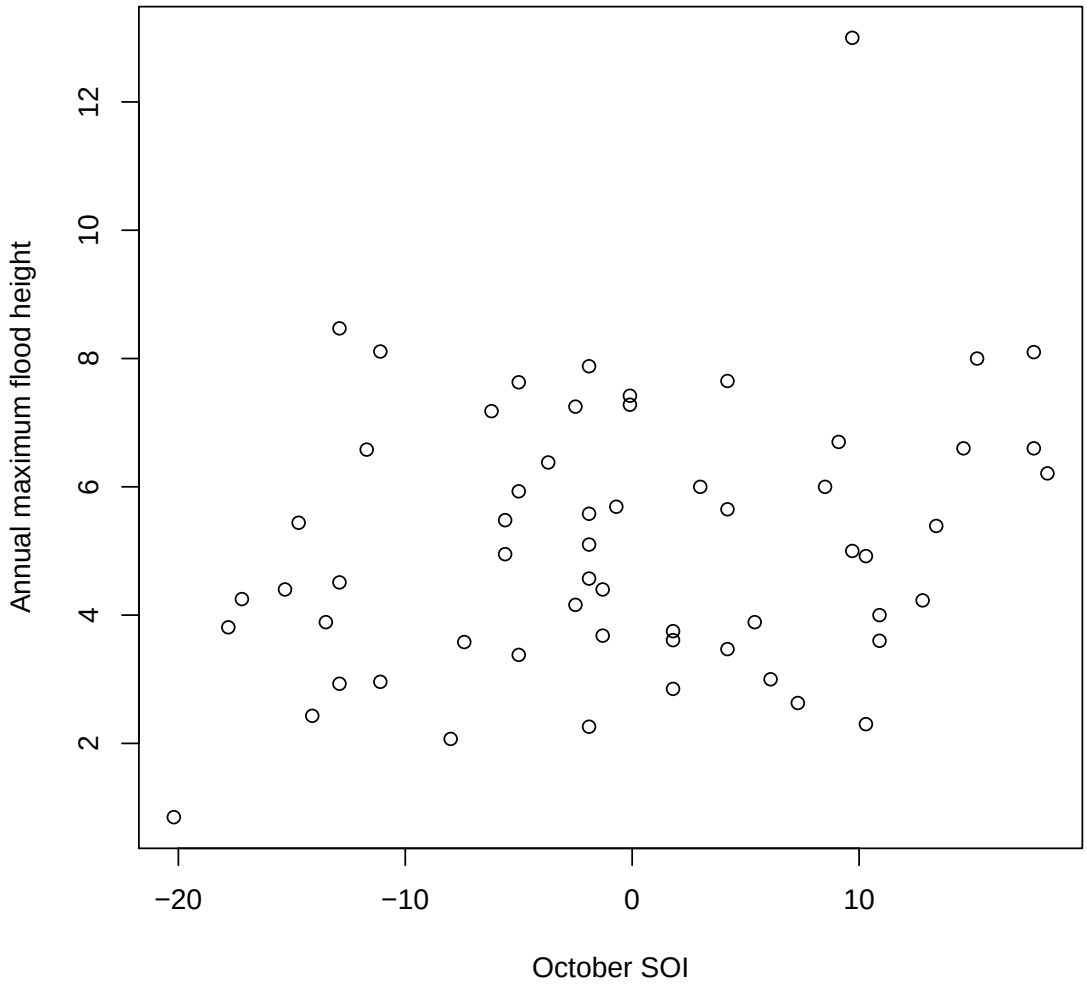


Figure 6.1: Scatter plot of annual maximum flood height and the southern oscillation index (SOI) at Chokwe

The model pair for Chokwe (M_0, M_5) in Table 6.4 has a critical value of $\chi^2_{1,0.05} = 3.841$ with a D statistic value of 8.502. The likelihood ratio test for $\mu_1 = 0$ has p-value $= 0.010$. This implies that the SOI term in location parameter is quite

significant (p-value < 0.05) and worthwhile ($D > 3.841$) for Chokwe. Thus, overall, model M_5 is significant and provides an improvement in fit over the time-homogeneous GEV model.

Table 6.4: Chokwe AM1 time-heterogeneous GEV models with SOI covariate included for the period 1951-2009.

Model	$\hat{\mu_0}$	$\hat{\mu_1}$	$\hat{\mu_2}$	$\hat{\sigma_0}$	$\hat{\sigma_1}$	$\hat{\xi}$	**NLLH**
M_0	4.248	0	0	1.785	-0.081	126.313	
M_5	4.284	0.053	0	1.690	-0.057	122.062	
M_6	4.256	0.053	0.001	1.689	-0.056	122.060	

The other model pair for Chokwe (M_0, M_6) in Table 6.4 has a critical value of $\chi^2_{2,0.05} = 5.991$ with a D statistic value of 8.506. The likelihood ratio test for $\mu_1 = 0$ has p-value $= 0.010$ and for $\mu_2 = 0$ has p-value $= 0.474$. This implies that the additional trend term in Model M_6 is not significant although the SOI term remains significant. Therefore model M_6 for Chokwe is worthwhile $(D > 5.991)$ but not significant in the additional trend term. In general, it can be concluded that model M_5 with a SOI term in the location parameter of the GEV distribution is a suitable model for Chokwe.

The general nonstationary GEV model for Chokwe with a SOI term is

$$G(\mu(t), \sigma, \xi; x_i) = \exp\left\{-\left(1 + \frac{-0.057\,(x_i - (4.284 + 0.053SOI))}{1.690}\right)_+^{\frac{1}{0.057}}\right\}, \quad (6.12)$$

where $x_{i,\forall i=1,2,\ldots,k}$ is the annual maximum flood height at the site.

The shape parameter (-0.057) for the prevailing GEV model with a SOI term in (6.12) is not significantly different from zero (p-value $= 0.265 > 0.05$) implying that it is not significantly negative. This suggests that the extreme floods at Chokwe hydrometric station follow a Gumbel distribution which is light-tailed,

hence consistent with time-homogeneous model in (6.8). The diagnostic plots for the time-heterogeneous GEV model in (6.12) are presented in Figure 6.8. The diagnostic plots in Figure 6.8 show that the model is of good fit.

Combomune site

Figure 6.2 presents a scatter plot between SOI and annual daily maximum flood heights. Results in Figure 6.2 showed that there is a probabilistic linear relationship between the annual daily maximum flood heights with considerable randomness. This relationship is further investigated using Table 6.5.

The model pair for Combomune (M_0, M_5) in Table 6.5 has a D statistic value of 6.738 with an associated critical value of $\chi^2_{1,0.05} = 3.841$. The likelihood ratio test for $\mu_1 = 0$ has p-value $= 0.029$. This implies that the SOI term in location parameter of the Combomune GEV model is significant (p-value < 0.05) and worthwhile ($D > 3.841$). Thus, overall, model M_5 is significant and provides an improvement in fit over the time-homogeneous GEV model for Chokwe.

Table 6.5: Combomune AM1 time-heterogeneous GEV models with SOI covariate included for the period 1966-2009.

Model	$\hat{\mu_0}$	$\hat{\mu_1}$	$\hat{\mu_2}$	$\hat{\sigma}$	$\hat{\xi}$	**NLLH**
M_0	5.163	0	0	1.660	-0.124	90.740
M_5	5.259	0.048	0	1.610	-0.124	87.371
M_6	5.472	0.048	-0.009	1.629	-0.145	87.294

The other model pair for Combomune (M_0, M_6) in Table 6.5 has a D statistic value of 6.892 with an associated critical value of $\chi^2_{2,0.05} = 5.991$. The likelihood ratio test for $\mu_1 = 0$ has p-value $= 0.031$ and for $\mu_2 = 0$ has p-value $= 0.350$. This implies that the additional trend term in Model M_6 is not significant although the SOI term remains significant in the model. Therefore model M_6 for

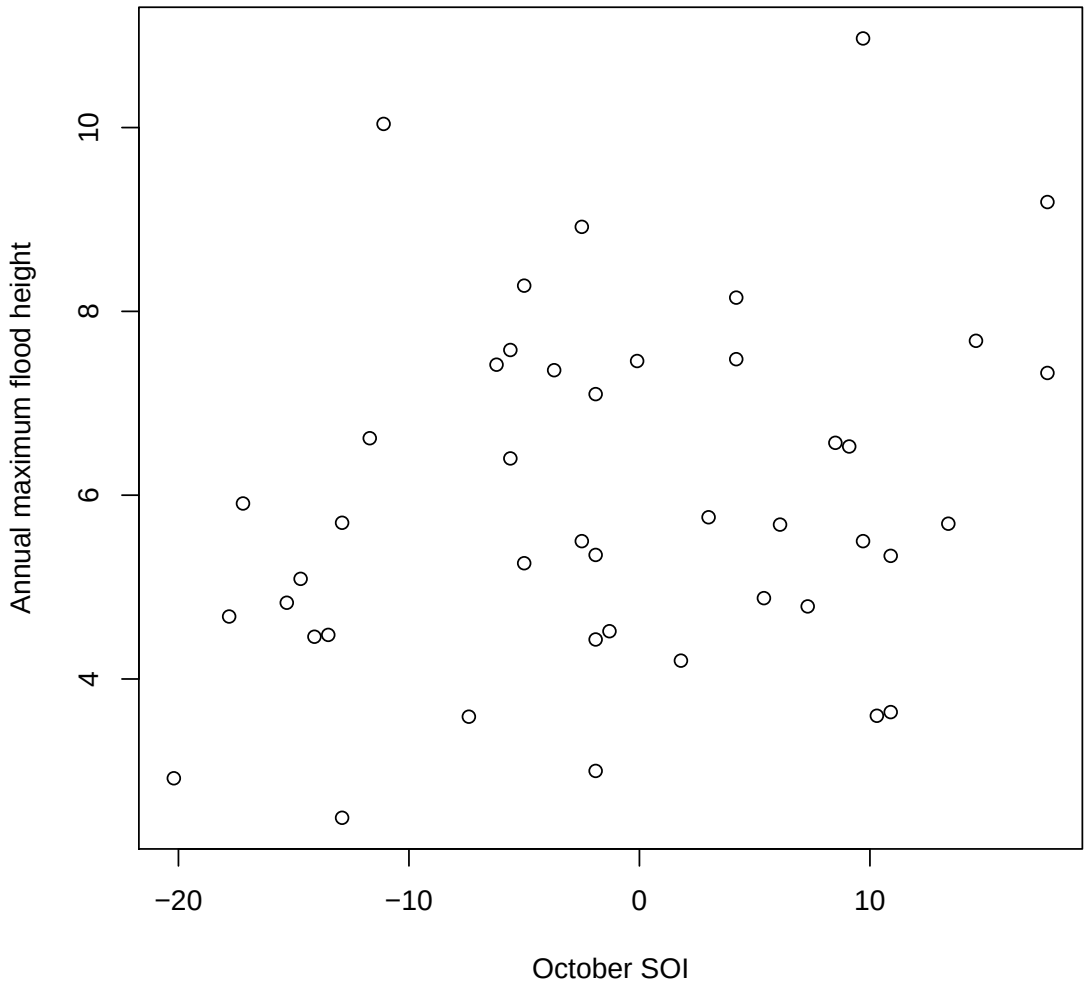


Figure 6.2: Scatter plot of annual maximum flood height and the southern oscillation index (SOI) at Combomune

Combomune is not significant in the additional trend although it is worthwhile $(D > 5.991)$. In general, it can be concluded that model M_5 with a SOI term in the location parameter of the GEV distribution is a suitable model for Combomune.

The general nonstationary GEV model for Combomune with a SOI term is

$$G(\mu(t),\sigma,\xi;x_i)=\exp\left\{-\left(1+\frac{-0.124\left(x_i-(5.259+0.048SOI)\right)}{1.610}\right)_+^{\frac{1}{0.124}}\right\}, \quad (6.13)$$

where $x_{i,\forall i=1,2,\ldots,k}$ is the annual maximum flood height at the site.

The shape parameter (-0.124) for the prevailing model in (6.13) is not significantly different from zero (p-value $= 0.1396 > 0.05$). This implies that the shape parameter (-0.124) is not significantly negative, suggesting that the extreme floods at Combomune hydrometric station follow the light-tailed Gumbel distribution family. The diagnostic plots for the time-heterogeneous model in (6.13) are presented in Figure 6.9. The residual probability plot suggests a good fit to the data.

Sicacate site

Figure 6.3 presents the scatter plot for the Sicacate site between SOI and annual daily maximum flood heights at the site. The scatter plot in Figure 6.3 showed that there is a no clear positive linear relationship between annual daily maximum flood heights and SOI. These results are investigated further using results in Table 6.6.

Based on Table 6.6 for Sicacate, the model pair (M_0, M_5) has a D statistic value of 7.570 with an associated critical value of $\chi^2_{1,0.05} = 3.841$. The likelihood ratio test for $\mu_1 = 0$ has p-value $= 0.020$. This implies that the SOI term in location parameter is significant (p-value < 0.05) and worthwhile ($D > 3.841$) for Sicacate. Thus, model M_5 is significant and provides an improvement in fit over the time-homogeneous GEV model.

The other model pair for Sicacate (M_0, M_6) in Table 6.6 has a D statistic value

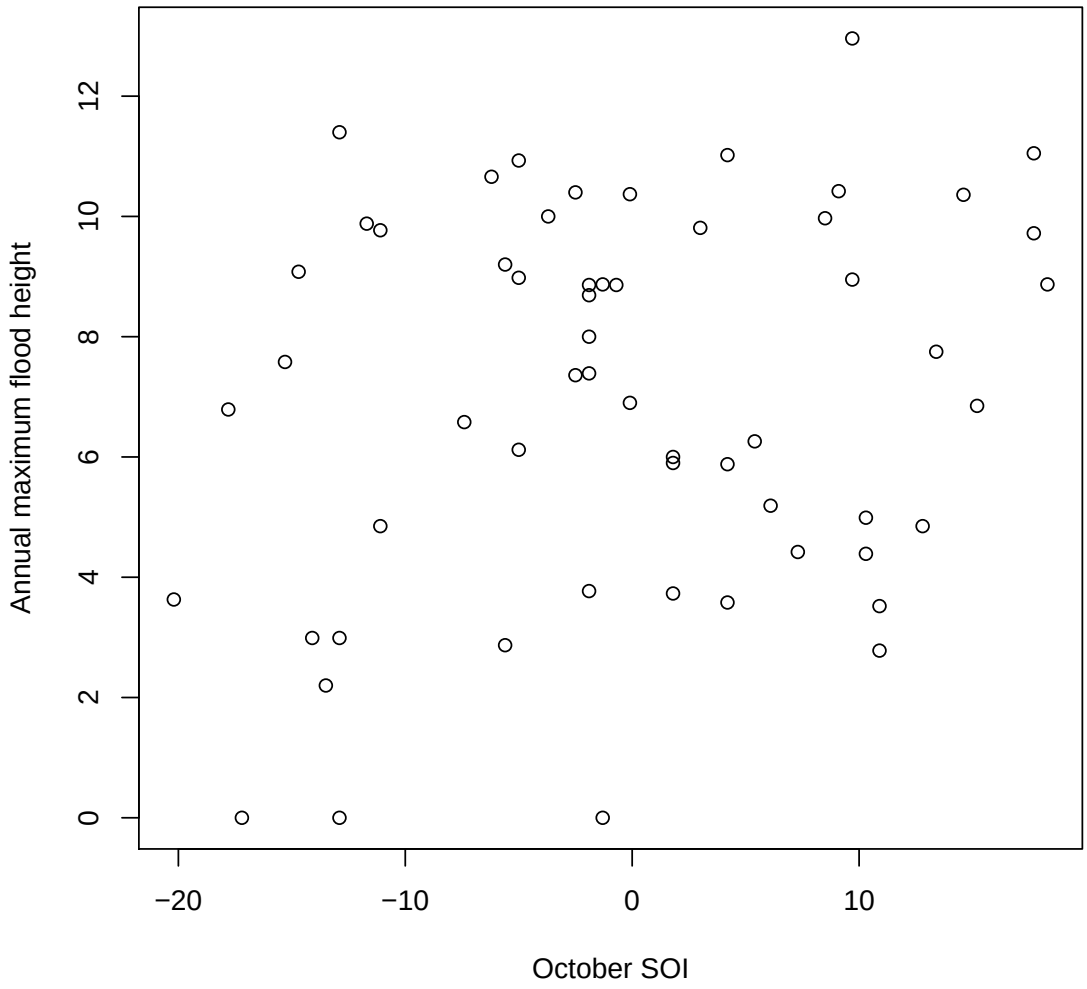


Figure 6.3: Scatter plot of annual maximum flood height and the southern oscillation index (SOI) at Sicacate

of 7.868 with an associated critical value of $\chi^2_{2,0.05} = 5.991$. The likelihood ratio test for $\mu_1 = 0$ has p-value $= 0.051$ and for $\mu_2 = 0$ has p-value $= 0.295$. This implies that the additional trend term in Model M_6 is quite insignificant although the SOI term remains significant in the model. Therefore model M_6 for Sicacate is not significant with respect to the additional trend term although it is worthwhile $(D > 5.991)$. In general, it can be concluded that model M_5 with a

Table 6.6: Sicacate AM1 time-heterogeneous GEV models with SOI covariate included for the period 1952-2009.

Model	$\hat{\mu_0}$	$\hat{\mu_1}$	$\hat{\mu_2}$	$\hat{\sigma}$	$\hat{\xi}$	**NLLH**
M_0	6.151	0	0	3.328	-0.454	148.547
M_5	6.244	0.062	0	3.307	-0.495	144.762
M_6	5.866	0.054	0.014	3.366	-0.529	144.613

SOI term in the location parameter of the GEV distribution is a suitable model for Sicacate.

The general nonstationary GEV model for Sicacate with a SOI term is

$$G(\mu(t), \sigma, \xi; x_i) = \exp\left\{-\left(1 + \frac{-0.495\left(x_i - (6.244 + 0.062SOI)\right)}{3.307}\right)_+^{\frac{1}{0.495}}\right\}, \quad (6.14)$$

where $x_{i,\forall i=1,2,\ldots,k}$ is the annual maximum flood height at the site.

The shape parameter in (6.14), that is, -0.495, is significantly different from zero (p-value < 0.0001) implying that the shape parameter for the model at Sicacate is significantly negative. This suggests that extreme floods at Sicacate hydrometric station can be modelled by a short-tailed Weibull distribution family. The diagnostic plots for the time-heterogeneous models in (6.14) are presented in Figure 6.10. The residual probability plots for the model suggest a good fit to the data.

6.4 Return level estimation

Once the models are built the interest turns to estimating the return levels and their corresponding return periods, that is, estimation of extreme flood heights and their corresponding return periods. In this thesis comprehensive return levels and their corresponding return periods are provided in Chapter 3 for

several time-homogeneous distributions for the three sites in the basin. This section presents the general equations for finding the return levels (high flood heights) of the models built in the previous sections.

The quantile function given in (6.5) can be re-written in terms of the return period such that the general equation for finding the high flood heights from a time-homogeneous GEV model becomes

$$X_T = G^{-1}(1 - T^{-1}) = \hat{\mu} + \frac{\hat{\sigma}}{\hat{\xi}} \left[\left(-\ln \left(1 - T^{-1} \right) \right)^{-\hat{\xi}} - 1 \right] \tag{6.15}$$

where T is the return period of the corresponding extreme flood height (return level), X_T, and $\hat{\mu}, \hat{\sigma}$ and $\hat{\xi}$ are the MLEs of the parameters μ, σ and ξ. For example, using (6.16), the 100-year flood height for Chokwe is given by

$$X_T = 4.248 - \frac{1.785}{0.081} \left[\left(-\ln \left(1 - 100^{-1} \right) \right)^{0.081} - 1 \right] = 11.10 \; m.$$

The general equation for finding the extreme flood heights from a nonstationary time-heterogeneous GEV model with a linear trend in the location and/or scale parameter is

$$X_T = (\hat{\mu}_0 + \hat{\mu}_1 t) + \frac{\exp\left(\hat{\sigma}_0 + \hat{\sigma}_1 t\right)}{\hat{\xi}} \left[\left(-\ln \left(1 - T^{-1} \right) \right)^{-\hat{\xi}} - 1 \right] \tag{6.16}$$

where $\hat{\mu}_0 + \hat{\mu}_1 t = \hat{\mu}$ if there is no linear trend in the location parameter and $\exp(\hat{\sigma}_0 + \hat{\sigma}_1 t) = \hat{\sigma}$ if there is no linear trend in the scale parameter. For example, using (6.17), the 100-year flood height for Sicacate for the year 2011 using (6.11) with a linear trend in the GEV scale parameter is

$$\begin{aligned} X_T &= 6.675 + \frac{\exp\left(1.966 + 0.55\left(2011 - 1952\right)\right)}{0.693} \left[\left(-\ln \left(1 - 100^{-1} \right) \right)^{0.693} - 1 \right] \\ &= 20.64 \; m. \end{aligned}$$

The general equation for finding the extreme flood heights from a nonstationary time-heterogeneous GEV model with a SOI term in the location parameter is

$$X_T = (\hat{\mu}_0 + \hat{\mu}_1 SOI) + \frac{\exp(\hat{\sigma}_0 + \hat{\sigma}_1 t)}{\hat{\xi}} \left[\left(-\ln\left(1 - T^{-1}\right)\right)^{-\hat{\xi}} - 1 \right] \tag{6.17}$$

Any flood height of interest can be calculated using (6.16) if the GEV is stationary, (6.17) if the GEV is nonstationary in time with a long-term trend in the location and/or scale parameter and (6.18) if the GEV is nonstationary with a SOI term in the location parameter.

6.4.1 General remarks on the results

The models developed at the three sites based on the existence of a linear trend in the scale and/or location parameters of the GEV distribution have comparable standard errors with those models developed using the existence of the SOI term in the GEV location parameter. However, the existence of a nonstationary GEV model with the combined effect of trend and SOI was not found to be significant in the study for this chapter.

The interesting findings are that while most studies in other regions have found a dominant linear trend in the location parameter of the GEV distribution for some rivers (Katz, 2010; Coles, 2001), this study has found no evidence of a significant linear trend in the location parameter of the GEV distribution for the LLRB of Mozambique. On the other hand, this study has revealed a dominant time dependent scale parameter for the river at Combomune upstream and Sicacate further downstream. The study also revealed evidence of a highly significant nonlinear quadratic term in the location parameter at Sicacate although the complexity of the overall model was not worthwhile with reference to the time-homogeneous GEV model.

A nonstationary GEV model with a SOI term in the location parameter was found to be a significant and worthwhile model at all the three sites. In other words, a nonstationary GEV model with a SOI term explains substantially more of the variability in the flood heights data series at the three sites than the simpler stationary GEV model. Thus, the ENSO effect or El Niño and La Niña effects, which are the descriptive measures of climate change, contribute substantially to variability of flood heights and frequency of floods and droughts in the LLRB of Mozambique. The SOI models developed in this chapter are comparable to the trend models developed earlier in the chapter and these models can be used as alternatives since the combined effect of both trend and SOI is not significant and not relatively worthwhile.

The findings in this chapter are in full support of a previous study by Aich et al. (2014) who used a geoscientific model called eco-hydrological SWIM model to compare the climate change impacts on streamflow in four large African river basins including the Limpopo, and found the Limpopo basin to be highly sensitive to climate change variability. The results obtained in this chapter, complemented by those of Aich et al. (2014), explain the reason for the increased frequency of extreme floods in the LLRB of Mozambique which can be attributed to the variability in climatic conditions.

6.5 Added value and importance of the study in this chapter

This study is significant in a number of ways. To begin with, the lower Limpopo River basin of Mozambique is an area in Southern Africa that has not been deeply studied, and yet it has reasonable quality data. In particular, studies on the application of statistics of extremes in the river basin are scarce,

let alone statistics of extremes in a changing climate. The lower Limpopo River basin suffers from extreme climate conditions, punctuated with frequent droughts and floods. This study showed the importance of incorporating climate change factors such as trend and SOI to the time-heterogeneous GEV model in the basin, rather than sticking to the time-homogeneous models in this era of global warming and a changing climate. This study complements and advances the geoscientific work that was recently done by Aich et al. (2014) in the basin which showed a strong impact of climate variability in the basin. The models developed in this study took into account this dominant climate variability in the basin to explain the frequency of floods which is currently on the rise possibly due to the La Niña effect. The models developed in this study will be communicated to the National Directorate of Water in Mozambique so that they can be recommended or considered in its disaster reduction efforts in the basin. It is hoped that this study will help reduce the associated risk and mitigate the deleterious impact of floods on humans and property in the basin. The Limpopo River basin is home to other countries in the SADC region, hence lesson for them to be learned from the models developed in the LLRB of Mozambique.

6.6 Concluding remarks

The study considered the use of statistics of extremes in a changing climate for the lower Limpopo River basin of Mozambique. Three hydrometric stations representing three sites along the lower Limpopo River were considered for the study. The maximum likelihood estimation method was used to estimate the parameters of the GEV distribution in the presence of a long-term trend covariate and an indicator of meteorological volatility known as SOI. The study has revealed the importance of considering nonstationary linear and nonlinear trend models, as well as SOI models, when using statistics of extremes in

a changing climate as these models provide a substantial improvement in fit over the time-homogeneous models. This improvement in fit is crucial for the planning and policy-making of the government of Mozambique and its partners in the lower Limpopo River basin, where the largest irrigation scheme of the country is situated. The importance of the developed models is attributed to the fact that these nonstationary models take into account the reasons for increased frequency of floods in the basin. Once the government and its partners are fully aware of the reasons behind the increased frequency of floods in the basin their planning can be much improved.

This study has successfully identified the prevailing models at the three sites such that Chokwe is the only site with a time-homogeneous GEV model. This can be attributed to the fact that Chokwe is the only station with an irrigation scheme and hence some of the water at the site is diverted to the Chokwe Irrigation Scheme for irrigation purposes. The other two sites Combomune and Sicacate have a prevailing time-heterogeneous GEV model with a dominant linear trend in the scale parameter. The site of Sicacate has an alternative nonstationary model with a linear trend in both the location and scale parameters of a GEV distribution. The prevailing trend models established in this study are consistent with cumulative (or moving sums) annual maximum series flood flows and therefore appear reliable to use for flood frequency analysis in the basin. The use of the identified time-dependent GEV models with a trend in the scale parameter in the basin would also reduce the sensitivity of the frequency of floods which is known to vary with changes in the scale parameter, and therefore lead to more reliable estimates in the frequency of floods.

Alternative models in the basin were also successfully identified through using SOI, an indicator of meteorological volatility. The SOI models substantially explained the effects of a changing climate in the variability of annual daily

maximum flood heights and increased frequency of floods in the basin. These models will indeed be of great value to the water management in Mozambique.

Future studies will attempt to advance this study to consider Bayesian MCMC methods in a changing climate for the lower Limpopo River basin of Mozambique. With more flood heights data sites available in the region or more rainfall data from several rain gauging stations across several catchments in the basin an attempt can be made to consider spatial extremes in the presence of these covariates in future studies involving statistics of extremes in a changing climate.

6.7 Summary of the chapter

In this chapter a nonstationary GEV distribution, with a trend covariate and an indicator of meteorological volatility variable, was fitted to annual daily maximum flood heights at three sites: Chokwe, Combomune and Sicacate in the lower Limpopo River basin of Mozambique. A GEV distribution was fitted to annual daily maximum flood heights and its moving sums. Nonstationary time-dependent GEV models with a linear trend in location and scale parameters were considered in the study for this chapter. The results showed lack of sufficient evidence to indicate a linear trend in the location parameter at Chokwe and Combomune sites. On the other hand, the findings in this chapter revealed strong evidence of the existence of a linear trend in the scale parameter at Combomune and Sicacate, while the scale parameter had no significant linear trend at Chokwe. Further investigation regarding the trend in this study revealed that the location parameter at Sicacate could be modelled by a nonlinear quadratic trend; however, the complexity of the overall model was not worthwhile in fit over a time-homogeneous model.

An alternative modelling approach involved fitting a GEV model with a SOI term in the location parameter of the model. In this study, the nonstationary GEV model with a SOI term was found to substantially account for more of the variability in the flood heights data than the stationary GEV model. No sufficient evidence of the existence of a combined effect of both trend and SOI was found in the study. The SOI models developed in this chapter can be used as alternative models to the trend models since their combined effect could not be established.

The study for this chapter showed the importance of extending the stationary GEV model to incorporate climate change factors such as trend and SOI in the lower Limpopo River basin of Mozambique, particularly in this era of a changing climate and suspected global warming.

APPENDIX 6.1: R PROGRAMMING IN ISMEV PACKAGE (SELECTED)

R program for Chokwe site with the SOI covariate included

```
Regg4=Chokwe.AM.SOI.2015
attach(regg4)
head(regg4)
tail(regg4)
summary(regg4)

# Creating a window
win.graph()
par(mfrow = c(1,1)) # graph window with 1 row and 1 column

install.packages('ismev')
library(ismev)

# Modelling with SOI # Creating variables for SOI
ti=matrix(ncol=1,nrow=59)
ti[,1]=seq(1,59,1)
covar=matrix(ncol=2,nrow=59)
covar[,1]=regg4[,3] #Stores the SOI values in column 1 of matrix named covar
covar[,2]=ti #Stores the time indicator in column 2 of covar

# fitting a GEV model with covariate SOI only, model:μ(t) = μ0 + μ1SOI(t)
soi =gev.fit(regg4[,2],ydat=covar,mul=1)

# fitting a GEV model with covariates SOI and time t, model: μ(t) = μ0 +
μ1SOI(t) + μ2t
```

```
Soi= gev.fit(regg4[,2],ydat=covar,mul=c(1,2))

# fitting the stationary GEV model with no covariates
soi = gev.fit(regg4[,2])

gev.diag(soi) # diagnostic plots
```

Output for Chokwe site

```
>ti=matrix(ncol=1,nrow=44)
> ti[,1]=seq(1,44,1)
> covar=matrix(ncol=2,nrow=44)
> covar[,1]=regg4[,3] #Stores the SOI values in column 1 of covar
> covar[,2]=ti

> soi=gev.fit(regg4[,2],ydat=covar,mul=1)
model
model[[1]] [1] 1

model[[2]]
NULL
model[[3]]
NULL

link [1] "c(identity, identity, identity)"

conv [1] 0

nllh [1] 122.0618
```

mle [1] 4.28366989 0.05306004 1.68974792 -0.05657706

se [1] 0.24665440 0.02232374 0.17704067 0.08991068

> gev.diag(soi) # diagnostic plots

> soi=gev.fit(regg4[,2],ydat=covar,mul=c(1,2))

model

model[[1]] [1] 1 2

model[[2]]

NULL

model[[3]]

NULL

link [1] "c(identity, identity, identity)"

conv [1] 0

nllh [1] 122.0599

mle [1] 4.2562210490 0.0531114448 0.0008740923 1.6887483029 -0.0561617964

se [1] 0.48374681 0.02231430 0.01354799 0.17753870 0.09065817

> gev.diag(soi)

APPENDIX 6.2: STATIONARY AND NONSTATIONARY GEV MODELS DIAGNOSTIC PLOTS

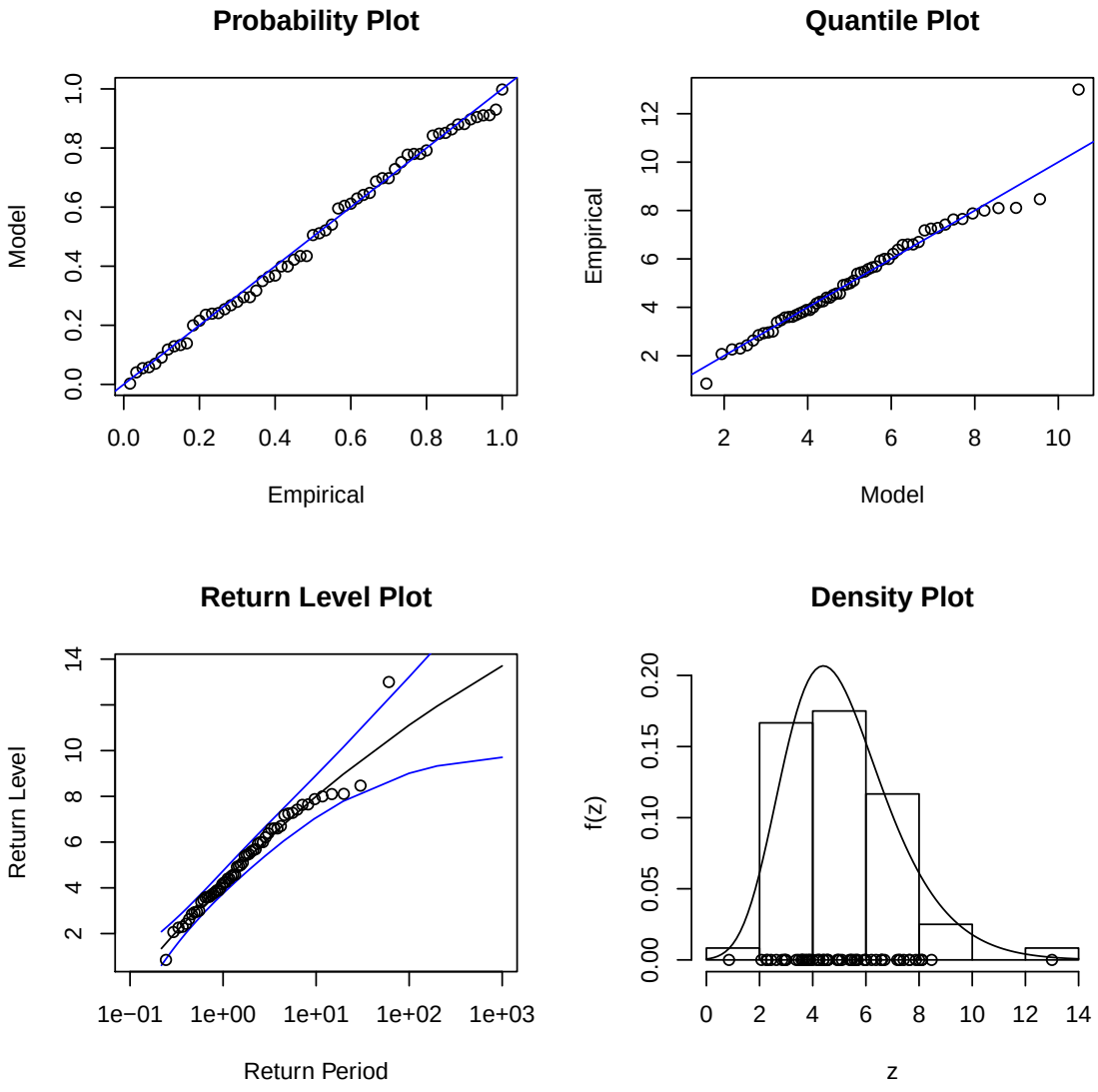


Figure 6.4: Diagnostic plots for the time-homogeneous GEV best fitting model at Chokwe hydrometric station

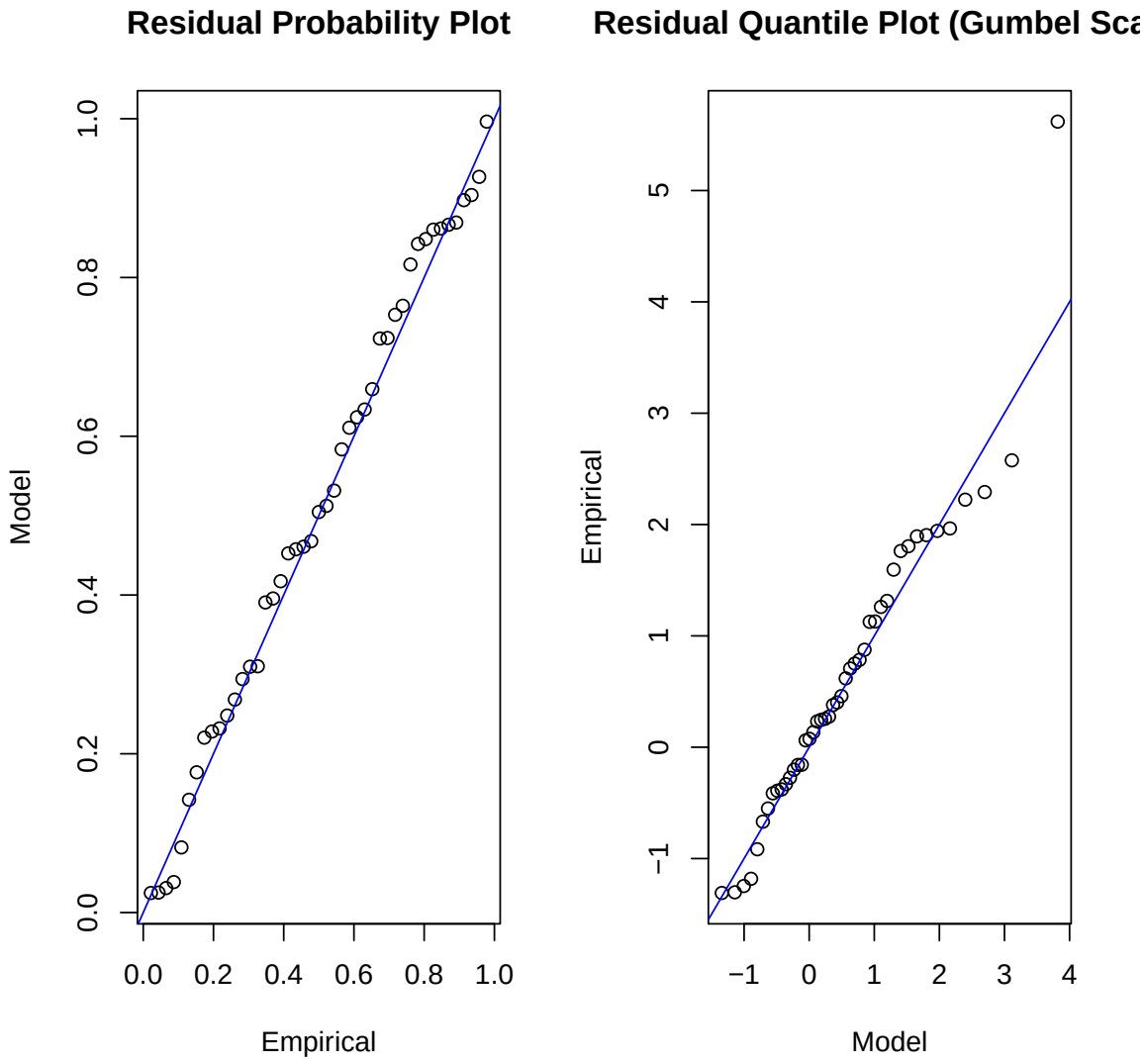


Figure 6.5: Diagnostic plots for the time-heterogeneous GEV best fitting model (with a trend term in the scale parameter) at Combomune hydrometric station

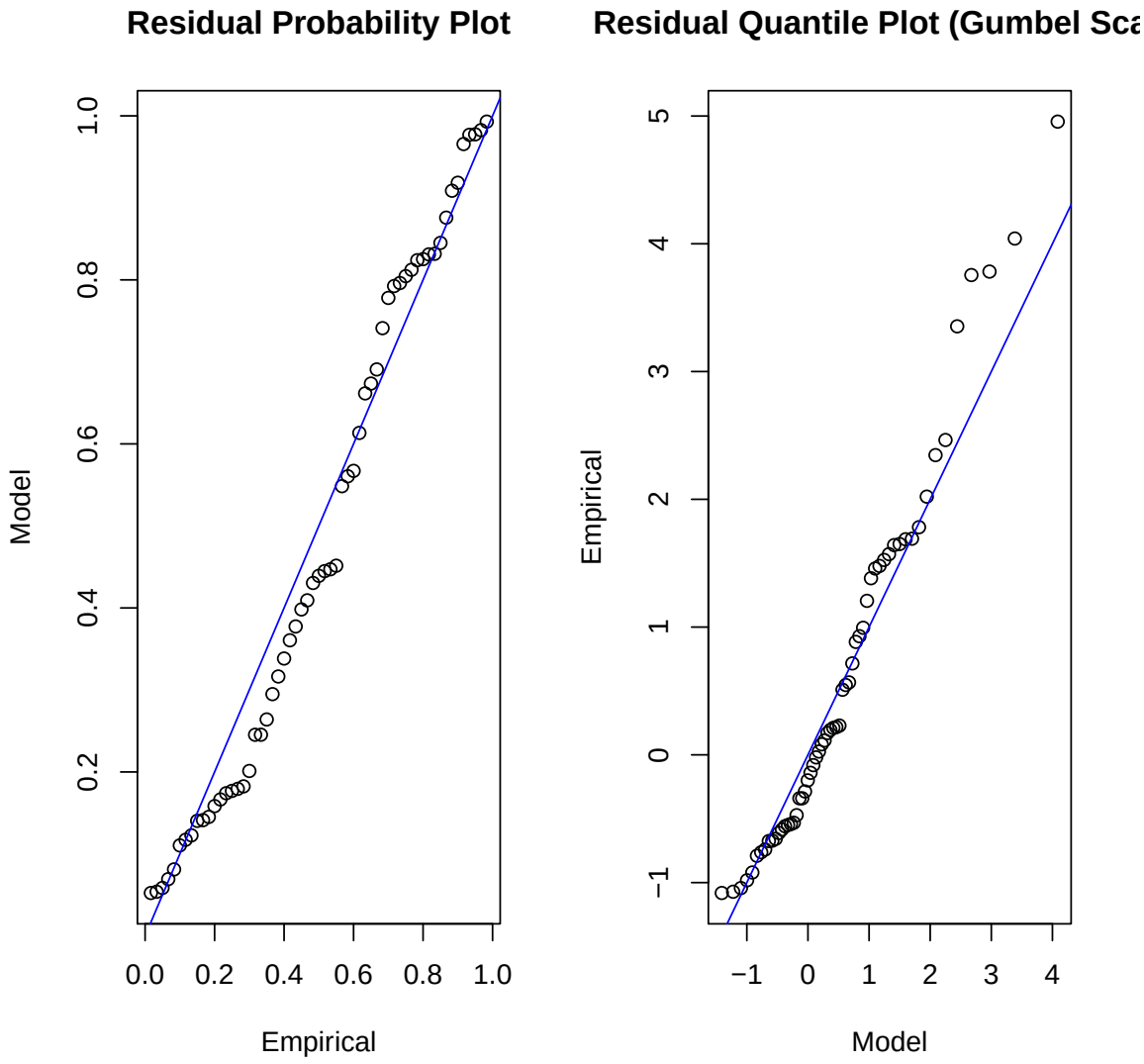


Figure 6.6: Diagnostic plots for the time-heterogeneous GEV best fitting model (with a trend term in the scale parameter) at Sicacate hydrometric station

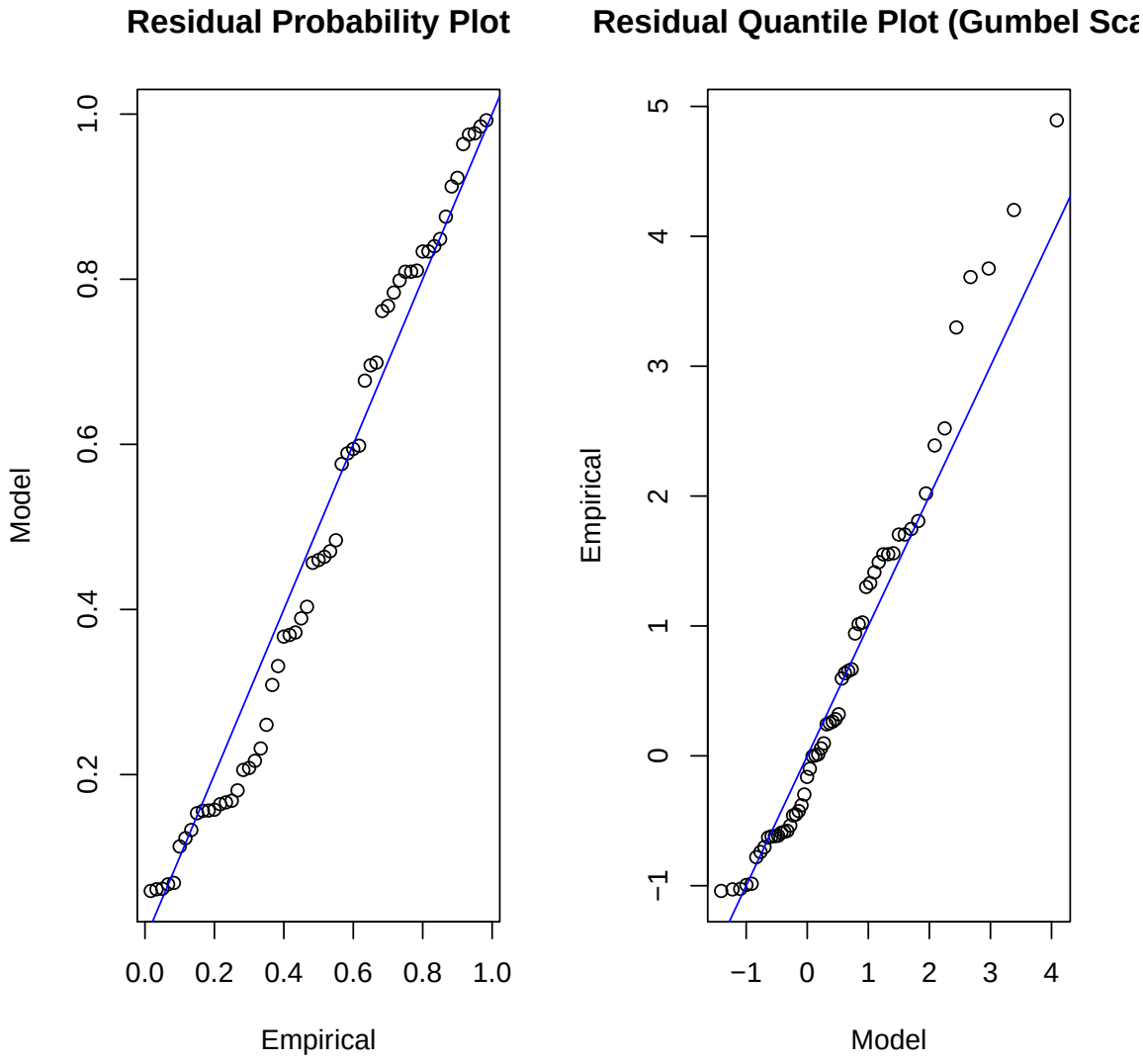


Figure 6.7: Diagnostic plots for the time-heterogeneous GEV best fitting model (with a trend term in both location & scale parameters) at Sicacate hydrometric station

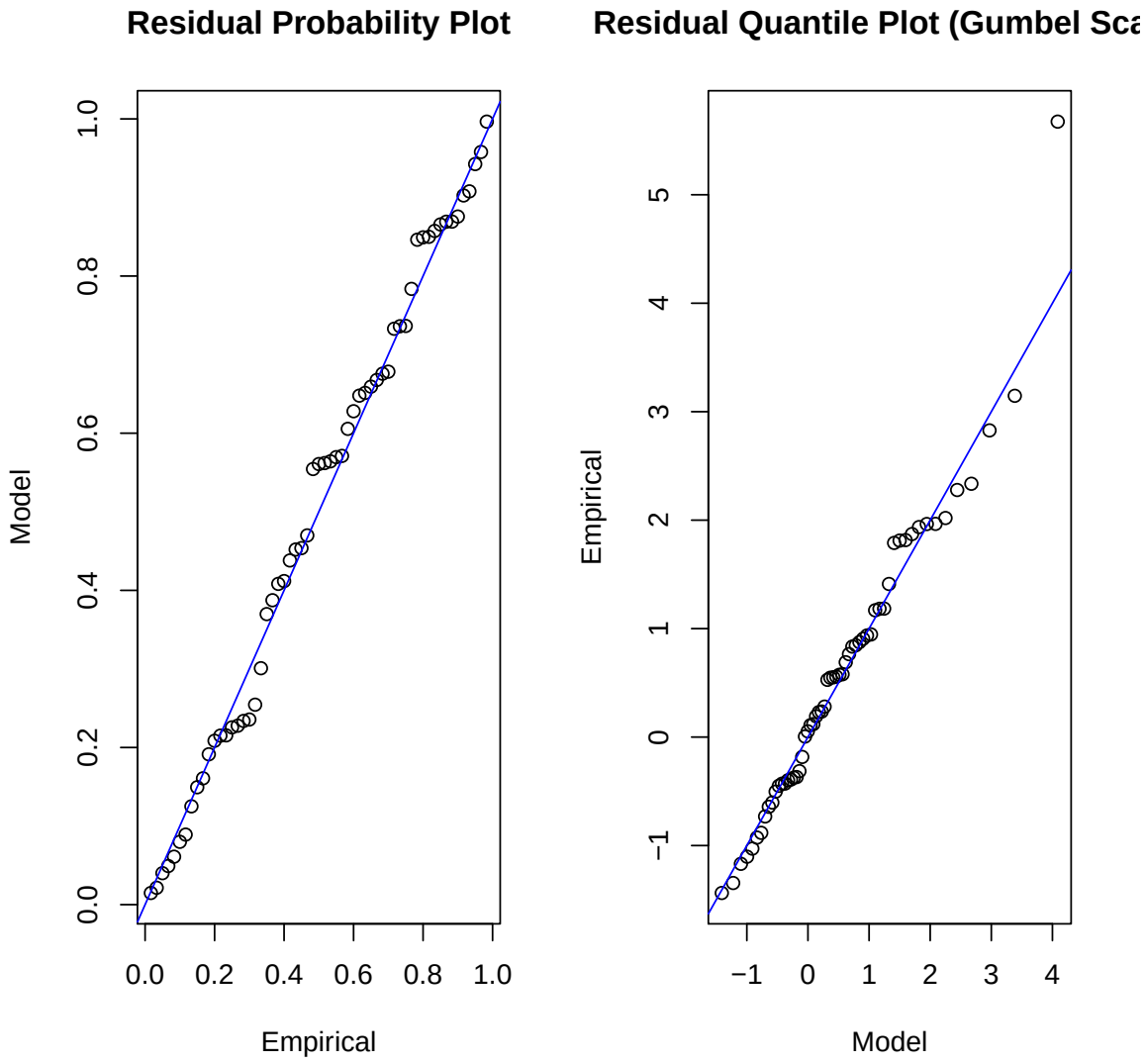


Figure 6.8: Diagnostic plots for the time-heterogeneous GEV best fitting model (with a SOI term in location parameter) at Chokwe hydrometric station

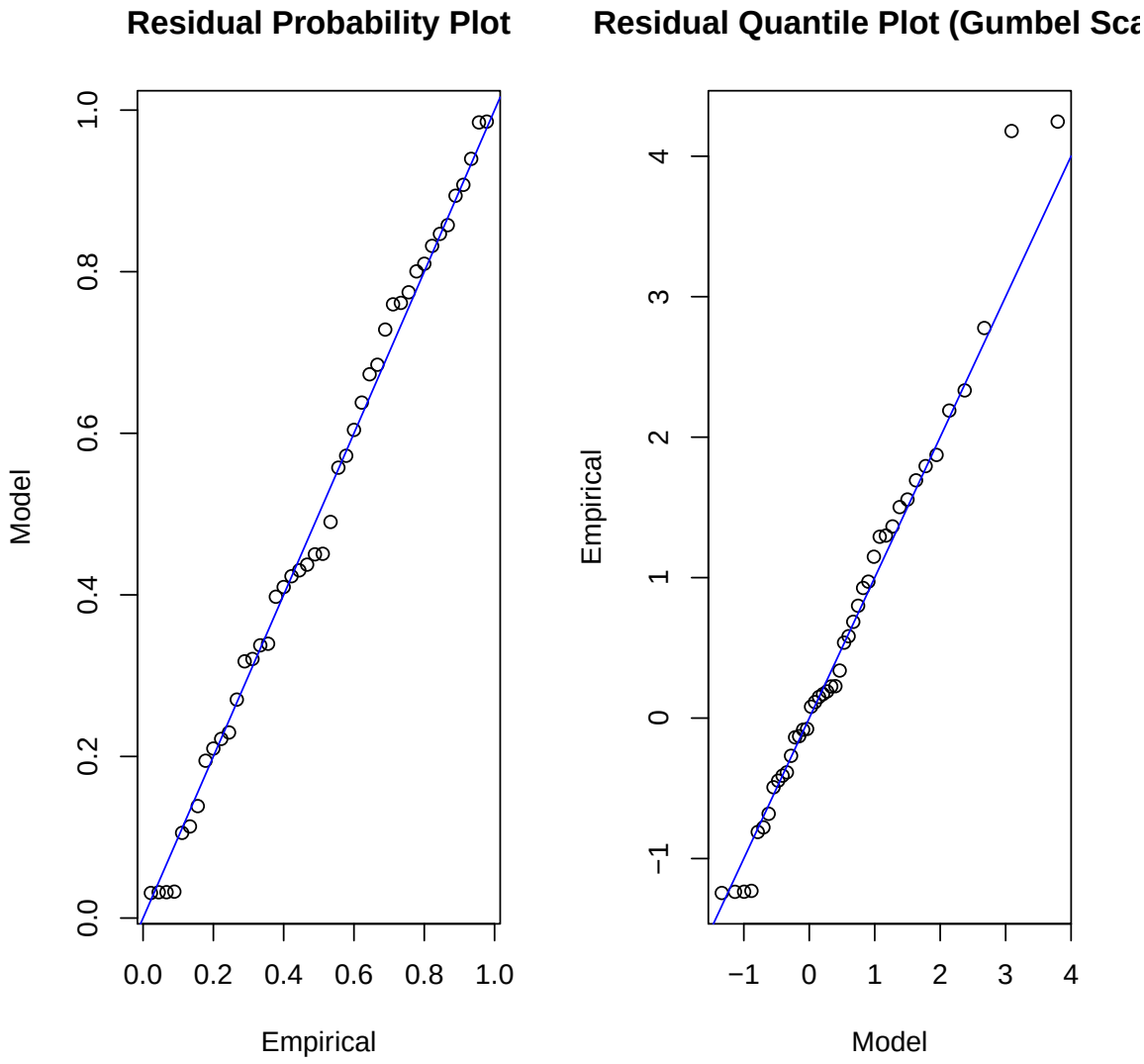


Figure 6.9: Diagnostic plots for the time-heterogeneous GEV best fitting model (with a SOI term in location parameter) at Combomune hydrometric station

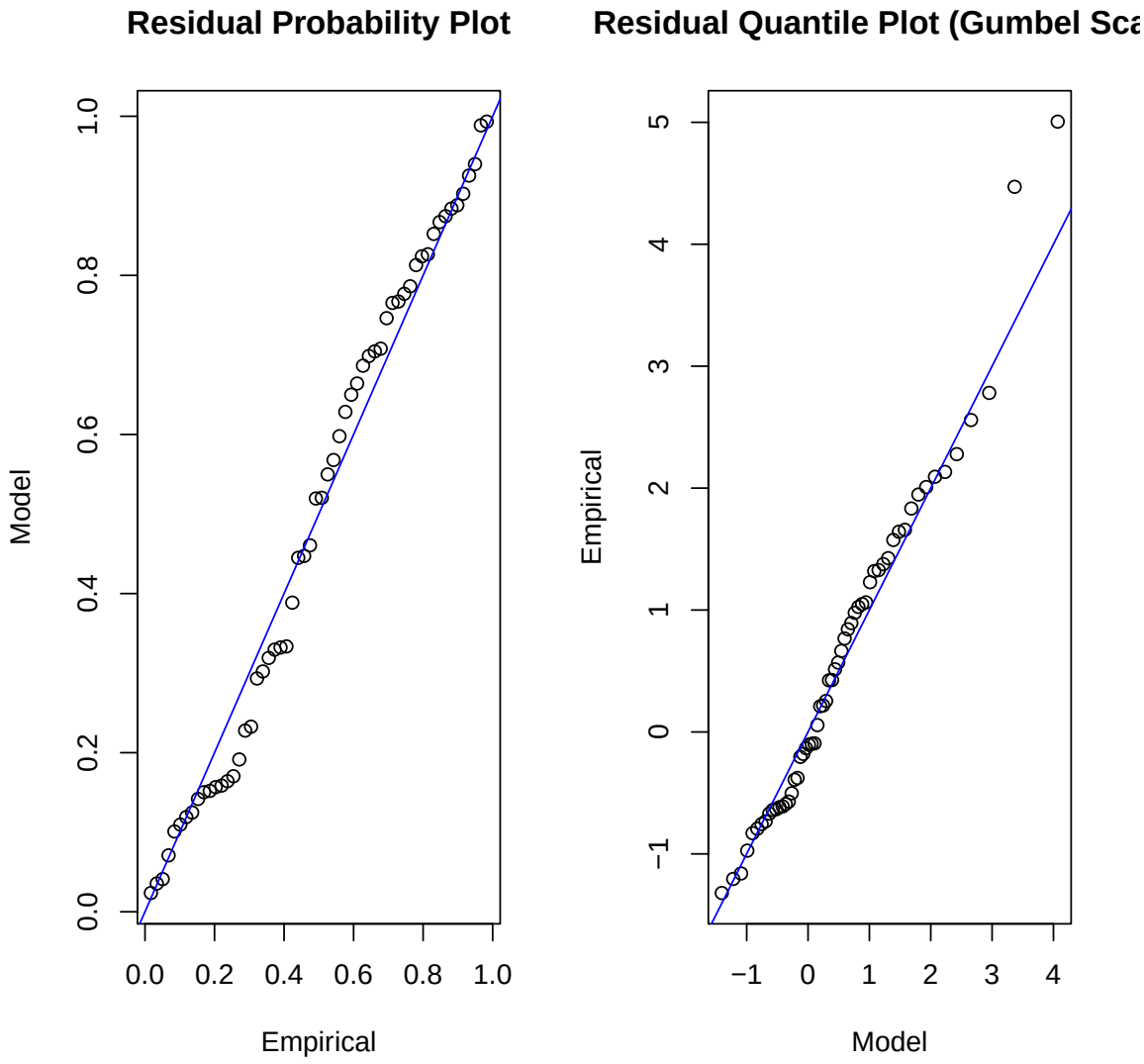


Figure 6.10: Diagnostic plots for the time-heterogeneous GEV best fitting model (with a SOI term in location parameter) at Sicacate hydrometric station

Chapter 7

Modelling extreme flood heights in the lower Limpopo River basin of Mozambique using a time-heterogeneous generalised Pareto distribution

7.1 Introduction

The presence of long-term trends attributed to climate variability in extreme events such as annual maximum river flow or precipitation data has become an active area of research interest for hydrologists and climatologists in the 21^{st} century in order to investigate climate change scenarios and improve research on climate impact on weather extremes (Saidi et al., 2013; Velasco et al., 2013;

Towler et al., 2010). These trends result in nonstationary processes which have the characteristic to systematically change with time (Beirlant et al., 2004; Coles, 2001). Extreme value theory is the branch of statistics used extensively to study very low or very high values in the tail of some distribution. However, when the process is nonstationary the usual extreme value limit models are not applicable (Coles and Davison, 2008). The usual procedure when dealing with nonstationary processes is to adopt a pragmatic approach of using the standard extreme value models as basic templates that can be enhanced by statistical modelling (Coles and Davison, 2008).

It is argued that although the atmosphere-ocean general circulation models from an Intergovernmental Panel on Climate Change assessment report (IPCC AR4) were projecting increases in intense precipitation and flooding at a large spatial scale, the models are limited in terms of their ability to quantify extreme events at regional and at-site scales which are crucial for decision making (Towler et al., 2010). The use of statistical techniques in river flood heights or large scale precipitation extremes associated with climate change has been limited (Towler et al., 2010, with references therein).

According to Velasco et al. (2013) some researchers have established that global change related extreme events are expected to be on the rise all over Europe although there is no general agreement in concluding that the frequency and magnitude of floods have increased due to climate related changes. The problem in lack of a general agreement is attributed to shortage of instrumental data records in many regions of the continent of Europe. In the view of the researcher these findings and problems encountered in Europe should draw similarities with climate related extreme events in Africa, particularly in Southern Africa where the study area in this chapter is situated although the climatic conditions are very different. In general, several studies in Europe project in-

creases in floods in some regions, for example, in Catalonia the northeast of Spain (Velasco et al., 2013).

In a separate study based on climate simulations, Saidi et al. (2013) argued that a warmer climate could increase the proportion of floods. It is generally accepted that the anticipated climatic changes are associated with a higher frequency of occurrence of extreme floods but not associated with a higher intensity of extreme floods (Saidi et al., 2013).

According to Velasco et al. (2013) with further reference to a review by IPCC (2012) "there are already a few studies regarding flood magnitude and occurrence changes at river basin level". However scientific literature on these climatic change related extremes is mainly focused on the continent of Europe, North America and the UK (Velasco et al., 2013, with references therein). In Australia Verdon-Kidd and Kiem (2015) reiterated that climate change has been a subject of concern over the past 15 years, particularly in Australia. In another climate change perspective, Vanem (2015b) performed an extreme value analysis of wave data in order to investigate the relationship between uncertainties due to extreme value analysis and variability in climate. This author found a discernable shift towards higher extremes in a changing future wave climate.

Given that the climatic conditions in these areas (rest of Europe, North America and the UK) differ strongly from the current climatic conditions in Southern Africa, and that future projections may also be very different, this study in this chapter aims to assess and develop climate related models for the future flood trends in the lower Limpopo River basin (LLRB) of Mozambique, an area in Southern Africa that has not been deeply studied. Recently Aich et al. (2014) studied the impact of climate change on streamflow in four large African river

basins including Limpopo River basin (LRB) using a geo-scientific model and found that the LLRB is highly affected by climate variability. Our statistical approach will complement the work by Aich et al. (2014) through developing statistical climate change models which will help in decision making for the basin.

According to Ferreira and de Haan (2015) there are two fundamental approaches widely used in statistics of extremes namely peaks-over-threshold (POT) (or partial duration series) and block maxima (also called annual maximum series). The block maxima approach in extreme value theory (EVT) consists of dividing the observation period into blocks (non-overlapping periods of equal size) and restricts attention to the maximum observation in each block (naturally years in the case of floods). The new observations (maxima series) follow approximately the generalised extreme value (GEV) distribution (Ferreira and de Haan, 2015). In the POT approach in EVT, one chooses a reasonably high threshold and selects those observations from the initial data series that exceed the predetermined threshold. The probability distribution of the new observations (exceedances) follows approximately a generalised Pareto distribution (GPD) (Ferreira and de Haan, 2015, 2014). The exact conditions under which the POT statistical method is justified are described by a second order term (de Haan and Ferreira, 2006, Section 2.3). In the case of block maxima approach it is generally accepted that the maxima follow very well an extreme value distribution and Ferreira and de Haan (2015) give more theoretical details on the exact conditions of the block maxima statistical method.

It is argued that the POT method makes better use of the available information since it retains all 'relevant' high observations whereas the block maxima method on one hand misses some of these high observations and, on the other hand, might retain some lower observations (the latter 'hand' might also be

the block maxima merit over POT as presented by Laurens de Haan at the Extreme Value Analysis (EVA2013) conference, 8-12 July 2013, Shanghai, China) (Ferreira and de Haan, 2015; Cunnane, 1973).

The relevant merits of block maxima and POT are discussed in detail in Ferreira and de Haan (2015) with references to several papers based on simulated data. Among the merits are that POT estimates are better than block maxima estimates if the number of exceedances is larger than 1.65 times the number of blocks for the Gumbel family of distributions $(\xi = 0)$ when using maximum likelihood estimators (MLEs). The POT is as efficient as block maxima for high quantiles using probability weighted moments (PWM) or L-moments parameter estimators. Provided the number of exceedances is larger than the number of blocks, POT is more preferable for fat-tailed distributions (Fréchet type), whereas block maxima is more efficient for short-tailed distributions (Weibull type). When using historical data, the gains with the block maxima are in the range of the gains with the POT method based on MLEs (Ferreira and de Haan, 2015, with references therein). Based on simulation studies, the POT samples with an average of two or more observations above the threshold per block have more accurate estimates than the corresponding block maxima estimates and the accuracies become similar and rather good with more than 200 years of historical data (Ferreira and de Haan, 2015, with references therein).

All these studies on the merits of POT, some with mixed views, show in general that the POT is more efficient than the block maxima in many ways provided that the number of exceedances is greater than the number of blocks. The two methods have comparable performances when the sample sizes are large. However, the theoretical comparison performed in Ferreira and de Haan (2015) showed that block maxima is more efficient with lower asymptotic variances of both extreme value index and quantile estimators compared with POT. The

minimal mean square error is also lower for block maxima under normal circumstances (Ferreira and de Haan, 2015).

The present study uses the POT method in order to utilise the richness of information from the big data records contained at the three sites in the LLRB of Mozambique. The block maxima method was considered in Chapter 6. The parameters of the GPD in this chapter are estimated by the MLEs method.

The outline of the rest of the chapter is organised in the following structure. Section 7.2 presents the research methodology, Section 7.3 presents the results and discussion of the findings, Section 7.4 gives the general remarks of the results, Section 7.5 discusses the added value and significance of the study, Section 7.6 gives the concluding remarks and Section 7.7 summarises the chapter. Finally Appendix 7.1 presents selected R code programs used in modelling the data for the chapter, while Appendix 7.2 presents some figures and diagnostic plots for the chapter.

7.2 Research methodology

This section presents the study sites and the data used in the study, a brief probability framework of the POT approach including the extension of time-homogeneous GPD model to linear trend models. Analytic relationships of the link between POT and block maxima methods are presented through mathematical proofs of the relationship between the GEV distribution and GPD using the results in literature (Reiss and Thomas, 2007; Coles, 2001).

7.2.1 Study sites and data

The data used in this study was obtained from the Mozambique National Directorate of Water (DNA), the authority responsible for water management in

Mozambique in the Ministry of Public Works and Housing. The data are hydrometric daily flood heights (in metres) recorded at Chokwe (1951-2010), Combomune (1966-2010) and Sicacate (1952-2010) hydrometric stations for the LLRB of Mozambique as presented in Figures 1.1-1.3 in Chapter 1. The three sites are such that Combomune is located in the upper part of the LLRB about 162 km from the border with South Africa and Zimbabwe, Chokwe is located in the middle of the basin about 130 km downstream of Combomune and Sicacate is further downstream of Chokwe in the lower part of the basin on way to the Indian Ocean. The daily flood heights at each site were recorded three times a day, i.e. morning, afternoon and evening periods. There was a number of missing values in-between the years at each site but this number was counterbalanced by the fact that the data was recorded three times a day making the number of missing values negligible. Further scrutiny on the data reveals that most of the missing values occurred in years of severe droughts or during the winter season which is usually characterised by very low rainfall. Since the interest of this study is in higher values above a certain high threshold, it would mean that these missing values were still likely going to be below the threshold and therefore irrelevant in the study.

7.2.2 Theoretical overview of peaks-over-threshold and generalised Pareto distribution (GPD)

The approach used in this study is POT. The POT method considers only those of the initial observations that exceed a pre-specified high threshold (Ferreira and de Haan, 2015, 2014). In a more formal approach, let $X = X_1, ..., X_n$ be iid random variables representing flood heights. If F is a distribution function (commonly unknown) of the flood heights X, then the conditional excess $(X-u)$ distribution function is given by (2.22) in Chapter 2 and the approximated GPD model is given in (2.23).

More details on Pareto type models and their tests for GoF can be found in Berning (2010) and Beirlant et al. (2006). In EVT the GPD plays the role of a natural distribution model for excesses over a reasonably high threshold. In other words, the GPD plays the same role for POT as the GEV plays for block maxima.

Some key points in extreme value theory are that if block maxima series have a GEV distribution then POT series have an associated GPD. Additionally, the shape parameter ξ in the GPD exactly equals that of the corresponding GEV distribution (Fischer and Schumann, 2014; Magadia, 2010; Coles, 2001). The following theorem gives an analytic relationship between the GEV distribution and GPD (Reiss and Thomas, 2007).

Theorem 7.1. *(Existence of analytic relationship between GPD and GEV distribution) (Reiss and Thomas, 2007). Suppose $H_\xi(y)$ is a GPD, $G_\xi(x)$ is a GEV distribution and the excesses $y = x - u$, then*

$$H_\xi(y) = 1 + \ln G_\xi(x), \quad \textit{for } \ln G_\xi(x) > -1. \tag{7.1}$$

The outline proof of Theorem 7.1 follows. Consider the right-hand side of the theorem for $\xi \neq 0$

$$\begin{aligned} & 1 + \ln \left[\exp \left\{ - \left[1 + \xi \left(\frac{x-u}{\sigma} \right) \right]^{\frac{-1}{\xi}} \right\} \right] \\ = \; & 1 - \left[1 + \xi \left(\frac{x-u}{\sigma} \right) \right]^{\frac{-1}{\xi}} \\ = \; & 1 - \left[1 + \xi \left(\frac{y}{\sigma} \right) \right]^{\frac{-1}{\xi}} \\ = \; & H_\xi(y) \end{aligned} \tag{7.2}$$

Thus the theoretical relationship between GPD and GEV distribution has been proved, which implies the connection between POT and block maxima. It will

be interesting to find out whether the practical results at the three sites in this chapter will show a connection with those of the corresponding sites based on block maxima in Chapter 6.

7.2.3 Threshold selection

Two threshold selection techniques were used in this study namely: mean excess life plot and threshold choice plot. The mean excess life plot is an exploratory technique carried out prior to model selection while threshold choice plot is based on an assessment of the stability of parameter estimates through fitting of models (GPD in this study) across a range of different thresholds (Beirlant et al., 2004; Coles, 2001).

The choice of a threshold is critical to any POT analysis. It is based on a trade-off between bias and variance. Too high a threshold would discard too much data and generate a few exceedances leading to high variance of the estimate of the parameters. On the other hand, too low a threshold would necessitate using data that are no longer considered as being in the tails of the distribution and this will violate the asymptotic basis of the model, thereby leading to an increase in bias (Magadia, 2010; Coles, 2001). The standard practice is to choose as low a threshold as possible provided the limit model gives a reasonable approximation. The bias-variance trade-off principle is based on choosing a low enough threshold value to have sufficient data to estimate the parameters σ and ξ, high enough threshold value for the asymptotic theorem to be considered accurate (Coles and Davison, 2008; Coles, 2001).

Mean excess life plot or mean residual plot

Let $x_{(1)} < x_{(2)} < ... < x_{(n_u)}$ be the exceedances $(x_i : x_i > u)$ that are obtained from the researcher's sample and define threshold excesses by $y_{(j)} = x_{(j)} - u$,

for $j = 1, ..., n_u$, then the empirical mean excess life plot (or mean residual plot) is defined by the points

$$\left\{(u, e_{n_u}(u)) : u < x_{(n_u)}\right\}, \tag{7.3}$$

where n_u is the number of observations that exceed u, $x_{(n_u)}$ is the largest value of X_i, and

$$e_{(n_u)} = \frac{1}{n}\sum_{i=1}^{n_u}\left(x_{(i)} - u\right). \tag{7.4}$$

The mean residual life plot should be approximately linear above a threshold u_0 at which the GPD provides a valid approximation to the excess distribution (Magadia, 2010; Coles, 2001). The linearity of the empirical mean excess life plot forms the basis of deciding a threshold (Magadia, 2010). The interpretation of the mean residual life plot in practical situations is not always an easy task (Coles, 2001) and this complexity can be eased by complementing the mean residual life plot with other plots such as the threshold choice, L-moment, dispersion plots and Pareto quantile plots (Magadia, 2010; Beirlant et al., 2004).

Threshold choice plot or stability plot

The threshold choice plot is based on the result that if $X \sim GPD(u_0, \sigma_0, \xi_0)$, then let u_1 be another threshold such that $u_1 > u_0$. Then $X|X > u_1$ is also another GPD with updated parameters $\sigma_1 = \sigma_0 + \xi_0\left(u_1 - u_0\right)$ and $\xi_1 = \xi_0$. Threshold choice plots are given by the points defined by

$$\left\{(u_1, \sigma_*) : u_1 < x_{(n_u)}\right\} \quad \text{and} \quad \left\{(u_1, \xi_1) : u_1 < x_{(n_u)}\right\}, \tag{7.5}$$

where $\sigma_* = \sigma_1 - \xi_1 u_1$. Thus estimates of σ_* and ξ are constant for all $u_1 > u_0$ if u_0 is a suitable threshold for the excesses to follow a GPD (Magadia, 2010). The

estimates of σ_* and ξ will not be exactly constant, but instead, approximately constant due to sampling variability (Coles, 2001).

The L-moment, dispersion plots and Pareto quantile plots will not be considered in this study. For more details on Pareto quantile plots the reader is referred to Berning (2010) and Beirlant et al. (2004), and more details on L-moment and dispersion plots are found in Southworth and Heffernan (2013) and Magadia (2010). A brief description of the Pareto quantile plot is given in the succeeding subsection.

Pareto quantile plot

Beirlant et al. (2004) defines a Pareto quantile plot as a scatter plot of the points: $(-\ln(1-p_i), \ln x_i)$ where $i = 1, ..., n$. According to Beirlant et al. (2004) numerous choices of p_i are employed in literature. The commonly used choice of p_i is $p_i = \frac{i}{n+1}$. The threshold in Pareto quantile plot is considered to be the observation on the vertical or y-axis at which the plot begins to follow a linear pattern.

7.2.4 Declustering

One of the shortcomings of the POT method compared to block maxima is that it is prone to producing dependent data particularly when dealing with time series data which are usually known to be dependent (Yilmaz et al., 2014; Southworth and Heffernan, 2013; Ferro and Segers, 2003). Time series data are known to be strongly auto-correlated hence a naïve selection of exceedances above a given threshold may lead to events that are no longer considered independent, but dependent (Yilmaz et al., 2014; Hamidieh, 2008; Ribatet, 2006; Ferro and Segers, 2003). In order to deal with this problem of clustering of neighbouring events a technique called declustering is used to achieve independence in cases where the original data are dependent (Hamidieh, 2008; Ri-

batet, 2006). In Ribatet (2006) a function called *clust* in R software package is used to identify exceedances over a fixed threshold while meeting the independence criteria using two arguments: the threshold and a time condition code named *tim.cond*.

In Ribatet (2006) clusters are identified using the following procedure.

1. The first exceedance initiates the first cluster;
2. The first observation below the threshold ends the cluster unless tim.cond does not hold;
3. The next exceedance which holds tim.cond initiates a new cluster;
4. The process is iterated as necessary.

In all declustering procedures, the main purpose is to identify cluster maxima. Variations usually appear in the choice of the time condition. In Ribatet (2006) two flood events are considered to be independent if they do not lie within a window period of 8 days. In Yilmaz et al. (2014) and Jakob et al. (2011) two flood events are considered to be independent if they are a day (24 hours) apart, that is, the POT values a day prior to and after the peak rainfall event are removed from the data set. For example, if a peak rainfall value (or flood height) in a cluster is selected for 5 December 2015 then rainfall values over the threshold on 4 December 2015 and 6 December 2015 are not considered in the POT data set.

In Ferro and Segers (2003) the Ferro-Segers declustering involves the estimation of the extremal index, γ, and the method proposes an automatic selection of the run-length auxiliary parameter, r, used to identify independent clusters. In Ferro and Segers (2003) the exceedance times consist of two groups: one corresponding to inter-cluster times and the other corresponding to intra-cluster

times. Based on asymptotic theory Ferro and Segers (2003) postulates that the $1-\gamma$ proportion of the smallest interexceedance times belong to the intra-cluster times, and the rest belong to the inter-cluster times. Therefore given m sorted interexceedance times, one can take the $(\lfloor m\gamma \rfloor + 1)^{th}$ interexceedance time as the smallest interexceedance time that separates the clusters. Declustering proceeds with $r = \lfloor m\gamma \rfloor + 1$. Concerning the study in this chapter, the researcher uses the declustering based on Ferro and Segers (2003) and programmed in R software package by Southworth and Heffernan (2013). The declustering results for the three sites are presented in Figures 7.2, 7.4 & 7.6 for Chokwe, Combomune and Sicacate, respectively.

7.2.5 Parameter estimation

The MLE method is used for estimation of the parameters of the GPD. Let $\boldsymbol{\alpha} = (\xi, \sigma)$ then ML of $\boldsymbol{\alpha}$ is given by

$$\begin{aligned} L(\boldsymbol{\alpha}, \mathbf{y}) &= \prod_{i=1}^{N_u} H_\xi(y_i) \\ &= \prod_{i=1}^{N_u} \frac{1}{\sigma}\left[1+\xi\left(\frac{y_i}{\sigma}\right)\right]^{-\left(1+\frac{1}{\xi}\right)} \end{aligned} \tag{7.6}$$

where N_u is the number of observations above the threshold u. The log-likelihood is

$$\ell(\boldsymbol{\alpha}, \mathbf{y}) = -n\ln\sigma - \left(1+\frac{1}{\xi}\right)\sum_{i=1}^{N_u}\ln\left[1+\frac{\xi(y_i)}{\sigma}\right] \tag{7.7}$$

7.2.6 GPD general models

Consider the GPD model in (7.2) for $\xi \neq 0$, that is

$$H(y) = 1 - \left(1+\frac{\xi}{\sigma}y\right)_+^{\frac{-1}{\xi}} \tag{7.8}$$

Let the model in (7.8) be M_0. In the present study, the researcher proposes one more model, M_1. The model M_1 has a linear trend in the scale parameter such that $\ln \sigma(t) = \sigma_0 + \sigma_1 t$ and $\xi(t) = \xi$. Hence model M_1 and its log-likelihood are of the form $G\left(\sigma(t), \xi; \mathbf{y}, t\right)$ and $l\left(\sigma_0, \sigma_1, \xi; \mathbf{y}, t\right)$, respectively. In its general form, the nonstationary model, M_1, is given by (7.9)

$$G\left(\sigma(t), \xi; y, t\right) = 1 - \left(1 + \frac{\xi y}{\exp\left(\sigma_0 + \sigma_1 t\right)}\right)_+^{\frac{-1}{\xi}}, \quad \text{for } \xi \neq 0,\ 0 \leq y \leq x_F - u. \quad (7.9)$$

7.2.7 Model choice

Like in the preceding chapter, an important question to answer is whether the nonstationary model is valid, i.e. is it worthwhile to have the nonstationary model? This is equivalent to testing whether the nonstationary model provides an improvement in fit over the time-homogeneous (usually simpler) model M_0. The MLE estimation of nested models uses a simple procedure called the deviance (D) statistic to compare one model against the other (Coles, 2001). In this study the time-homogeneous GPD model, M_0, is a special case of the time-dependent model M_1. Consider $M_0 \subset M_1$, then the D statistic is

$$D = 2\left\{l_1\left(M_1\right) - l_0\left(M_0\right)\right\},$$

where $l_1(M_1)$ and $l_0(M_0)$ are maximised negative log-likelihood (NLLH) for the GPD models M_1 and M_0, respectively (Coles, 2001). Like in the previous chapter, $D \sim \chi^2_{k,\alpha}$, where the degrees of freedom k is the difference in dimensionality of M_1 and M_0. A value of $D > \chi^2_{k,\alpha}$ suggests that M_1 explains substantially more of the variability in the data than M_0.

7.3 Results and discussion

This section presents the results of the study as well as discussing the results. The results in this section were obtained using R statistical programming package and R Studio (Southworth and Heffernan, 2013). Results for the time-homogeneous GPD model (M_0) and time-dependent GPD model (M_1) are presented in Tables 7.1, 7.2 and 7.3 for Chokwe, Combomune and Sicacate, respectively. The MLEs method was used to estimate the parameters of all GPD models.

7.3.1 Chokwe models

The time series plot for Chokwe showed that, with the exception of a very rare extreme 13 m flood height, the majority of flood heights at Chokwe were below 9 m (Figure 1.1, Chapter 1). The mean residual plot and the threshold choice plots (Figure 7.1) were used to come up with a reasonably high threshold of 4.8 m for Chokwe hydrometric station. The threshold of 4.8 m was chosen in order to meet the requirements of the bias-variance threshold trade-off balance such that it was high enough for the asymptotic theorem to be considered accurate and low enough to have sufficient data to estimate the GPD parameters.

Since the exceedances above the threshold could not be assumed to be independent from each other, declustering of the cluster maxima was performed and the results are presented in Figure 7.2 for Chokwe cluster maxima.

The shape parameter ξ was significantly different from zero (p-value < 0.0001) for both the time-homogeneous and nonstationary GPD models (Table 7.1), suggesting that the distribution of exceedances over the 4.8 m threshold at Chokwe was short-tailed (negative Weibull) and did not come from a Gumbel (exponential) distribution family. The D statistic value for model pair (M_0, M_1) in Table

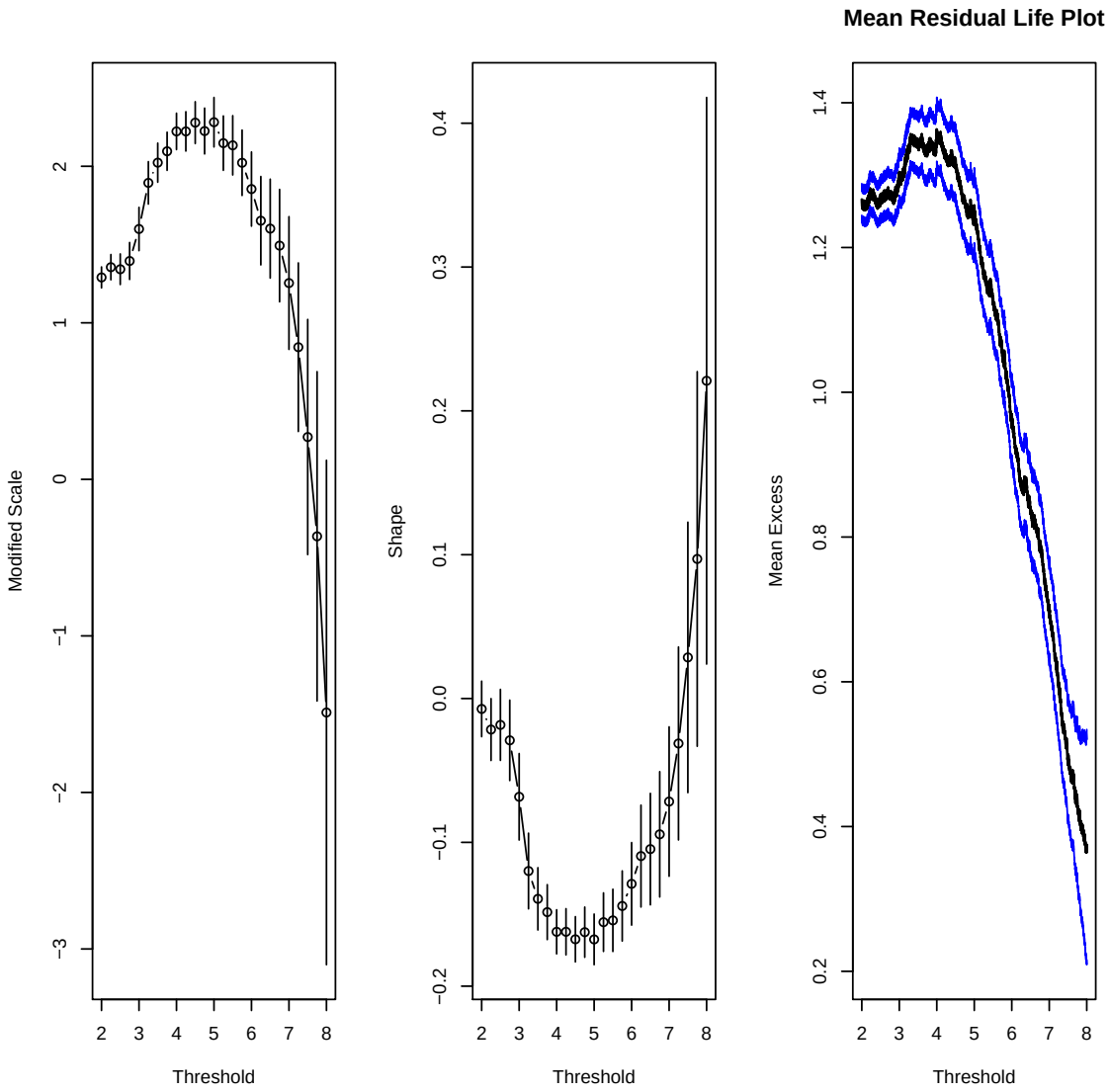


Figure 7.1: Chokwe threshold selection (from left to right): Panel (a). First two plots: Threshold choice plots or parameter stability plots; Panel (b). Mean residual life plot for the daily flood height data at Chokwe. Both panels for Chokwe show MLE estimates and 95% confidence intervals for the transformed parameters in GPD model

7.1 was 2(1803.637-1794.366) = 18.542 and the critical value for the pair was $\chi^2_{1,0.05} = 3.841$. The likelihood ratio test from the NLLH values in Table 7.1 for the test of $\sigma_1 = 0$ was highly significant at 5% level of significance (t-ratio $= 3.067$, p-value < 0.005). These results showed that the nonstationary GPD

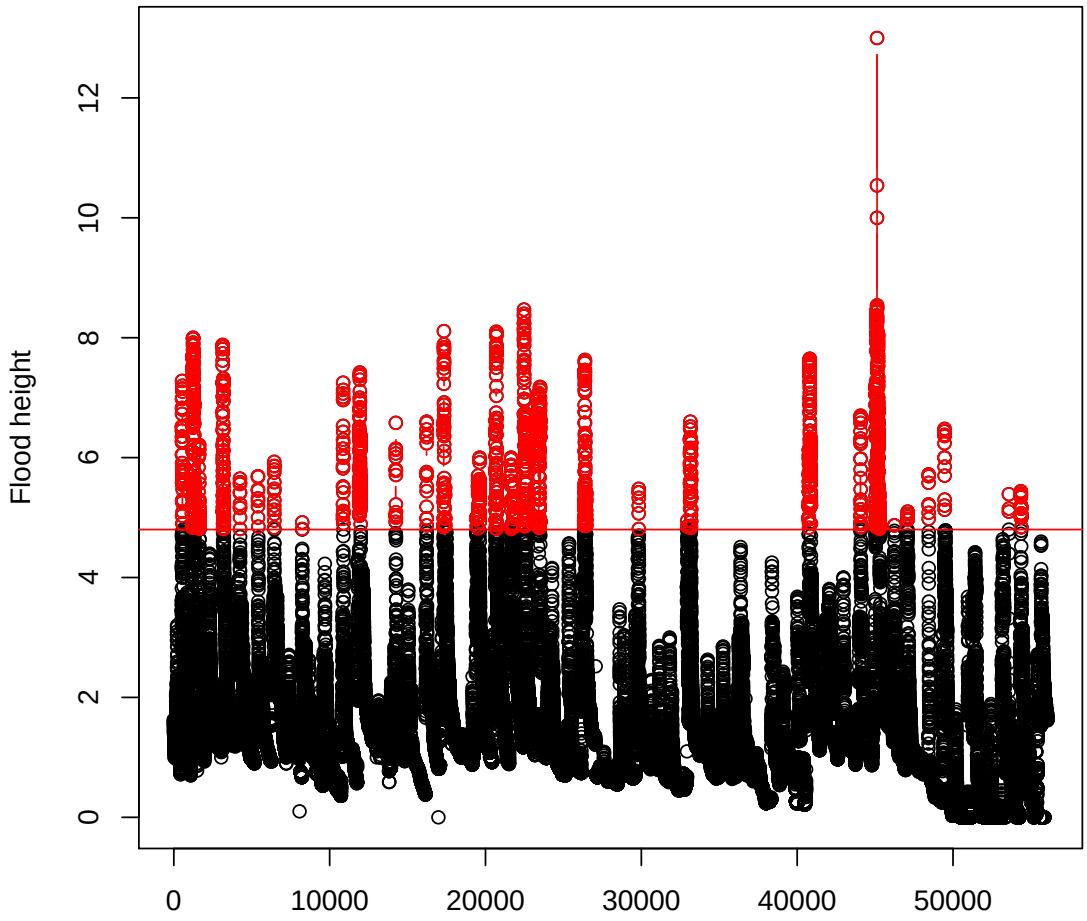


Figure 7.2: Chokwe declustered flood heights showing cluster maxima above 4.8 m threshold

model, M_1 was both significant and worthwhile over the time-homogeneous GPD model, M_0 in fitting the daily flood heights at Chokwe. These findings suggested that the nonstationary GPD model with a linear trend in the scale parameter provided an improvement in fit to the daily flood heights at Chokwe over the time-homogeneous GPD model since the D statistic value of 18.542 (> 3.841) was significantly large. The residual probability plot for the nonsta-

tionary GPD model suggested a good fit to the data at Chokwe (Figure 7.7).

Table 7.1: Parameter estimates and negative log-likelihood of the GPD models for Chokwe (1951-2010).

Model	$\hat{\sigma_0}$	$\hat{\sigma_1}$	$\hat{\xi}$	**NLLH**
M_0	1.4478677	0	-0.1628891	1803.637
M_1	0.2528215	0.0000059	-0.1940787	1794.366

On the contrary, however, the residual quantile plot showed a poor fit towards high quantiles indicating that extremely high flood heights at Chokwe may not be adequately modelled by the nonstationary model. The failure of the nonstationary GPD to model extremely high quantiles (flood heights) at Chokwe such as the 13 m flood height of the year 2000 was also highlighted in the time-homogeneous GPD model at the site (Figure 7.7). The return level plot based on the time-homogeneous GPD model for Chokwe revealed that the 13 m flood height which occurred in the year 2000 had a return period, on average, in excess of 1000 years (Figure 7.7) which appeared to be a ridiculously high return period, and much higher than that obtained in Chapter 6. In general, the residual diagnostic plots (Figures 7.7 and 7.10) suggest that the GPD models, both stationary and nonstationary, may not be quite suitable to model extreme floods at Chokwe implying that an alternative distribution such as the GEV in Chapter 6 may be necessary.

Despite the shortcomings in the GPD models for Chokwe presented in this study, there is strong evidence that the nonstationary GPD model outperforms the time-homogeneous GPD model. Therefore, based on the results of this study, the proposed model for Chokwe is the nonstationary GPD model with a linear trend in the scale parameter.

The general nonstationary GPD model for Chokwe is given in (7.10)

$$G\left(\sigma(t),\xi;y_t,t\right) \;=\; 1-\left(1+\frac{-0.1940787y_t}{\exp\left(0.2528215+0.0000059t\right)}\right)^{\frac{1}{0.1940787}} \qquad (7.10)$$

where $y_t = x_t - 4.8$ are the exceedances over the 4.8 threshold, x_t is cluster maxima of daily flood heights, and $t = 1, 2, ..., 56058, ...$, where 56058 is the time of the last observed flood height value over the period 19 June 1951 to 31 August 2010. Note also that the daily flood height data were recorded three times a day, i.e. morning, afternoon and evening meaning that each day had three values of t. This however, does not pose a problem in the independence of data since declustering is performed in the analysis.

7.3.2 Combomune models

The time series plot for Combomune shows that the majority of flood heights are below 10 metres except for a few rare extreme flood heights of about 11 m (Figure 1.2, Chapter 1). The mean residual plot and the threshold choice plots (Figure 7.3) were used to come up with a reasonably high threshold of 5.8 m for Combomune hydrometric station which was chosen in such a way that it was high enough for the asymptotic theorem to be considered accurate and low enough to have sufficient data to estimate the GPD parameters.

Since the exceedances above the threshold, 5.8 m, for Combomune could not be assumed to be independent from each other, declustering of the cluster maxima was performed and the results are presented in Figure 7.4 for Combomune cluster maxima.

The shape parameter ξ was significantly different from zero (p-value < 0.0001) for both the time-homogeneous and nonstationary GPD models (Table 7.2), suggesting that the distribution of exceedances over the 5.8 m threshold at

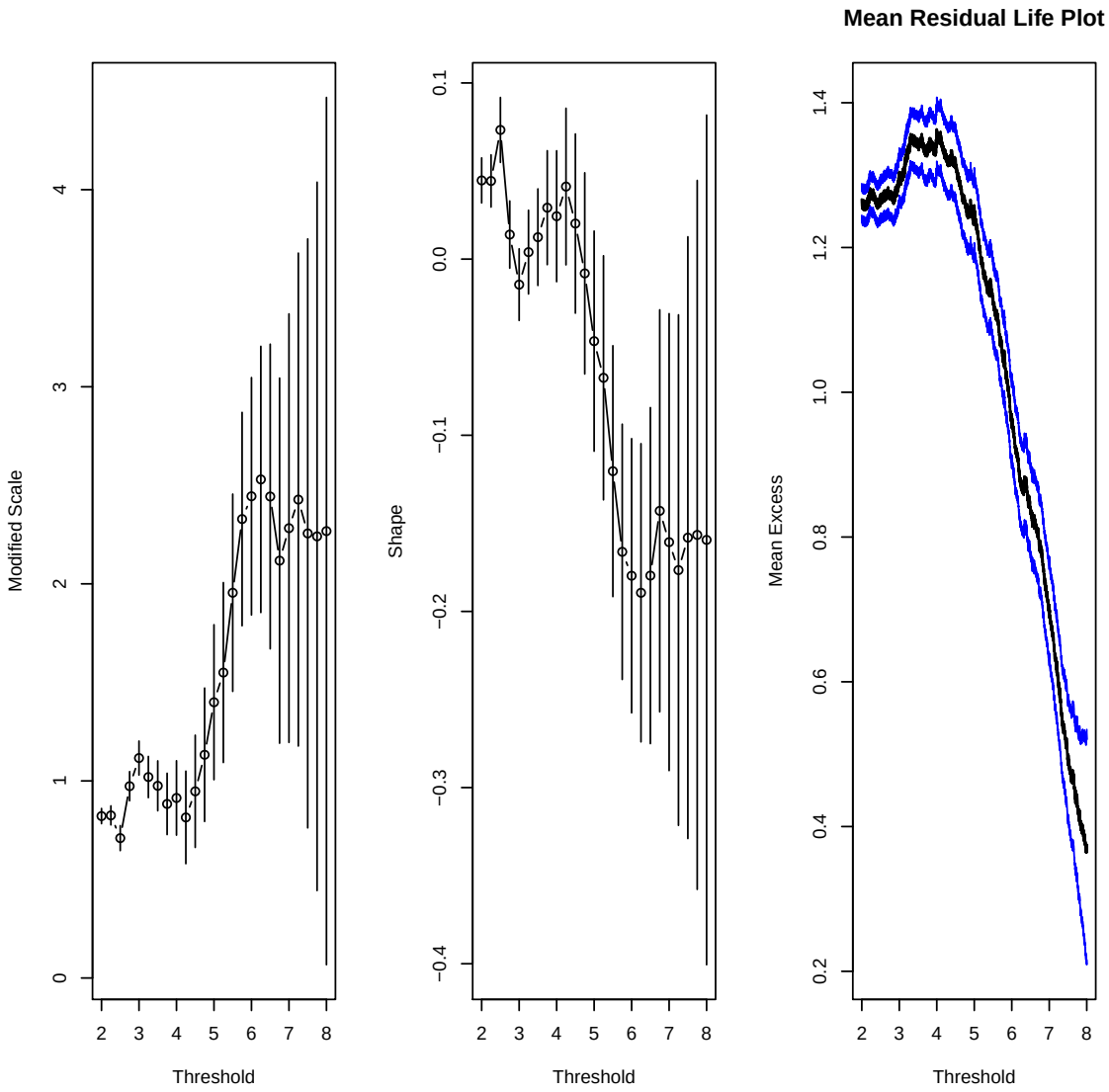


Figure 7.3: Combomune threshold selection (from left to right): Panel (a). First two plots: Threshold choice plots or parameter stability plots; Panel (b). Mean residual life plot for the daily flood height data at Combomune. Both panels for Combomune show MLE estimates and 95% confidence intervals for the transformed parameters in GPD model

Combomune was short-tailed (negative Weibull) and did not come from a Gumbel (exponential) distribution family. The D statistic value for the model pair (M_0, M_1) in Table 7.2 was 33.512 which was too high compared to the critical value of $\chi^2_{1,0.05} = 3.841$. The likelihood ratio test from the NLLH values in Ta-

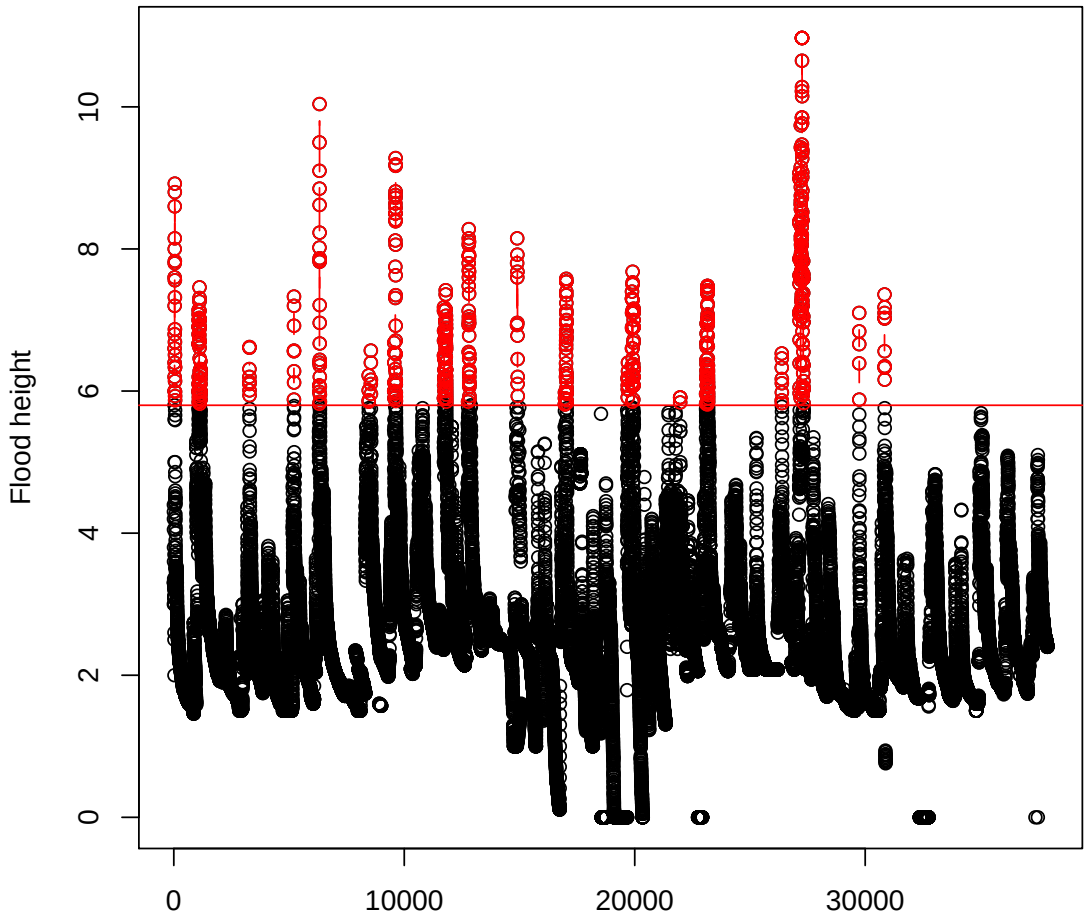


Figure 7.4: Combomune declustered flood heights showing cluster maxima above 5.8 m threshold

ble 7.2 for the test of $\sigma_1 = 0$ was highly significant at 5% level of significance (t-ratio $= 9.481$, p-value < 0.0001). These results showed that the nonstationary GPD model, M_1 was both highly significant and worthwhile over the time-homogeneous GPD model, M_0 in fitting the daily flood heights at Combomune. This suggested that the nonstationary GPD model with a linear trend in scale parameter provided an improvement in fit to the daily flood heights at Com-

bomune over the time-homogeneous GPD model since the D statistic value of 33.512 (> 3.841) was significantly large. The residual diagnostic plots for the nonstationary GPD model suggested a good fit to the data (Figure 7.11). Likewise, the residual diagnostics for the time-homogeneous model also suggested a good fit to the data (Figure 7.8). However, results in this study have revealed overwhelming evidence that the nonstationary GPD model outperformed the time-homogeneous GPD model and provided an improvement in fit over the time-homogeneous GPD model.

Table 7.2: Parameter estimates and negative log-likelihood (NLLH) of the GPD models for Combomune (1966-2010).

Model	$\hat{\sigma_0}$	$\hat{\sigma_1}$	$\hat{\xi}$	**NLLH**
M_0	1.3974615	0	-0.1785666	635.826
M_1	0.0482690	0.0000188	-0.2254202	619.070

The proposed model for Combomune based on the findings of this study is the nonstationary GPD model with a linear trend in the scale parameter. The nonstationary GPD model for Combomune is given in (7.11)

$$G\left(\sigma(t), \xi; y_t, t\right) = 1 - \left(1 + \frac{-0.2254202 y_t}{\exp\left(0.0482690 + 0.0000188t\right)}\right)^{\frac{1}{0.2254202}} \quad (7.11)$$

where $y_t = x_t - 5.8$ are the exceedances over the 5.8 threshold, x_t is cluster maxima of daily flood heights, and $t = 1, 2, ..., 37907, ...$, where 37907 is the time of the last observed flood height value over the period 3 February 1966 to 31 August 2010. Also note that the daily flood height data at Combomune were recorded three times a day, i.e. morning, afternoon and evening meaning that each day had three values of t.

7.3.3 Sicacate models

The time series plot for Sicacate showed that the flood heights are generally high with quite a number of them above 10 m (Figure 1.3, Chapter 1). The graph also showed one extremely high flood height of about 13 m in magnitude. The mean residual life plot and the threshold choice plots (Figure 7.5) were used to come up with a reasonably high threshold of 7.4 m for Sicacate hydrometric station which was chosen to meet the bias-variance threshold trade-off balance such that it was high enough for the asymptotic theorem to be considered accurate and low enough to have sufficient data to estimate the GPD parameters.

Since the exceedances above the 7.4 m threshold for Sicacate appeared in clusters and could not be assumed to be independent from each other, declustering of the cluster maxima was performed and the results are presented in Figure 7.6 for Sicacate cluster maxima.

The shape parameter ξ was significantly different from zero (p-value < 0.0001) for both the time-homogeneous and nonstationary GPD models (Table 7.3), suggesting that the distribution of exceedances over the 7.4 m threshold at Sicacate was short-tailed (negative Weibull) and does not come from a Gumbel (exponential) distribution family. The model pair (M_0, M_1) from Table 7.3 had a D statistic value of 360.042 and a critical value of $\chi^2_{1,0.05} = 3.841$. The likelihood ratio test from the NLLH values in Table 7.3 for the test of $\sigma_1 = 0$ was highly significant at 5% level of significance (t-ratio $= 13.486$, p-value < 0.0001). These results revealed that the nonstationary GPD model, M_1 was both highly significant and worthwhile over the time-homogeneous GPD model, M_0 in fitting the daily flood heights at Sicacate. These findings suggested that the nonstationary GPD model with a linear trend in the scale parameter provided an improvement in fit to the daily flood heights at Sicacate over the time-homogeneous

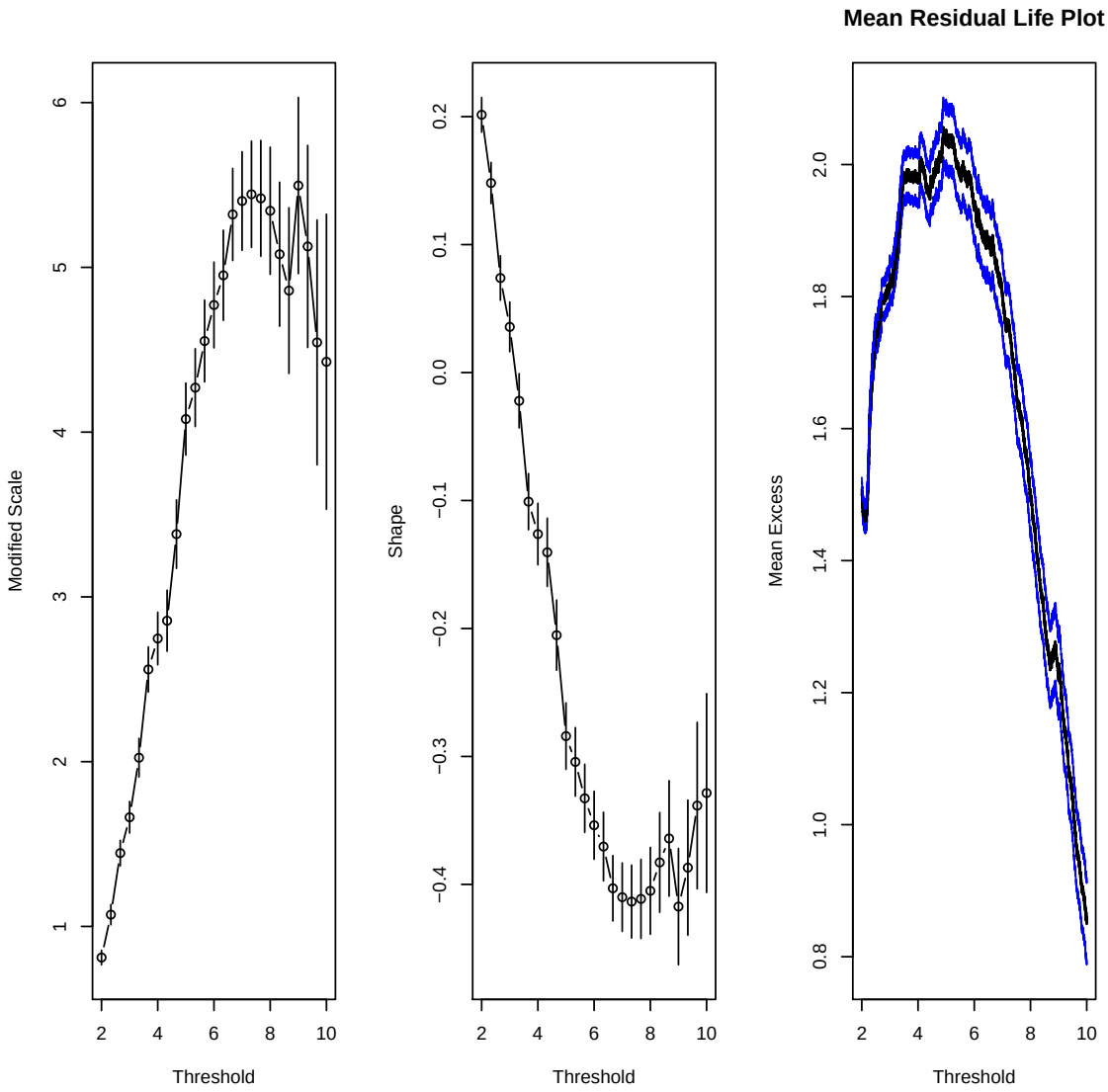


Figure 7.5: Sicacate threshold selection (from left to right): Panel (a). First two plots: Threshold choice plots or parameter stability plots; Panel (b). Mean residual life plot for the daily flood height data at Sicacate. Both panels for Sicacate show MLE estimates and 95% confidence intervals for the transformed parameters in GPD model

GPD model since the D statistic value of 360.042 (> 3.841) was significantly large. The residual diagnostic plots for both the time-homogeneous and the nonstationary GPD models suggested a good fit to the data (Figures 7.9 and 7.12). Since the results of the diagnostic plots from both Figures 7.9 and 7.12

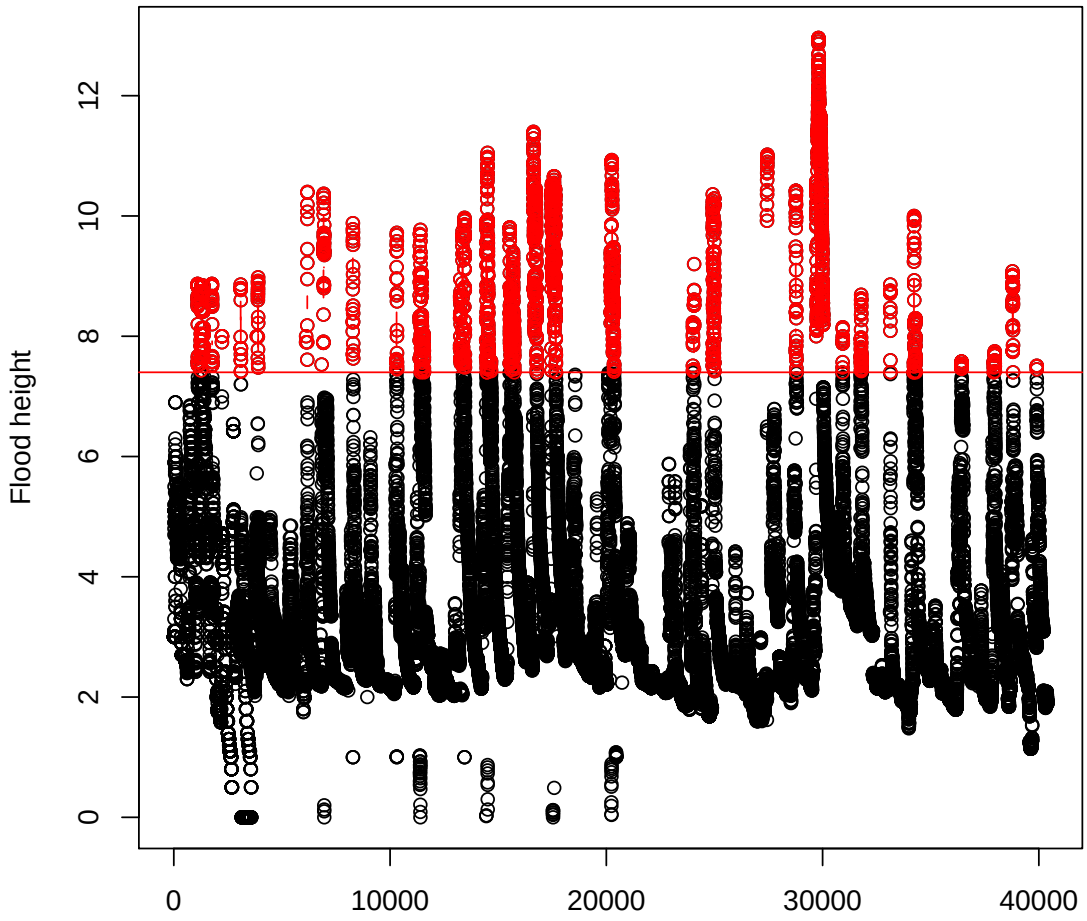


Figure 7.6: Sicacate declustered flood heights showing cluster maxima above 7.4 m threshold

suggested a reasonably good fit of the GPD to the data, this implied that both models could be recommended for modelling the daily flood heights at Sicacate. Nevertheless, overwhelming evidence from the analytical goodness of fit tests suggested that the nonstationary GPD model was more appropriate at the site and is worth proposing because it adds more information in fit over the time-homogeneous model.

Table 7.3: Parameter estimates and negative log-likelihood (NLLH) of the GPD models for Sicacate (1952-2010)

Model	$\hat{\sigma_0}$	$\hat{\sigma_1}$	$\hat{\xi}$	**NLLH**
M_0	2.3711619	0	-0.4105294	2702.413
M_1	0.4627278	0.0000261	-0.6105592	2522.392

The proposed model for Sicacate based on the findings is the nonstationary GPD model with a linear trend in the scale parameter. The nonstationary GPD model for Sicacate is given in (7.12)

$$G\left(\sigma(t),\xi;y_t,t\right) = 1-\left(1+\frac{-0.6105592y_t}{\exp\left(0.4627278+0.0000261t\right)}\right)^{\frac{1}{0.6105592}} \quad (7.12)$$

where $y_t = x_t - 7.4$ are the exceedances over the 7.4 threshold, x_t is cluster maxima of daily flood heights, and $t = 1, 2, ..., 40396, ...$, where 40396 is the time of the last observed flood height value over the period 16 December 1952 to 31 August 2010. Once again, note that the daily flood height data at Sicacate were recorded at most three times a day, i.e. morning, afternoon and evening meaning that each day had at most three values of t.

7.4 Further discussion and general remarks of the results

The data series at all the three sites considered in this study had some missing values in-between the years during the period considered for the study. However, there are two main reasons to be content with the data used to develop the models in this study: (1). In most years where there are missing values they appeared during the winter period when the flood heights are generally low and would have likely missed out on the high threshold even if they were

recorded, (2). The fact that the data was recorded three times a day means we have more data than we would have expected if the data was recorded once a day, for instance, the number of exceedances above high thresholds were 1 494, 550 and 1 860 for Chokwe, Combomune and Sicacate, respectively. The percentages of the number of exceedances over a prescribed threshold compared to the total number of observations at each site were 2.7%, 1.5% and 4.6% for Chokwe, Combomune and Sicacate, respectively. According to Ferreira and de Haan (2015), the POT method is more efficient if the number of exceedances is much larger than the number of blocks. In our case there are 60, 45, and 59 blocks for Chokwe, Combomune, and Sicacate, which means that the number of exceedances at each site are 24.9, 12.2, and 31.5 times the number of blocks, respectively. Literature states that the estimation of high quantiles becomes better when the number of exceedances is 1.65 times the number of blocks for $\xi = 0$, particularly when using maximum likelihood estimators (Ferreira and de Haan, 2015; Cunnane, 1973). Although this study has revealed that the shape parameter is non-zero at all the three sites, the number of exceedances at all the sites is comparatively higher than 1.65 times the number of blocks. All facts pertaining to the data and the findings in this study suggest that the models developed based on this data can be relied upon.

In Chapter 6, a GEV distribution estimated by the MLE method was fitted to block maxima data for the same sites Chokwe, Combomune and Sicacate, and it was found that the distribution of annual daily maximum flood heights at Combomune could be modelled by a nonstationary GEV distribution with a linear trend in the scale parameter, while for Sicacate the proposed GEV models had a linear trend in both location and scale parameters, whereas no evidence of a linear trend in the GEV model was found at Chokwe. The present study found similar results for Combomune and Sicacate except for the additional presence of a linear trend in the location parameter at Sicacate. The major dif-

ferences were found at Chokwe where the present study found a linear trend in the scale parameter of the GPD model which is completely different from the time-homogeneous GEV model recommended based on block maxima data.

The findings in this study concur with the results of Aich et al. (2014) which found climate variability to have a very big impact on the Limpopo River basin streamflow. However, the return periods estimates based on the GPD models found in this chapter appear to be quite higher than those based on the GEV distribution models found in Chapter 6. The reasons for this discrepancy are still unclear and may need further investigation in the future.

7.5 Added value and importance of the study in this chapter

The study in this chapter advances the work of Aich et al. (2014) through incorporating the dominant impact of climate variability in the basin into time-heterogeneous GPD extreme value models. The results of this chapter can contribute as a basis for comparison with the existing models in an attempt to explain the climate change related frequency of floods in the basin, as well as the increased intensity of these floods.

These findings may also be useful in helping the lower Limpopo River basin community prepare and protect itself from future disastrous extreme floods. The Limpopo River basin is very important to the economy of Mozambique because it houses the largest irrigation scheme in the country, Chokwe Irrigation Scheme. The basin also forms the backbone of the economy of the country in terms of agriculture as most of the agricultural activities in the country such as rice production are done in the basin mainly in Chokwe district. This implies that a single disastrous extreme flood such as the one that occurred in

the basin in the year 2000 may bring the economy of the country to its knees. The major highlights of this chapter are in the application of statistics of extremes methods to large volumes of existing data that is untapped in the basin in order to complement the existing methods in the basin used to control and reduce flood disasters.

7.6 Concluding remarks

Statistics of extremes in a changing climate is considered for the lower Limpopo River basin of Mozambique at three sites in an attempt to develop future flood trends for an area that has not been fully studied in Southern Africa. This is the first time climate change extreme value statistics models are applied to the data in the basin. It is hoped that the findings in this study will contribute towards decision making in the basin and help reduce the impact of floods on humans and properties, as well as reduce the amount of aid money required for post disaster recovery and rehabilitation assistance in the basin.

The findings in this study revealed a very strong impact of climate change in the basin which can be modelled by a nonstationary GPD model with a linear trend in the scale parameter. The time-heterogeneous GPD models outperformed the time-homogeneous GPD models at all the three sites suggesting that the nonstationary GPD models are substantially worthwhile and provide an improvement in fit over the time-homogeneous GPD models. This improvement in fit is very important for the planning and policy-making of the government of Mozambique and its partners in the lower Limpopo River basin, where the largest irrigation scheme of the country is situated. The developed time-dependent GPD models would also likely produce more reliable estimates in the frequency of floods since the new models in the basin take into account of the trend in the scale parameter.

Future research will attempt to advance this study to consider Bayesian MCMC inference in a changing climate for the lower Limpopo River basin of Mozambique. Covariates in the form of cycles and/or a physical variable such as a meteorological volatility indicator like SOI which indicates the variability in the ENSO effect in the region will also be considered in future studies involving statistics of extremes with a GPD model in a changing climate.

7.7 Summary of the chapter

In this chapter a time-heterogeneous GPD was fitted to the flood heights in the lower Limpopo River basin of Mozambique. The maximum likelihood method was used for parameter estimation of the nonstationary GPD. An in-depth review of the merits of peaks-over-threshold and block maxima was discussed in this chapter, as a compliment to the extensive discussion in Chapter 2. The relationship between the GEV distribution and the GPD was shown in a mathematical proof and the link between the mathematical proof and the findings was discussed. Nonstationary time-heterogenous GPD models with a trend in the scale parameter were considered in this study. The results show overwhelming evidence in support of the existence of a linear trend in the scale parameter of the GPD models at all the three sites in the basin. The time-heterogeneous GPD models developed in this study were found to be substantially worthwhile and therefore provide an improvement in fit over the time-homogeneous GPD models based on the goodness-of-fit tests evaluated. This study showed the importance of extending the time-homogeneous GPD models to incorporate climate change factors such as trend in the LLRB. The models developed in this study are expected to be more reliable than their stationary counterparts for planning and decision making processes in Mozambique.

APPENDIX 7.1: R PROGRAMMING IN TEXMEX PACKAGE (SELECTED)

R program for declustering exceedances over a high threshold with the trend covariate included

```
pot1=Chokwe.DFH
attach(pot1)
head(pot1)
tail(pot1)
summary(pot1)

#create a new window for graphs
win.graph()
par(mfrow = c(1,1))

#Plotting a time series plot of the data with a threshold line
plot(ts(FH),ylab="Flood Height(m)",main="Chokwe Ts-Plot")
abline(a=4.8,b=0,col="blue")

#Creating a new variable ti to use for GPD modelling
ti=matrix(ncol=1,nrow=56058)
ti[,1]=seq(1,56058,1)

#Calculating t-ratios and p-values
tb1=abs((5.937905e-06)/1.935993e-06); tb1
pt(tb1,56055,lower.tail=FALSE)

#Mean residual life plot and Threshold choice plots
#mean life residual plot
```

```
mrp=mrl.plot(FH)

# Threshold choice plots command used for producing a plot of 30 (in this case) even spaced thresholds over the interval [4,8], nint specifies intervals
gpd.fitrange(FH,4,8,nint=30)

# Declustering excesses over a high threshold with a covariate a trend covariate included
# Installing a package called texmex in R statistical software
install.packages('texmex')
library(texmex)
dc=declust(FH,threshold=4.8)
decl.gpd=evm(dc)
Cho=decl.gpd=evm(dc,ydat=ti,sigl=NULL,siglink=exp)
decl.gpd
plot(Cho)
plot(decl.gpd)
xcl=ts(dc, start=1951,freq=934.3)
plot(dc, col="blue", xlab="Time",ylab="Flood height")
```

APPENDIX 7.2: DIAGNOSTIC PLOTS

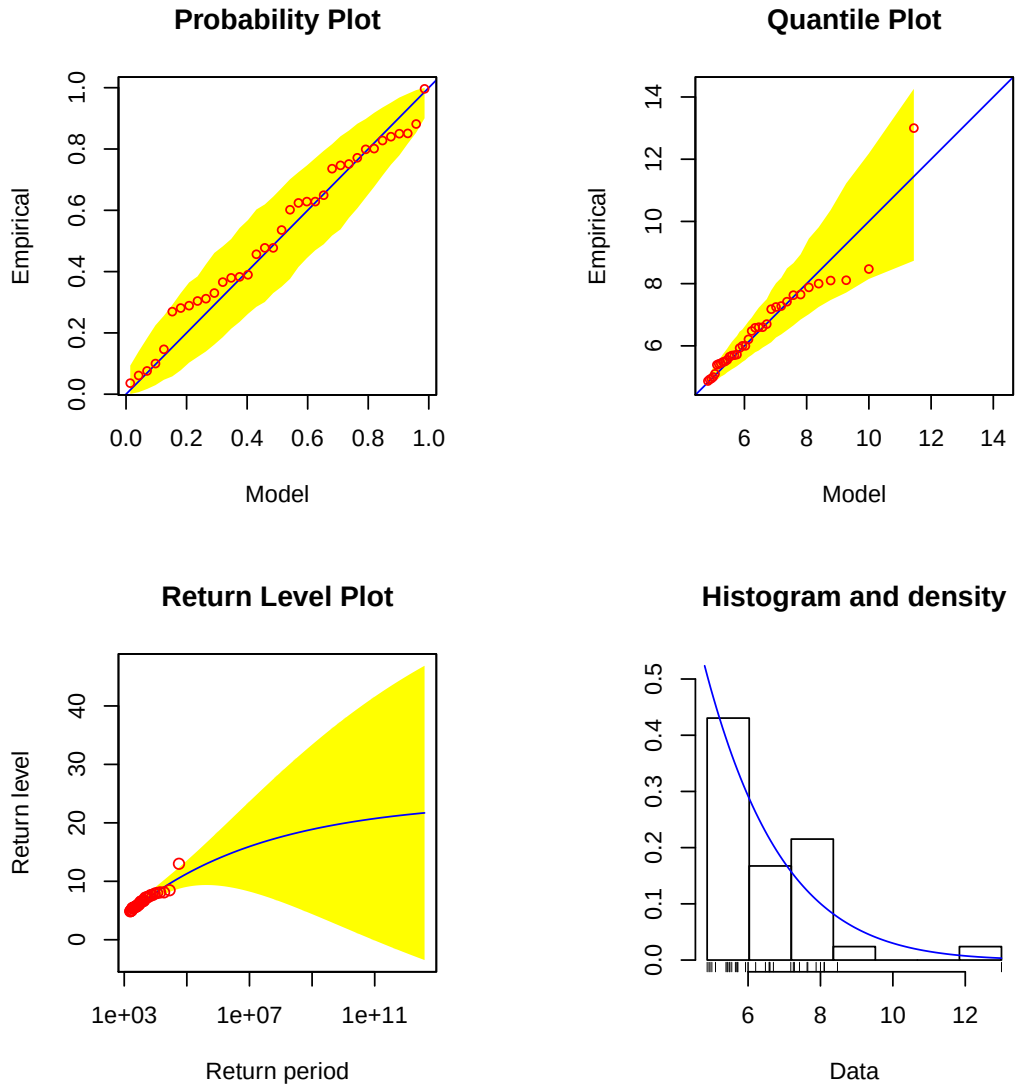


Figure 7.7: Chokwe time-homogeneous GPD diagnostic plots

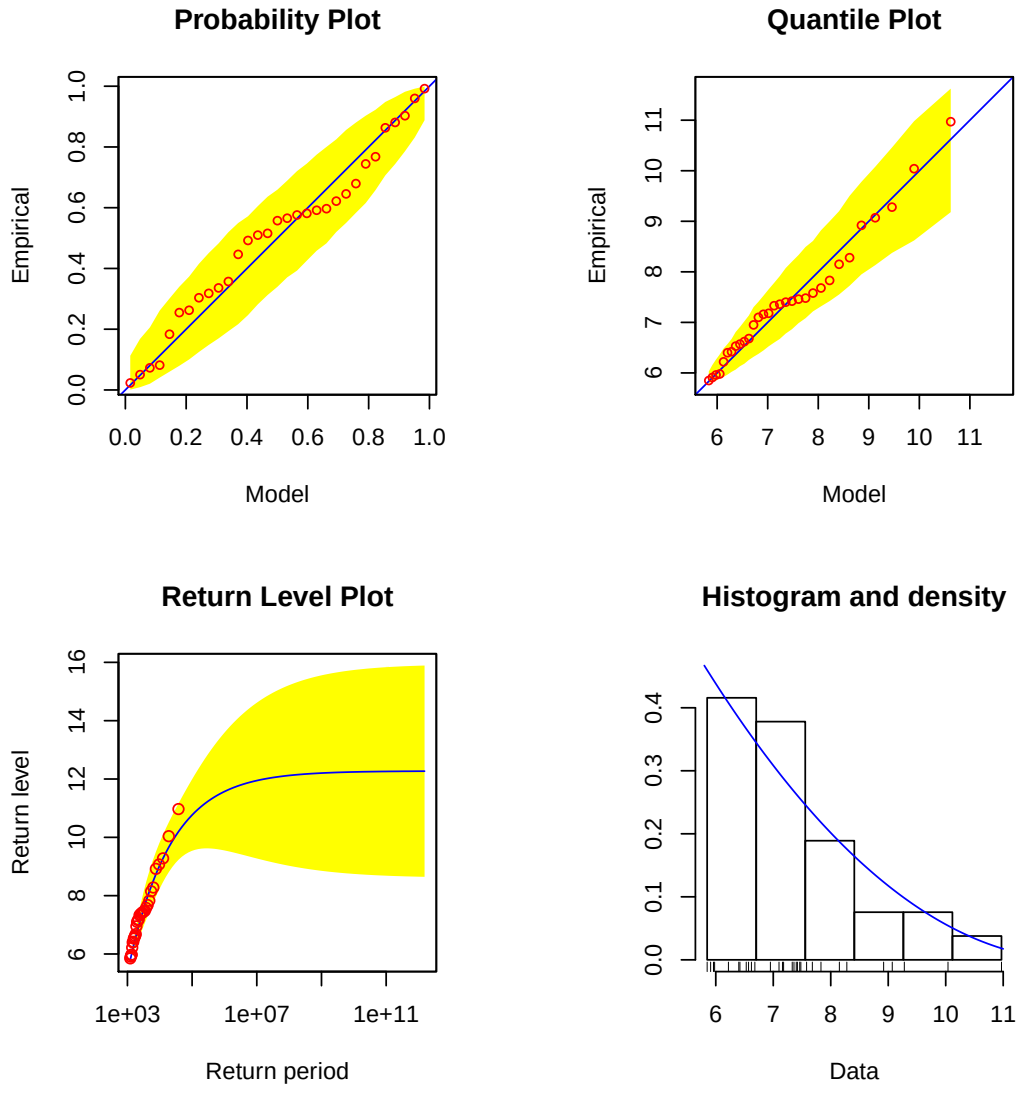


Figure 7.8: Combomune time-homogeneous GPD diagnostic plots

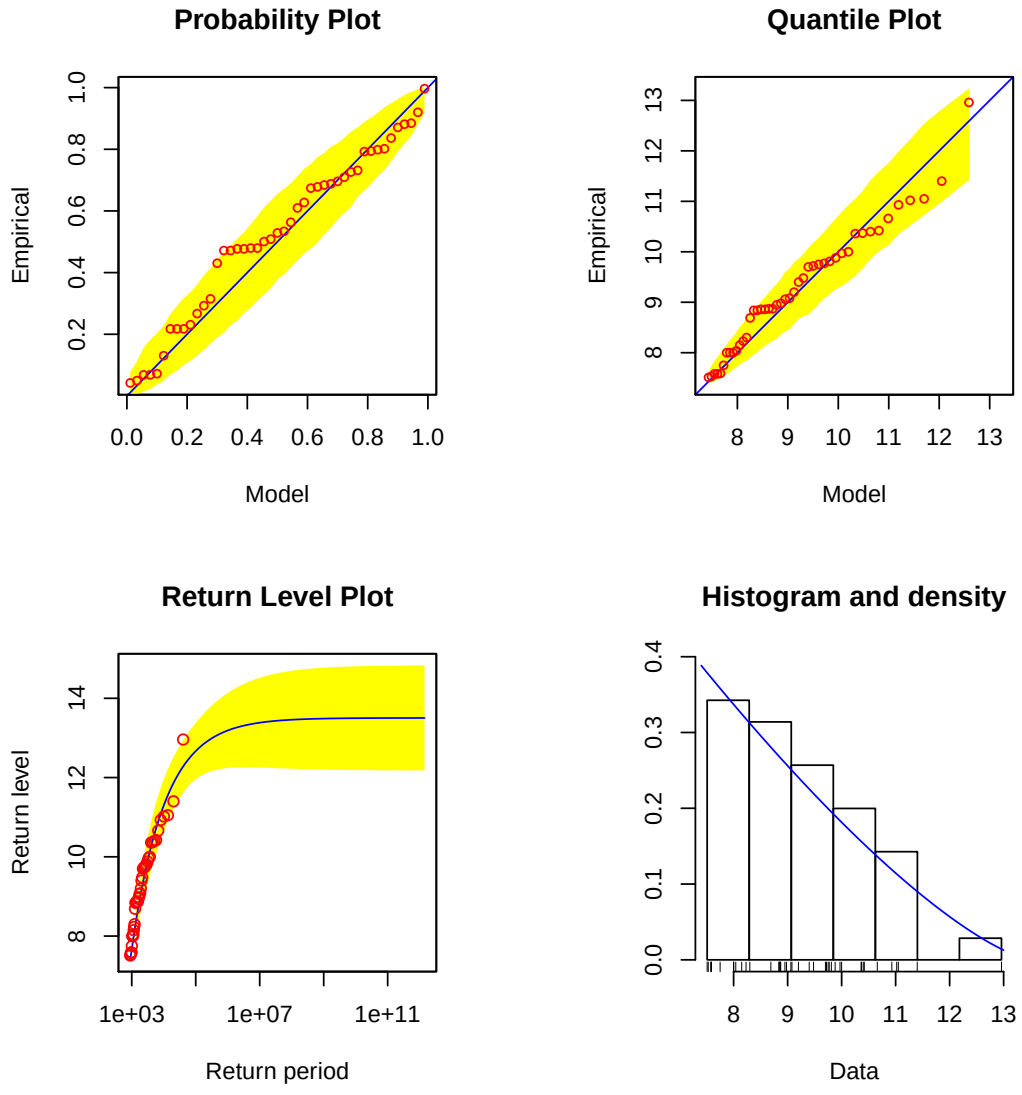


Figure 7.9: Sicacate time-homogeneous GPD diagnostic plots

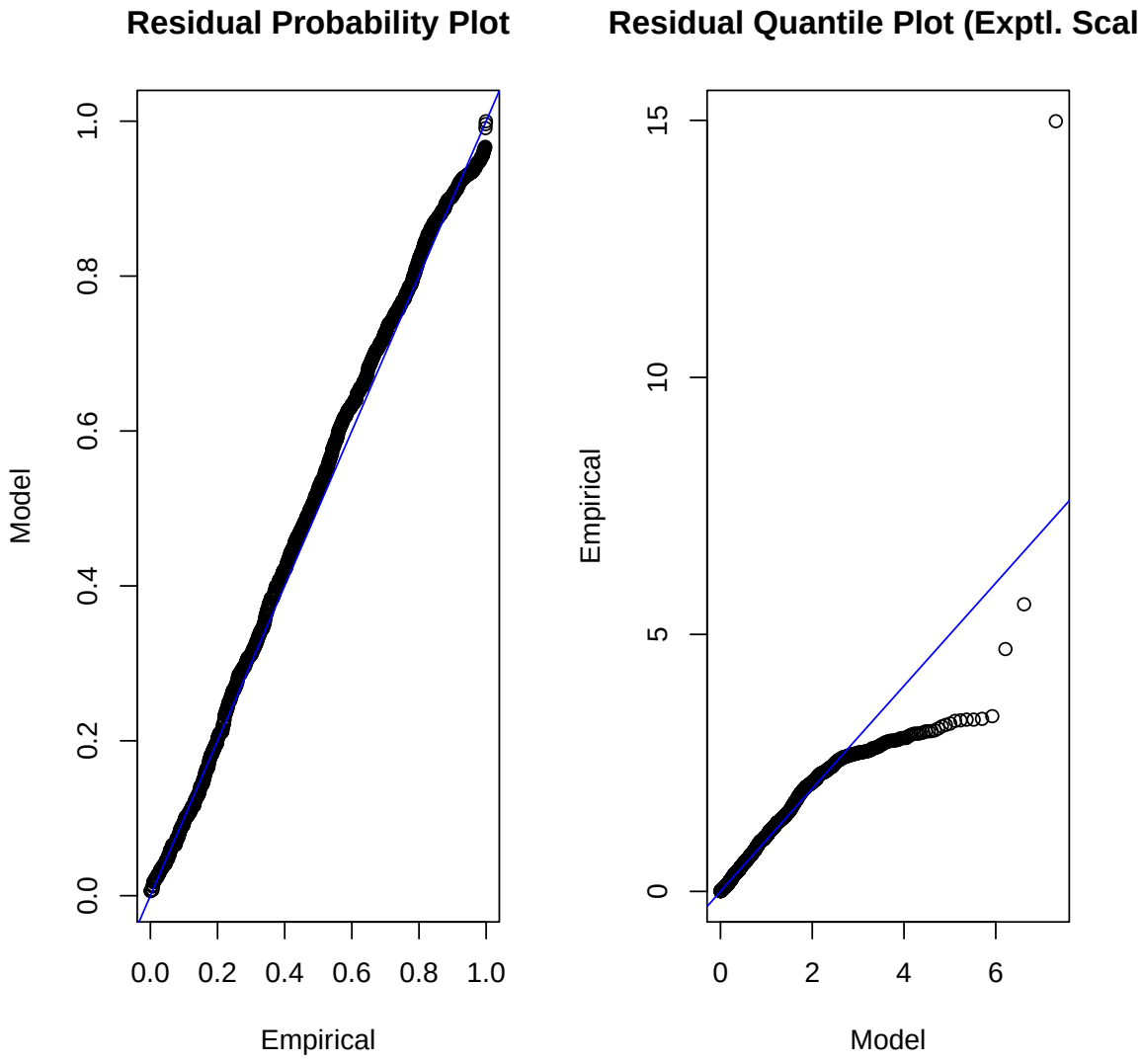


Figure 7.10: Nonstationary GPD diagnostic plots for Chokwe

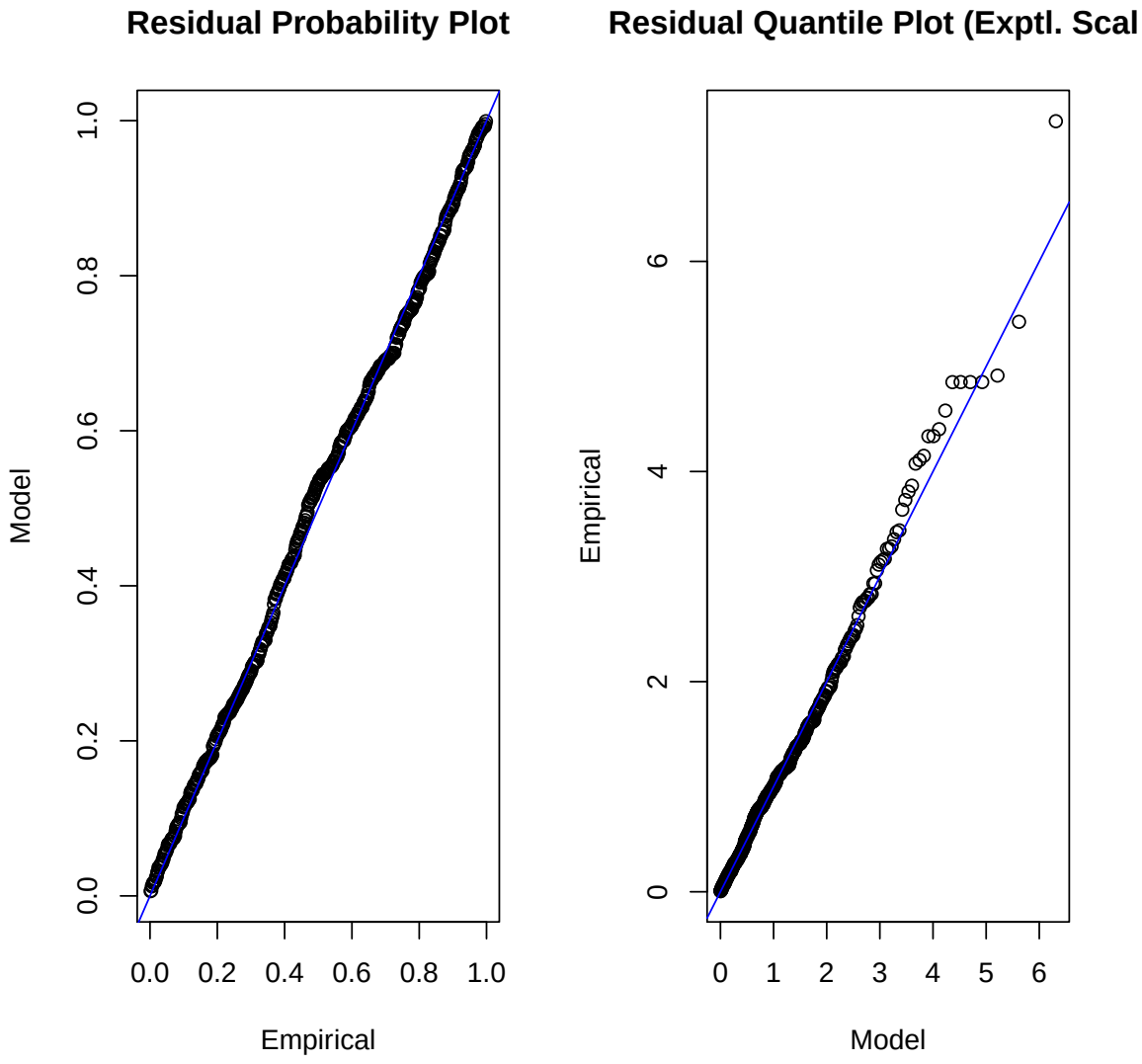


Figure 7.11: Nonstationary GPD diagnostic plots for Combomune

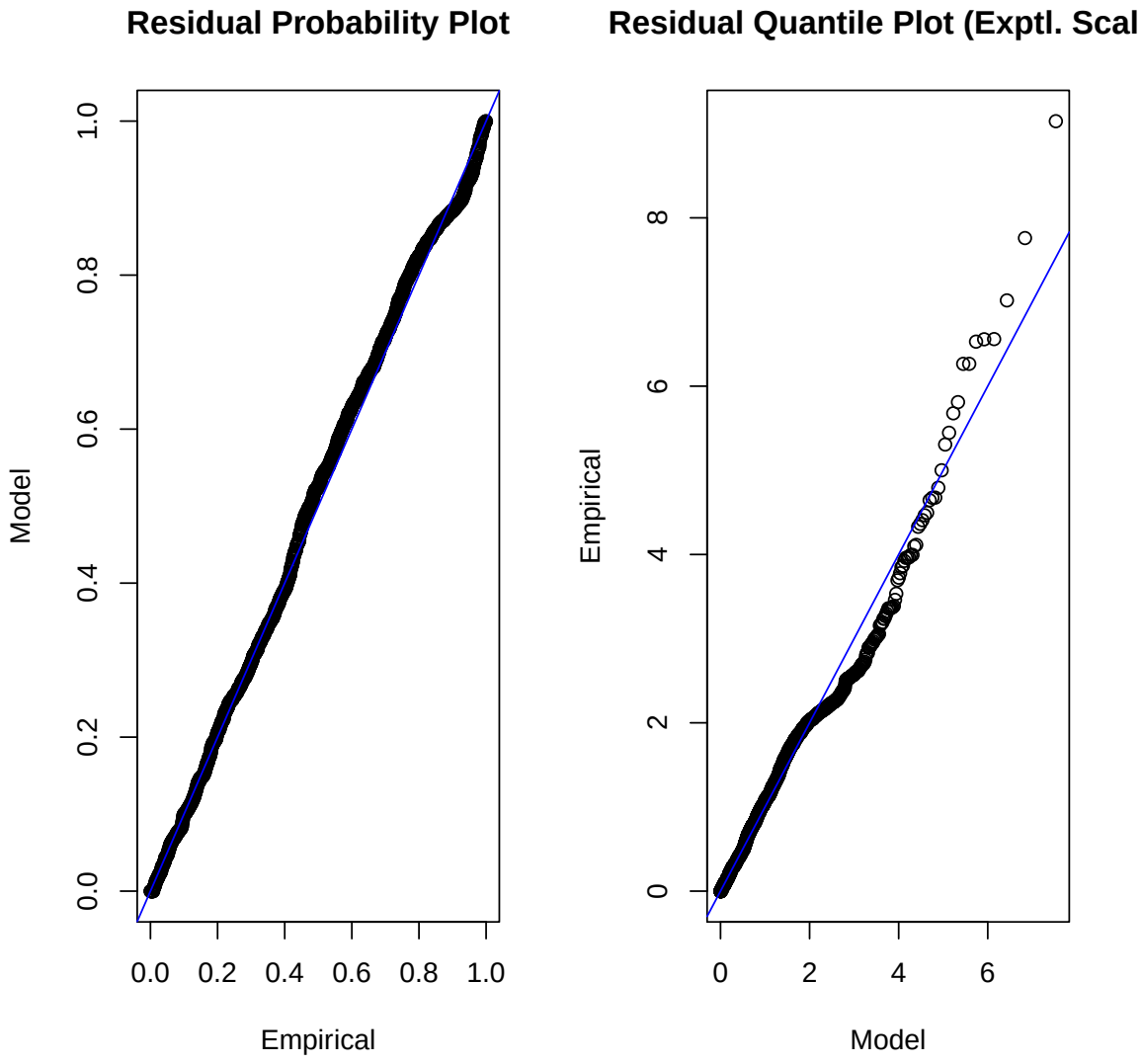


Figure 7.12: Nonstationary GPD diagnostic plots for Sicacate

Chapter 8

On the use of r largest annual maxima order statistics in estimating high quantiles of extreme flood heights in the lower Limpopo River basin of Mozambique

8.1 Introduction

In EVT there are two main realisations namely POT and block maxima. The POT is based on exceedances over a high threshold, while the block maxima is based on the selection of the highest order statistics from a block, for small r,

where r is the order of statistics in decreasing rank order (Beirlant et al., 2004; Coles, 2001). Since the r largest order statistics is selected from a block, this implies that the standard block maxima approach in which only one observation is selected from a block arises when $r = 1$ (Beirlant et al., 2004).

One of the main advantages of POT approach over block maxima is its efficient utilisation of data. On the contrary, the standard block maxima approach discards a lot of data that could add value to the analysis. For instance, in a rainy year in which there is a series of higher order statistics very close to each other in magnitude, only the highest of all is selected. The retention of only a few observations in standard block maxima usually poses a problem particularly in small samples where it is more evident. In order to overcome this shortcoming of standard block maxima, a small number of r largest observations from a block is selected, for $r \geq 1$ (Coles, 2001). It should, however, be noted that order statistics are not independent while in block maxima observations are independent. Extensive literature on r largest is discussed in Chapter 2, Section 10 of this thesis.

The choice of r is based on a trade-off between variance and bias. Too large a value of r will violate the asymptotic theory leading to bias and too small a value of r would lead to large variance (An and Pandey, 2007; Beirlant et al., 2004; Coles, 2001). Extensive theoretical proofs for the complete stability of the largest order statistics are provided in Li and Tomkins (1991). Detailed theoretical stochastic comparisons of the largest order statistics using the generalised gamma distribution family which encompass Weibull, gamma and exponential random variables are provided in Kochar and Torrado (2015). Zhao and Balakrishnan (2015) expanded the topic of stochastic comparisons of largest order statistics to multiple-outlier gamma models. These authors, Zhao and Balakrishnan (2015), made their comparisons in terms of different stochastic

ordering that included the likelihood ratio order, hazard rate order, star order and dispersive order. They also presented a general sufficient condition for the star order.

The main purpose of the present study is the application of the joint asymptotic distribution of the r largest order statistics to model the frequency of annual maximum daily flood heights and to extrapolate their extreme values. A comparison of the analysis with that based on standard block maxima is made in order to investigate the worthiness of r largest order statistics in the analysis of annual maximum flood heights for the LLRB of Mozambique.

The rest of the chapter is arranged in the following order. Section 8.2 gives an overview of the methodology, while Section 8.3 presents the results and discussion. Section 8.4 gives a brief discussion of the value added by the study, while Section 8.5 gives the concluding remarks of the chapter and finally Section 8.6 summarises the chapter.

8.2 Research methodology

Consider $X_1, X_2, ...$ to be a sequence of iid random variables with common distribution function $F(x)$ such that $F(x) < 1 \ \forall x \in \mathbb{R}$ and suppose $M_n^{(r)} = r^{th}$ largest order statistics of $\{X_1, X_2, ...X_n\}$ for r a fixed positive integer and $n \geq r$. Thus $M_n^{(r)}$ is the r^{th} largest of the n observations (Li and Tomkins, 1991, see also Chapter 2, Section 10). The joint pdf is given in Chapter 2, Section 10, Theorem 2.6.

The return level estimate, X_T, corresponding to the return period of T years is

obtained by

$$X_T = \hat{\mu} + \frac{\hat{\sigma}}{\hat{\xi}}\left[\left(-\ln\left(1 - T^{-1}\right)\right)^{-\hat{\xi}} - 1\right]. \tag{8.1}$$

As discussed earlier in literature review in Chapter 2, the choice of r is a major task in r largest order statistics analysis. In this study the researcher chooses $r = 5$ (An and Pandey, 2007; Soares and Scotto, 2004). The argument of the choice of r is based on the findings in Tawn (1988) that the results for the range $r = 3$ up to 7 are very stable and that the method gives very consistent results when r is within that range.

The statistical model presented in (8.1) is based on two principal assumptions: the first assumption being that observations across the years are independent, and the second being that the distribution of r largest observations in a single year follows the joint distribution in (2.32) of Theorem 2.6. According to Soares and Scotto (2004), the first assumption is easily tenable, but the second assumption is not satisfied since the observations within the same year are prone to seasonality and dependence between observed daily flood heights belonging to the same cluster.

In order to satisfy the second assumption, the r largest values are chosen in such a way that dependence between the chosen values is minimised. An assumption was made that the two flood events can be considered independent if they are separated by a window period of 8 days (Ribatet, 2006). Thus after selecting the first largest flood height, sequential steps were taken to select the next r largest values maintaining that the next largest value is at least 8 days away from the preceding extracted value. For instance, suppose the second largest r value falls on 20 April, then values within 8 days prior and within 8 days after cannot be selected, that is , the next value to be extracted must be from 11 April backward or 29 April forward.

8.3 Results and discussion

This section presents the results of the analysis based on r largest order statistics and standard block maxima.

Chokwe

Table 8.1 presents results for the parameter estimates of the limiting joint GEV distribution for the r largest order statistics and results of the GEV parameter estimates of the standard block maxima. An r value of 5 was chosen in this study for the r largest order statistics. The results in Table 8.1 revealed that the standard errors are substantially reduced in model $r5$ compared with model $r1$. The deviance statistic, D for the comparison of the two models is given by

$$D_{(i,j)} = 2\left\{\ell(r_j) - \ell(r_i)\right\} \sim \chi^2_{1,0.05}, \quad \text{for } i \neq j, \tag{8.2}$$

where $\ell(r_i)$ is the log-likelihood function of r_i. To test for the validity of the model r_j relative to r_i at the α level of significance is to reject r_i if $D_{i,j} > \chi^2_{1,0.05}(= 3.841)$ (Soares and Scotto, 2004). From Table 8.1, $D = 2(169.29 - 126.31) = 85.96 > 3.841$, which implies that the r largest order statistics is substantially worthwhile over the standard block maxima model.

Table 8.1: Maximised log-likelihoods ℓ, parameter estimates and standard errors (in parentheses) of r largest order statistics model fitted to the Chokwe flood height data with different values of r.

r	ℓ	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\xi}$
1	126.31	4.26(0.25)	1.79(0.18)	-0.084(0.073)
5	169.29	6.37(0.16)	1.47(0.08)	-0.123(0.037)

In Table 8.2 the results for the tail quantile estimates and predicted extreme flood heights for Chokwe are presented. The results showed that the predicted

extreme flood heights values are comparable and both models show that the 13 m flood height that occurred at the site in 2000 is above the 500-year flood level. It can be observed from Table 8.2 that the predicted extreme flood heights at Chokwe based on the r largest order statistics are substantially higher than those based on standard block maxima ML approach up to the 250-year flood level and thereafter, on the contrary, the predicted flood height values based on r largest order statistics become increasingly less than those based on standard block maxima ML method.

Table 8.2: r largest order statistics tail quantile estimation and prediction of extreme flood heights for Chokwe

F(x)	**p**	**T**	**Standard ML estimate** (Exceedances)*	**r largest estimate** (Exceedances)*
98^{th}	0.02	50 years	10.22 m (1)	10.94 m (1)
99^{th}	0.01	100 years	11.10 m (1)	11.55 m (1)
99.5^{th}	0.005	200 years	11.92 m (1)	12.11 m (1)
99.6^{th}	0.004	250 years	12.18 m (1)	12.28 m (1)
99.8^{th}	0.002	500 years	12.94 m (1)	12.78 m (1)
99.9^{th}	0.001	1,000 years	13.65 m (0)	13.24 m (0)
99.99^{th}	0.0001	10,000 years	15.76 m (0)	14.51 m (0)

Key: $F(x)$ represents non-exceedance probability.
p represents exceedance Probability.
T represents return period.
(Exceedance)* represents the number of empirical observations above the flood level at the site.

Combomune

The results in Table 8.3 are the Combomune parameter estimates of the limiting joint GEV distribution for the r largest order statistics for Combomune and the results of the estimates revealed that the standard errors for model $r5$ are substantially lower with reference to those of $r1$ model. Additionally the deviance statistic, $D = 29.62 > 3.841$, which indicates that the model $r5$ is sub-

Table 8.3: Maximised log-likelihoods ℓ, parameter estimates and standard errors (in parentheses) of r largest order statistics model fitted to the Combomune flood height data with different values of r.

r	ℓ	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\xi}$
1	90.80	5.16(0.28)	1.66(0.20)	-0.123(0.110)
5	105.61	6.96(0.17)	1.38(0.11)	-0.083(0.056)

stantially worthwhile with reference to model $r1$.

Table 8.4: r largest order statistics tail quantile estimation and prediction of extreme flood heights for Combomune

F(x)	**p**	**T**	**Standard ML estimate** (Exceedances)*	**r-largest estimate** (Exceedances)*
98^{th}	0.02	50 years	10.31 m (1)	11.55 m (0)
99^{th}	0.01	100 years	11.00 m (0)	12.23 m (0)
99.5^{th}	0.005	200 years	11.63 m (0)	12.87 m (0)
99.6^{th}	0.004	250 years	11.82 m (0)	13.06 m (0)
99.8^{th}	0.002	500 years	12.38 m (0)	13.65 m (0)
99.9^{th}	0.001	1,000 years	12.89 m (0)	14.21 m (0)
99.99^{th}	0.0001	10,000 years	14.32 m (0)	15.83 m (0)

Key: $F(x)$ represents non-exceedance probability.
p represents exceedance Probability.
T represents return period.
(Exceedance)* represents the number of empirical observations above the flood level at the site.

Table 8.4 presents results for the tail quantile estimation and predicted extreme flood height values for Combomune. The results in Table 8.4 showed that the predicted extreme flood heights based on r largest order statistics are quite higher than those based on standard block maxima ML method. It is further observed that the 10.97 m flood height that occurred at Combomune during the year 2000 floods is below the 50-year flood height level based on the r largest order statistics approach. However, the 13 m flood height which

occurred at downstream Chokwe and Sicacate is just above the 200-year flood height level.

Sicacate

Table 8.5: Maximised log-likelihoods ℓ, parameter estimates and standard errors (in parentheses) of r largest order statistics model fitted to the Sicacate flood height data with different values of r.

r	ℓ	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\xi}$
1	148.55	6.15(0.47)	3.33(0.35)	-0.454(0.071)
5	298.17	9.63(0.19)	1.71(0.06)	-0.462(0.032)

Table 8.5 presents results for the parameter estimates of the limiting joint GEV distribution for the r largest order statistics for Sicacate and the results of the estimates revealed that the standard errors for model $r5$ are substantially lower compared to those of $r1$ model. Moreover, the deviance statistic, $D = 299.24 > 3.841$, is very large which indicates that the model $r5$ is substantially worthwhile over the model $r1$.

Table 8.6 presents results for the tail quantile estimation and predicted extreme flood heights for the r largest order statistics for Sicacate. The results show that the predicted extreme flood heights based on r largest order statistics are quite comparable to those based on standard block maxima ML method. Based on the r largest order statistics results the 13 m flood height of the year 2000 has a return period of 200 years.

8.4 General remarks and added value of the study

In this chapter the r largest order statistics was discussed and applied to the Limpopo River data at three sites. The results of the r largest order statistics

Table 8.6: r largest order statistics tail quantile estimation and prediction of extreme flood heights for Sicacate

F(x)	p	T	**Standard ML estimate** (Exceedances)*	**r-largest estimate** (Exceedances)*
98^{th}	0.02	50 years	12.23 m (1)	12.71 m (1)
99^{th}	0.01	100 years	12.57 m (1)	12.88 m (1)
99.5^{th}	0.005	200 years	12.82 m (1)	13.00 m (0)
99.6^{th}	0.004	250 years	12.88 m (1)	13.03 m (0)
99.8^{th}	0.002	500 years	13.04 m (0)	13.11 m (0)
99.9^{th}	0.001	1,000 years	13.16 m (0)	13.17 m (0)
99.99^{th}	0.0001	10,000 years	13.37 m (0)	13.27 m (0)

Key: $F(x)$ represents non-exceedance probability.
p represents exceedance Probability.
T represents return period.
(Exceedance)* represents the number of empirical observations above the flood level at the site.

were compared to those of standard block maxima based on ML method. The results across all the sites considered in this study showed consistently that the r largest order statistics leads to a substantial reduction of standard errors of the parameter estimates compared to the standard errors based on the standard block maxima ML estimates. In general, the predicted extreme flood heights are substantially higher for r largest order statistics than the associated predicted extreme flood heights based on standard block maxima across all sites.

The main contribution of this chapter was the introduction of an approach similar to that used for declustering in POT (Ribatet, 2006) to ensure that the r largest order values within the same year are not dependent, but instead they are independent or nearly independent. This was done by considering observed flood heights within an 8 days' window period (which is just over a week) to have some dependence structure between them and, on the other hand, observed flood heights outside this window period are considered to be indepen-

dent. The 8 days' window period used in this chapter is supported by results in Chapter 4 where the characteristics of the distributions of time series annual maxima moving sums showed a tendency to differ after 7 days. This approach can be important in many applications concerning r largest order statistics.

8.5 Concluding remarks

In this chapter an investigation is made on the use of r largest order statistics in flood frequency analysis in a block maxima realisation. The probability density function of the limiting joint GEV distribution is presented. Since the limiting joint GEV distribution is an extension of the standard GEV distribution the parameters are interpreted in the same manner and can be easily compared as if they were coming from the same distribution. The addition of more data in r largest improves precision in the results of the estimates and predictions by substantially reducing the standard errors, which are usually higher in the standard block maxima.

It can be concluded that the r largest order statistics model (for $r = 5$) is a worthwhile model and substantially reduces the standard errors in the analysis. However, it is crucial to emphasise that when considering r largest order statistics it is important to use some practical procedure to ensure that r largest order values within the same block are independent.

8.6 Summary of the chapter

The study for this chapter considers two approaches in a block maxima realisation for flood frequency analysis. The two approaches studied were the r largest order statistics and standard block maxima ML methods. The findings in this study have revealed that the r largest order statistics improves the

precision of the results in an analysis by substantially reducing the standard errors of the estimates. The results also showed that the r largest order statistics approach produces predicted extreme flood heights (return levels) that are generally greater than corresponding flood heights based on standard block maxima ML approach. A technique to achieve the independence of r largest order values within the same year is discussed in the chapter and a window period is suggested. It was concluded that the r largest order statistics model is worth applying to the Limpopo flood heights data.

Chapter 9

Conclusion

9.1 Introduction

In this thesis the researcher has analysed, presented and discussed numerous approaches to flood frequency analysis, fundamental approaches of EVT and several techniques of data analysis, among other issues. The main thrust of this thesis is in the extensive application of methods based on statistics of extremes to flood heights data series in the lower Limpopo River basin of Mozambique, specifically at the sites of Chokwe, Combomune and Sicacate. Since the data in its original form was daily flood heights, this allowed for the data to be organised into various forms which consequently allowed for various EVT techniques to be applied to the data.

Mozambique is still in the category of economically challenged countries and the lower Limpopo River basin of Mozambique is characterised by extreme natural hazards, alternating between extreme floods and severe droughts. The Limpopo River basin hosts the Chokwe Irrigation Scheme which forms the cor-

nerstone of the agricultural economy of Mozambique. Modelling extreme floods in the basin can play a vital role through reducing the prediction uncertainties and enhancing disaster preparedness and consequently save the nation billions of American dollars in post disaster relief operations. A summary of the main conclusions reached in the various parts of the thesis is presented in this chapter.

The rest of the chapter is organised in the following order. Section 9.2 gives the overall summary and concluding remarks of the thesis, while Section 9.3 summarises the key findings and contributions of the thesis. Section 9.4 presents the limitations of the thesis and Section 9.5 gives the future research directions of the thesis.

9.2 Thesis summary and concluding remarks

The thesis was organised into various chapters in order to be able to tackle the problems step by step to answer the objectives set in Chapter 1, Section 1.4.2. Chapter 1 dealt with the historical background of the Limpopo River, its origins, benefits and drawbacks. The chapter also gave a brief literature on historical developments of extreme value theory (EVT). A list of the main contributions of the thesis including an outline of the chapters of this thesis that were published in peer-reviewed journals are given in Chapter 1.

An extensive review of relevant literature was performed and discussed in Chapter 2. The review started with a discussion of the existing methods in LLRB of Mozambique as well as the new existing methods in other countries, including developed countries. Implementations from worldwide disaster conferences were discussed as well as some disaster risk reduction efforts being made worldwide to reduce the detrimental effects of these natural disasters

on humans and property. An extensive review of literature was performed on the fundamental approaches of flood frequency analysis (FFA) that include at-site and regional FFA. Moreover, an intensive literature was also performed on the fundamental realisations of EVT which include block maxima and POT. The parameter estimation techniques used in the two FFA approaches were discussed and reviewed. The probability models currently used in the LLRB were reviewed as well as the flood forecasting and early warning systems in the basin. A lot if gaps were exposed in literature including the scarcity (or lack) of statistical models in the LLRB under study.

The recent articles by Bücher and Segers (2016), Dombry (2015), Ferreira and de Haan (2015) and Ferreira and de Haan (2014) form the backbone of the theories applied in this thesis. Although the main focus of this thesis is in application of these recently developed (or revisited) techniques, the theorems stated and proved in this thesis help our understanding of the theoretical developments which greatly improves our application of the methods. The Bayesian MCMC and r largest order statistics approach were also extensively reviewed in a block maxima realisation. Based on theoretical arguments of Ferreira and de Haan (2015) it was concluded that the block maxima approach is more efficient in a number of situations compared to the POT approach. The result by Ferreira and de Haan (2015) is a major turning point in current research as majority of the previous research would claim that the POT method is more efficient than the block maxima since it optimises the use of available data.

In general, the literature reviewed in Chapter 2 revealed several gaps in the theory and applications of EVT. Several authors have advocated for the improvement of the existing methods in literature and the need to evaluate the new methods adopted in other countries (usually developed countries), with the view that these methods may also work in the rivers and catchments of de-

veloping countries, given different operating characteristics and climatic conditions.

Chapter 3 investigated a number of candidate distributions over their goodness-of-fit for the LLRB flood height data at three sites. Among the candidate distributions investigated were the GEV, generalised gamma, Gumbel, Weibull and many others. The main objective of Chapter was to identify at most three suitable distributions for a particular site. The GEV and Gumbel distributions prevailed at Sicacate, while at Combomune the GEV, Ga2 and LP3 were the most suitable distributions. Furthermore, the Ga3, LN3 and GEV were the best distributions for Chokwe. The findings in Chapter 3 revealed that the GEV distributions prevailed at all the three sites suggesting its dominance in modelling the Limpopo River extreme flood heights. The prevailing distributions for the study sites in Chapter 3 were eventually used to predict the extreme flood heights and their corresponding return periods at each site. The findings in Chapter 3 concerning the return period of the disastrous 13 m flood height that occurred in the year 2000 were in agreement with those of Smithers et al. (2001, with references therein) that this flood height is above the 100-year flood height. Specifically, the 13 m flood height based on the results of Chapter 3 had a return period in excess of 250 years at some sites and in excess of 500 years at other sites.

Chapter 4 explored the characteristics of the moving sums of annual maximum flood heights in a block maxima approach. Six annual maxima time series models were considered at each site and compared with respect to skewness, excess kurtosis, coefficient of variation, and the goodness-of-fit of the GEV distribution as measured by Anderson-Darling and Kolmogorov-Smirnov test statistics. The six annual maxima time series models considered at each site were the annual daily (AM1), two-day (AM2), five-day (AM5), seven-day (AM7), ten-

day (AM10) and thirty-day (AM30) maximum flood heights. The findings in the study for Chapter 4 revealed no sufficient evidence of significant differences among the six annual maxima time series models at all the three sites with respect to the characteristics tested. The study also revealed notable differences between the three sites considered in the study in overall skewness of the annual maxima time series models, with Chokwe and Combomune (upstream) dominated by positive skewness and Sicacate (downstream) dominated by negative skewness. This is reasonable in the real world since it must be expected that there is usually more water downstream than upstream and consequently the flood heights become higher downstream than upstream. Although not statistically significant, there were indications in this study that the characteristics of the moving sums begin to show changes after the 7^{th} day across all the sites.

Chapter 4 established that using the annual daily maximum flood heights model or any of the other five annual maxima time series models in flood frequency analysis has no significant effect on the forecasts of extreme return levels and their corresponding return periods. Without loss of generality, the researcher recommends the use of the annual daily maximum flood heights model in flood frequency analysis in the construct flood frequency curves mainly due to its simplicity and the relative ease of obtaining it.

Chapter 5 is an extension of Chapters 3 and 4. The GEV distribution was used as the likelihood function and its parameters were estimated using the Bayesian MCMC approach and maximum likelihood method (for the frequentist approach). The choice of the GEV distribution over other distributions was inspired by the findings in Chapter 3. The findings in Chapter 5 revealed the importance of the inclusion of uncertainties through a prior distribution in Bayesian MCMC analysis. Chapter 5 established that the inclusion of a

prior distribution substantially improves the precision of tail quantile estimates and at-site predicted extreme flood heights. Furthermore, it was established that the return levels based on Bayesian MCMC approach are substantially higher than their corresponding maximum likelihood based return levels. These Bayesian estimates of maximum flood heights and their associated return periods appear to be closer to reality than those based on ML approach. Modelling the upper tail efficiently may help in reducing the impact of a flood event through disaster preparedness and management. These findings are in agreement with those in Reis and Stedinger (2005). In general, the Bayesian MCMC and frequentist approaches produce results that are somewhat comparable. Nevertheless, the Bayesian MCMC realisation offers more attractive results due to its ability to take into account the uncertainties involved in the hydrological processes of flood heights through a prior distribution.

Chapter 6 is an extension of Chapters 3, 4 and 5 with reference to the maximum likelihood estimation of the GEV parameters. The chapter considered modelling annual maximum flood heights using of statistics of extremes in a changing climate for the lower Limpopo River basin of Mozambique. The maximum likelihood estimation method was used to estimate the parameters of the GEV distribution in the presence of a long-term trend covariate and an indicator of meteorological volatility known as seasonal oscillation index (SOI). Chapter 6 revealed that using statistics of extremes in a changing climate provides a substantial improvement in fit over the time-homogeneous models. This improvement in fit is very important for the planning and policy-making of the government of Mozambique and its partners in the lower Limpopo River basin, where the largest irrigation scheme of the country is situated. The importance of the developed models is attributed to the fact that these nonstationary models take into account the reasons for increased frequency of floods in the basin such as El Niño an La Nińa effects. Once the government and its partners are

fully aware of the reasons behind the increased frequency of floods in the basin their planning can be greatly improved.

In order to make efficient use of the abundance of data available, Chapter 7 considered the POT approach of EVT. Again, statistics of extremes in a changing climate was considered for the lower Limpopo River basin of Mozambique in a POT approach in an attempt to develop future flood trends for an area that has not been fully studied in Southern Africa. Chapter 7 established that a nonstationary GPD model with a linear trend in the scale parameter can be used to model extreme flood heights in the LLRB. It is hoped that the findings in Chapter 7 will contribute towards decision making in the basin and help reduce the impact of floods on humans and properties, as well as reduce the amount of aid money required for post disaster recovery and rehabilitation assistance in the basin. The developed nonstationary GPD models would also likely produce more reliable estimates in the frequency of floods since the new models in the basin take into account of the trend in the scale parameter.

In Chapter 8, the standard block maxima approach in Chapters 3, 4, 5, and 6 is extended to the r largest order statistics. Chapter 8 established that the r largest order statistics substantially reduces the standard errors of the estimates. The use of standard block maxima and r largest order statistics produce results which are fairly comparable. However, the r largest order statistics is more attractive since it substantially reduces the standard errors due to its ability to include more data in a block maxima realisation.

9.3 Summary of the key findings and contributions

The following were the key findings of this thesis:

1. The GEV distribution is the most suitable distribution to model extreme flood heights across all sites in the lower Limpopo River basin of Mozambique. The other alternative distributions that are site specific are the Gumbel distribution for Sicacate, two-parameter gamma and log-Pearson type 3 for Combomune, and three-parameter gamma and three-parameter lognormal for Chokwe.

2. The 13 m flood height which occurred in the year 2000 in the lower Limpopo River basin of Mozambique and parts of South Africa and Zimbabwe was a very rare flood event with a return period in excess of 100 years based on all approaches used to analyse the data in this thesis. Specifically the 13 m flood height based on the frequentist maximum likelihood and L-moments approaches has a return period in excess 250 years at Sicacate and a return period in excess 500 years at Chokwe and Combomune, whereas based on the Bayesian MCMC approach the return period reduced to about 100 years for Sicacate, and in excess of 250 years for Chokwe and Combomune.

3. The Bayesian MCMC and frequentist approaches give results that are relatively comparable. Nevertheless, the Bayesian approach is more attractive since it has a provision of taking into account uncertainties in hydrological processes in the estimation of the parameters. Thus the Bayesian MCMC approach is more informative when compared to the frequentist (classical) approach.

4. The moving sums of the annual maximum flood heights explored in this study revealed no significant difference in characteristics. However, changes

in the skewness characteristics of the moving sums begin to be more noticeable after the 7^{th} day annual maxima moving sum. This prompted the use of the 8 days window in the choice of r independent largest order statistics.

5. For a suitably chosen r, the r largest order statistics substantially reduces standard errors of the parameter estimates in a block maxima realisation.

6. For suitably declustered exceedances, nonstationary GPD models with a linear long-term trend in the log-scale parameter can provide appropriate models for the daily flood heights in a POT realisation for the lower Limpopo River basin.

7. Nonstationary models with linear trend in the scale and location parameters of the GEV distribution provide substantially better models than the stationary GEV models for the Combomune and Sicacate annual maxima flood heights data. Annual maximum flood heights at Chokwe can be modelled by a stationary GEV model since the long-term trend covariate is not worthwhile at the site.

8. The SOI covariate can provide substantially improved and worthwhile nonstationary GEV models that can explain the effect of climate variability on annual maxima flood heights at all the three sites in the lower Limpopo River basin.

9. Spatial variability was found to exist in the lower Limpopo River basin in this thesis with reference to the 13 m flood height of the year 2000 disastrous floods.

9.4 Limitations of the thesis

The limitations for this thesis stem from the number of study sites used in the analysis. This thesis concentrated on three hydrometric stations Chokwe, Combomune and Sicacate due to limited number of hydrometric stations with quality data in the lower Limpopo River basin of Mozambique. It would have been much desirable to use data from a large number of hydrometric stations in the basin which would allow the researcher to explore other research methods such as spatial extremes. Missing data has been a major issue in the majority of the available stations in the basin and this has not only limited the number of hydrometric stations used in this thesis, but also cut short the length of the series in the stations used. For instance, some hydrometric stations started collecting data in as early as the 1930s but due to missing data the series had to be cut to commence in the 1950s for the analysis. The other data provided by DNA was rainfall data (in millimetres) collected at rain gauge stations in the catchment areas of Chokwe, Combomune, Pafuri and Sicacate. The rainfall data would have been ideal for spatial analysis because it is collected over all catchments surrounding a hydrometric station, nonetheless, the rainfall data had serious cases of missing values such that not even one station had quality data and this rendered the rainfall data unusable. The other recent limitation was the difficulty encountered in trying to obtain updated flood heights data for the period 2011-2016 from the DNA Mozambique mainly due to staff turn over and changes in management personnel.

9.5 Future research directions

The findings of this thesis provide possible areas for further research in the future. The following possible future research directions are suggested:

- One possible direction for future research may include the use of statistics of spatial extremes in the basin with the GEV distribution as the likelihood. This may help accommodate the problem of spatial variability found in the thesis. This will require that more sites (stations) with quality data become available.

- Regional flood frequency analysis is also proposed for future research in the basin. This will accommodate the ungauged sites in the region or catchment of interest.

- Future studies may also attempt to advance this study to consider Bayesian MCMC methods in a changing climate in both GEV distribution and GPD models for the lower Limpopo River basin of Mozambique.

- Research in the future may also consider modelling the stochastic behaviour of extreme flood heights using the Poisson point processes.

- Covariates in the form of cycles and/or climate change indicator variables such as SOI which indicates the variability in the ENSO effect in a region may also be considered in future studies involving statistics of extremes with a GPD model in a changing climate.

- Future studies may also explore the use of a Bayesian prior distribution proposed by Martins and Stedinger (2000), to restrict ξ values to a statistically or physically reasonable range in a generalised maximum likelihood (GML) analysis that will eliminate the problem of regularity conditions encountered when using MLEs, MOM and L-moments parameter estimation methods.

In general, the search for improved and reliable statistical techniques in long-term flood frequency forecasting should be considered as an ongoing process in EVT and disaster risk reduction.

References

ABDUL-KARIM, M. D. AND CHOWDHURY, J. U. (1995). A comparison of four distributions used in flood frequency analysis in Bangladesh. *Hydrological Sciences Journal*, **40** (1), 55–56.

ABIDA, H. AND ALLOUZE, M. (2008). Probability distribution of flood flows in Tunisia. *Hydrology and Earth System Sciences*, **12**, 703–714.

AHAMMED, F. AND HEWA, G. A. (2012). Development of hydrological tools using extreme rainfall events for Dhaka, Bangladesh. *Water International*, **37** (1), 43–52.

AICH, V., LIERSCH, S., VETTER, T., HUANG, S., TECKLENBURG, J., HOFFMANN, P., KOCH, H., MÜLLER, N., AND HATTERMANN, F. F. (2014). Comparing impacts of climate change on streamflow in four large African river basins. *Hydrology and Earth System Sciences*, **18**, 1305–1321.

ALAM, S. AND KHAN, M. S. M. (2014). Statistical characterization of extreme hydrologic parameters for the pheripheral river system of Dhaka city. *Journal of Water Resources and Ocean Science*, **3** (3), 30–37.

ALEXANDER, W. J. R. (1990). *Flood hydrology for Southern Africa*. SANCOLD, Pretoria, South Africa.

ALEXANDER, W. J. R. (2002). Statistical analysis of extreme floods. *Journal of the South African Institution of Civil Engineering*, **44** (1), 20–25.

AN, Y. AND PANDEY, M. D. (2007). The r largest order statistics model for extreme wind speed estimation. *Journal of Wind Engineering and Industrial Aerodynamics*, **95**, 165–182.

ARNDT, C., JAMES, R. C., AND SIMLER, K. R. (2006). Has economic growth in Mozambique been pro-poor? *Journal of African Economies*, **15**, 571–602.

ARNDT, C. AND SIMLER, K. R. (2007). Consistent poverty comparisons and inference. *Agricultural Economics*, **37**, 133–143.

ARNDT, C. AND THURLOW, J. (2014). Climate uncertainty and economic development: Evaluating the case of Mozambique to 2050. *Climate Change*, doi 10.1007/s10584–014–1294–x.

ARNELL, N. W., BROWN, S., GOSLING, S. N., GOTTSCHALK, P., HINKEL, J., HUNTINGFORD, C., LIOYD-HUGHES, B., LOWE, J. A., NICHOLLS, R. J., OSBORN, T. J., OSBORNE, T. M., ROSE, G. A., SMITH, P., WHEELER, T. R., AND ZELAZOWSKI, P. (2014). The impacts of climate change across the globe: A multi-sectorial assessment. *Climate Change*, doi 10.1007/s10584–014–1281–2.

ARYA, A. S., BOEN, T., AND ISHIYAMA, Y. (2014). *Guidelines for earthquake resistant non-engineered construction*. UNESCO, Paris.

ASANTE, K. O., MACUACUA, R. D., ARTAN, G. A., LIETZOW, R. W., AND VERDIN, J. P. (2007). Developing a flood monitoring system from remotely sensed data for the Limpopo basin. *IEEE Transactions on Geoscience and Remote Sensing*, **45** (6), 1709–1714.

ATROOSH, K. B. AND MUSTAFA, A. T. (2012). An estimation of the probability distribution of Wadi Bana flow in the Abyan Delta of Yemen. *Journal of Agricultural Science*, **4** (6), 80–89.

BAKER (2003). *A bright future for old flows: origins, status and future of paleoflood hydrology. In: Thorndycraft, V.R., Benito, G., Barriendos, M., and Llasat, M.C. (Eds.) (2002): Paleofloods, historical floods and climatic variability: Applications in flood risk assessment*. Proceedings of the PHERA Workshop, 16-19 October, Barcelona. CSIC, Madrid: 13-18.

BALKEMA, A. A. AND DE HAAN, L. (1974). Residual life time at great age. *Annals of Probability*, **2**, 792–894.

BARATTI, E., MONTANARI, A., CASTELLARIN, A., SALINAS, J. L., VIGLIONE, A., AND BEZZI, A. (2012). Estimating the flood frequency distribution at seasonal and annual time scales. *Hydrology and Earth System Sciences*, **16**, 4651–4660.

BAYARRI, M. J. AND BERGER, J. O. (2004). The interplay of Bayesian and frequentist analysis. *Statistical Science*, **19** (1), 58–80.

BBC (2005). *Mourners mark tsunami anniversary*. BBC News. Last accessed: 2015-10-15.
URL: *http://news.bbc.co.uk/2/hi/asia-pacific/4559404.stm*

BEIRLANT, J., DE WET, T., AND GOEGEBEUR, Y. (2006). A goodness-of-fit statistic for Pareto-type-behaviour. *Journal of Computational and Applied Mathematics*, **186**, 99–116.

BEIRLANT, J., GOEGEBEUR, Y., SEGERS, J., DE WAAL, D., AND FERRO, C. (2004). *Statistics of extremes: Theory and applications*. John Wiley & Sons Ltd, West Sussex.

BERNADARA, P., MAZAS, F., KERGADALLAN, X., AND HAMM, L. (2014). A two-step framework for over-the-threshold modelling of environmental extremes. *Natural Hazards and Earth System Sciences*, **14**, 635–647.

BERNING, T. L. (2010). *Improved estimation procedures for a positive extreme value index*. PhD Thesis, SUNScholar, Stellenbosch University, Cape Town.

BLAIN, G. C. AND MESCHIATTI, M. C. (2014). Using multi-parameters distributions to assess the probability of occurrence of extreme rainfall data. *Revista Brasileira de Engenhairia Agricola e Ambiental*, **18** (3), 307–313.

BÜCHER, A. AND SEGERS, J. (2016). On the maximum likelihood estimator for the generalised extreme value distribution. *arXiv:1601.05702v1*.

CHENG, L. AND AGHAKOUCHAK, A. (2014). Nonstationary precipitation intensity-duration-frequency curves for infrastructure design in a changing climate. *Scientific Reports*, **4** (7093), doi 10.1038/srep07093.

CHILUNDO, M., KELDERMAN, P., AND KEEFFE, J. H. O. (2008). Design of a water quality monitoring network for the Limpopo River basin in Mozambique. *Physics and Chemistry of the Earth*, **33** (8-13), 655–665.

CHINOWSKY, P. S., SCHWEIKERT, A. E., STRZEPEK, N. L., AND STRZEPEK, K. (2014). Infrastructure and climate change: A study of impacts and adaptations in Malawi, Mozambique and Zambia. *Climate Change*, doi 10.1007/s10584–014–1219–8.

CII (2009). *Coping with climate change: risks and opportunities for insurers*. The Chartered Insurance Institute, Chapter 5, Market Failure and Climate Change: Climate Change Research Report 2009-2009.

CNN (2015). *At least 66 dead after another powerful earthquake hits Nepal*. Last accessed: 2015-10-20.
URL: *http://edition.cnn.com/2015/05/12/asia/nepal-earthquake/*

COLES, S. (2001). *An introduction to statistical modelling of extreme values*. Springer-Verlag, London.

COLES, S. AND DAVISON, A. (2008). *Statistical modelling of extreme values. Based on 'An introduction to statistical modelling of extreme values' by Stuart Coles, Springer 2001*. Copyright 2008.

COLES, S. G. AND PERICCHI, L. (2003). Anticipating catastrophes through extreme value modelling. *Journal of the Royal Statistical Society C: Applied Statistics*, **52**, 405–416.

COLES, S. G., PERICCHI, L., AND SISSON, S. (2003). A fully probabilistic approach to extreme rainfall modelling. *Journal of Hydrology*, **273**, 35–50.

COLES, S. G. AND POWELL, E. A. (1996). Bayesian methods in extreme value amodelling: A review and new developments. *Int. Statist. Rev.*, **64**, 119–136.

COOLEY, D. (2009). Extreme value analysis and the study of climate change: A commentary on Wigley 1988. *Climatic Change*, **97**, 77–83.

COOLEY, S. D. (2005). *Statistical analysis of extremes motivated by weather and climate studies: Applied and theoretical advances*. PhD Thesis, Department of Applied Mathematics, University of Colorado.

CORDERY, I. AND PILGRIM, D. H. (2000). *The state of the art flood prediction. In: Parker, D. J. (ed) (2000): Floods. Volume II*. Routledge, London.

CUAMBA, B. C. AND MAÚRE, G. A. (2008). *Challenges to managing floods and droughts in transboundary river basins in Mozambique. In: Petermann T. (Ed.) (2008): Towards climate change adaptation-building adaptive capacity in managing African transboundary river basins*. InWEnt, Zschortau, Germany.

CUNNANE, C. (1973). A particular comparison of annual maxima and partial duration series methods of flood frequency prediction. *Journal of Hydrology*, **18**, 257–271.

DE HAAN, L. (1970). *On regular variation and its application to the weak convergence of sample extremes*. Mathematical Centre Tracts 32, Mathematisch Centrum, Amsterdam.

DE HAAN, L. AND FERREIRA, A. (2006). *Extreme value theory: An introduction*. Springer, New York.

DE WET, T. (2004). Statistical adjustment of tidal gauge data. *South African Statistical Journal*, **38** (1), 79–91.

DE WET, T., GOEGEBEUR, Y., AND MUNCH, M. R. (2012). Asymptotically unbiased estimation of second order tail parameter. *Statistics and Probability Letters*, **83** (3), 565–573.

DECARLO, L. T. (1997). On the meaning and use of kurtosis. *Psychological methods*, **2** (3), 292–307.

DODD, E. L. (1923). The greatest and least variate under general limit laws of error. *Transactions of the American Mathematical Society*, **25**, 525–539.

DOMBRY, C. (2015). Existence and consistency of maximum likelihood estimators for the extreme value index within the block maxima framework. *Bernoulli*, **21** (1), 420–436.

DOMBRY, C. AND RIBATET, M. (2015). Functional regular variations, Pareto processes and peaks over threshold. *Statistics and Its Interface*, **8** (1), 9–17.

EASTOE, E. F. AND TAWN, J. A. (2009). Modelling non-stationary extremes with appliaction to surface level ozone. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, **58** (1), 25–45.

FARADA, D., LUCARINI, V., TURCHETTI, G., AND VAIENTI, S. (2011). Numerical convergence of the block-maxima approach to the generalised extreme value distribution. *Journal of Statistical Physiscs*, **145** (5), 1156–1180.

FERNÁNDEZ, N. C., GARCIA, R. R., HERRERA, R. G., PUYOL, D. G., PRESA, L. G., MARTIN, E. H., AND RODRIQUEZ, P. R. (2004). Analysis of the ENSO signal in tropospheric and stratospheric temperatures observed by Microwave Sounding Unit (MSU). *Journal of Climate*, **17**, 3934–3946.

FERREIRA, A. AND DE HAAN, L. (2014). The generalised Pareto process; with a view towards application and simulation. *Bernoulli*, **20** (4), 1717–1737.

FERREIRA, A. AND DE HAAN, L. (2015). On the block maxima method in extreme value theory: PWM estimators. *The Annals of Statistics*, **43** (1), 276–298.

FERRO, C. A. T. AND SEGERS, J. (2003). Inference for clusters of extreme values. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, **65**, 545–556.

FISCHER, S. AND SCHUMANN, A. (2014). *Comparison between classical annual maxima and peak over threshold approach concerning robustness*. SFB 823 Discussion Paper Nr.26/2014.

FISHER, R. A. AND TIPPETT, L. H. C. (1928). Limiting forms of frequency distribution of the largest or smallest member of a sample. *Cambridge Philosophical Society*, **24**, 180–190.

FRÉCHET, M. (1927). Sur la loi de probabilite de Íecart maximum. *Société Polonaise de Mathematique. Annales*, **6**, 93–116.

FULLER, W. E. (1914). Flood flows. *Transactions of the American Society of Civil Engineers*, **77**, 564–617.

GAIONI, E., DEY, D., AND RUGGER, F. (2010). Baysian modelling of flash floods using generalised extreme value distribution with prior elicitation. *Chilean Journal of Statistics*, **1** (1), 75–90.

GAUME, E., GAÁL, L., VIGLIONE, A., SZOLGAY, J., KOHNOVA, S., AND BL¨SCHL, G. (2010). Bayesian MCMC approach to regional flood frequency analyses involving extraordinary flood events at ungauged sites. *Journal of Hydrology*, **394**, 101–117.

GELMAN, A., CARLIN, J., STERN, H., AND RUBIN, D. (2004). *Bayesian data analysis. $2^{nd} Edition$.* Chapman and Hall, Boca Raton, FL.

GEMAN, S. AND GEMAN, D. (1984). Stochastic relaxation, gibss ditributions and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **8**, 721–741.

GENCAY, R. AND SELCUK, K. (2004). Extreme value theory and value-at-risk: relative performance in emerging markets. *International Journal of Forecasting*, **20**, 287–303.

GNEDENKO, B. V. (1943). Sur la distribution limite du terme maximum ́dune serie aleatoire. *Annals of Mathematics*, **44**, 423-453. Translated and reprinted in: *Breakthroughs in Statistics*. Eds. Kotz, S. and Johnson, N. L. (1992). *Springer Verlag*, **1**, 195–225.

GOEGEBEUR, Y., BEIRLANT, J., AND DE WET, T. (2008). Linking pareto-tail kernel goodness-of-fit statistics with tail index at optimal threshold and second order estimation. *REVSTAT Statistical Jornal*, **6**, 51–69.

GOHIL, R. B. AND CHOWDHARY, D. R. (2013). Study of flood frequency for Tan River at station Amba, Gujarat. *PARIPEX - Indian Journal of Research*, **12**, 132–133.

GRIFFITH, A. A. (1920). The phenomena of rupture and flow in solids. *Philosophical Transactions of the Royal Society of London, Series A*, **221**, 163–198.

GUMBEL, E. J. (1941). The return period of flood flows. *Annals of Mathematical Statistics*, **12**, 163–190.

GUMBEL, E. J. (1954). *Statistical theory of extreme values and some practical applications*. Applied Mathematics Series 33. U.S. Department of Commerce, National Bureau of Standards.

GUMBEL, E. J. (1958). *Statistics of extremes*. Columbia University Press, New York.

HAKALA, A. AND PEKONEN, L. (2008). *The impact of the irrigation system and agricultural production on water quality in Chokwe Irrigation Scheme*. Thesis, Savonia University, Finland in collaboration with Universidade Eduardo Mondlane, Mozambique.

HAKTANIR, T., COBANER, M., AND GORKEMLI, B. (2013). Assessment of right-tail prediction ability of some distributions by Monte Carlo analyses. *Journal of Hydrologic Engineering*, **8** (5), 499–517.

HAMIDIEH, K. (2008). *Topics in statistical modelling and estimation of extremes and their dependence*. PhD Thesis, University of Michigan, UMI Number 3343082.

HASTINGS, W. K. (1970). Monte carlo sampling methods using markov chains and their applications. *Biometrika*, **57**, 97–109.

HEFFERNAN, J. E. AND STEPHENSON, A. G. (2015). *Ismev: R package version 1.40*.
URL: *http://www.ral.ucar.edu/ericg/softextreme.php*

HOSKING, J. R. M. AND WALLIS, J. R. (1997). *Regional frequency analysis: An approach based on L-moments*. Cambridge University Press, Cambridge.

IBRAHIM, J. AND CHEN, M. (2000). Power prior distributions for regression models. *Statistical Sciences*, **15**, 46–60.

IPCC (2012). *Managing the risks of extreme events and disasters to advance climate change adaptation (SREX)*. A special Report on Working Groups I

and II of the Intergovernmental Panel on Climate Change, Cambridge University Press, UK and USA.

IZINYON, O. C. AND EHIOROBO, J. O. (2015). L-moments method for flood frequency analysis of River Owan at Owan in Benin Owena River basin in Nigeria. *Current Advances in Civil Engineering (CACE)*, **3** (1), 1–10.

JACKSON, J. (2013a). *150,000 displaced as Mozambique floods spread*. Modern Ghana. Last accessed: 2015-10-15.
URL: *http://www.modernghana.com/news/442083/1/150000-displaced-as-mozambique-floods-spread.html*

JACKSON, J. (2013b). *Mozambique floods spur roof births, ruin and diarrhoea*. Agence France-Presse (AFP). Last accessed: 2015-10-15.
URL: *http://reliefweb.int/report/mozambique/mozambique-floods-spur-roof-births-ruin-and-diarrhoea*

JÄGER, J., FRÜHMANN, J., GÜNBERGER, S., AND VAG, A. (2009). *Environmental change and forced migration scenarios project synthesis report*. May 14, Deliverable 044468.

JAKOB, D., KAROLY, D. J., AND SEED, A. (2011). Non-stationary in daily and sub-daily intense rainfall - Part I: Sydney, Australia. *Natural hazards and Earth System Sciences*, **11**, 2263–2271.

JONATHAN, P., RANDELL, D., WU, Y., AND EWANS, K. (2014). Return level estimation from non-stationary spatial data exhibiting multidimensional covariate effects. *Ocean Engineering*, **88**, 520–532.

KACHROO, R. K., MKHANDI, S. H., AND PARIDA, B. P. (2000). Flood frequency analysis of Southern Africa: I. Delineation of homogeneous regions. *Hydrological Sciences Journal*, **45** (3), 437–447.

KASS, R. E., WASSERMAN, L., AND PARIDA, B. P. (1996). Formal rules for selecting prior distrubutions: A review and annoted bibliograpgy. *Journal of the American Statistical Association*, **91**, 343–370.

KATZ, R. W. (2010). Statistics of extremes in climate change. *Climatic Change*, **100**, 71–76.

KATZ, W. K., PARLANGE, M. B., AND NAVEAU, P. (2002). Statistics of extremes in hydrology. *Advances in Water Resources*, **25**, 1287–1304.

KHODABIN, M. AND AHMADABADI, A. (2010). Some properties of the generalised gamma distribution. *Mathematical Sciences*, **4** (1), 9–28.

KHULUSE, S., DAS, S., DEBBA, P., AND ELPHINSTONE, C. (2009). *What can we infer from beyond the data? The statistics behind the analysis of risk events in the context of environmental studies*. In: Proceedings of the 2nd African Digital Scholarship and Curation Conference, 12-14 May, CSIR Conference Centre, Pretoria.

KOCH, K.-R. (2007). *Introduction to Bayesian statistics. 2^{nd} Edition*. Springer, Berlin.

KOCHANEK, K., RENARD, B., ARNAUD, P., AUBERT, Y., LANG, M., CIPRIANI, T., AND SAUQUET, E. (2013). A data-based comparison of flood frequency analysis methods used in France. *Natural Hazards and Earth System Sciences: Discussions*, **1**, 4445–4479.

KOCHAR, S. AND TORRADO, N. (2015). On stochastic comparisons of largest order statistics in the scale model [currently published by: Portland State University, PDXScholar, Mathematics and Statistics Faculty Publications]. *To appear in: Communications in Statistics - Theory and Methods*.

KOUTSOYIANNIS, D. (2004). Statistics of extremes and estimation of extreme

rainfall: I. Theoretical investigation. *Hydrological Sciences-Journal-des Sciences Hydrologiques*, **49** (4), 575–590.

KRON, W., STEUER, M., LÖW, P., AND WIRTZ, A. (2012). How to deal properly with a natural catastrophe database-analysis of flood losses. *Natural Hazards and Earth System Sciences*, **12**, 535–550.

KUCZERA, G. (1999). Comprehensive at-site flood frequency analysis using Monte Carlo Bayesian inference. *Water Resources Research*, **35** (5), 1551–1557.

KWON, H.-H., BROWN, C., AND LALL, U. (2008). Climate informed flood frequency analysis and prediction in Montana using hierarchical Bayesian modelling. *Geophysical Reesearch Letters*, **35**, L05404.

LEADBETTER, M., LINDGREN, G., AND ROOTZÉN, H. (1983). *Extremes and related properties of random sequences and processes*. Springer Verlag, New York.

LI, Z. F. AND TOMKINS, R. J. (1991). Complete stability of large order statistics. *Journal of Theoretical Probability*, **4** (1), 213–221.

LÓPEZ, J. AND FRANCÉS, F. (2013). Non-stationary flood frequency analysis in continental Spanish rivers, using climate and reservoir indices as external covariates. *Hydrology and Earth System Sciences*, **17**, 3189–3203.

LUCIO, F. D. F. (2007). *Predictability of extreme events associated with the inter-annual variability of rainfall. In: Agricultural Water Management: A critical factor in reduction of poverty and hunger*. Proceedings of the 2nd Regional Workshop on Agricultural Water Management in Eastern and Southern Africa, 18-22 September 2006, Maputo.

MABASO, E. AND MANYENA, S. B. (2013). Contigency planning in Southern

Africa: Events rather than processes. *Jàmbá: Journal of Disaster Risk Studies*, **5** (95), doi:10.4102/jamba.v5i1.95.

MABOTE, G. M. A. (2011). *Conception of a simplified model for the monitoring of flood wave: A case study of the Limpopo River basin*. MSc Thesis, Department of Agraria, Università Studi di Firenze, Italy.

MACDONALD, A. (2012). *Extreme value mixture models with medical and industrial applications*. PhD Thesis, University of Canterbury, New Zealand.

MACHIWAL, D. AND MADAN, K. J. (2008). Evaluation comparative de tests statistiques pour l'analyse de series temporelles: Application á des series temporelles hydrologiques [Comparative evaluation of statistical tests for time series analysis: Application to hydrological time series]. *Hydrological Sciences Journal*, **53** (2), 353–366.

MAGADIA, J. (2010). *Value-at-Risk modelling via the peaks-over-threshold approach*. Annual BSP-UP Professional Chair Lectures, 15-17 February 2010, Bangko Sentral ng Pilipinas, Malate, Manila.

MANUEL, I. R. AND VICENTE, W. M. (2002). The problem of flooding in Mozambique; the catastrophe of the year 2000. *Africa Geosciences Review*, **9** (4), 401–407.

MARAUN, D., RUST, H. W., AND OSBORN, T. J. (2009). The annual cycle of heavy precipitation across the United Kingdom: A model based on extreme value statistics. *International Journal of Climatology*, **29**, 1731–1744.

MAREE, M. (2011). *ENHANS International Workshop aims to reduce disaster risk due to extreme natural hazards in Africa*.
URL: *http://www.enhans.org*

MARTINS, E. S. AND STEDINGER, J. R. (2000). Generalised maximum-

likehood generalized extreme-value quantile estimators for hydrologic data. *Water Resources Research*, **36**, 737–744.

MCMAHON, T. A., VOGEL, R. M., PEEL, M. C., AND PEGRAM, G. G. S. (2007). Global streamflows - Part (i): Characteristics of Annual Streamflows. *Journal of Hydrology*, **347**, 243–259.

MEHRANNIA, H. AND POKGOHAR, A. (2014). Using easyFit software for goodness-of-fit test and data generation. *International Journal on Mathematical Archive*, **5**, 118–124.

MEJZLER, D. G. (1949). On a theorem of B. V. Gnedenko. *Sbornik Trudov Inst. Mat. Akao. Nauk Ukrain. SSR*, **12**, 31–35.

MKHANDI, S. H., KACHROO, R. K., AND GUNASEKARA, T. A. G. (2000). Flood frequency analysis of Southern Africa: II. Identification of regional distributions. *Hydrological Sciences Journal*, **45** (3), 449–464.

MÉLICE, J. L. AND REASON, C. J. C. (2007). Return period of extreme rainfall at George, South Africa. *South African Journal of Science*, **103** (11-12), 499–501.

MOHAMED, E. A. (2014). Comparing Africa's shared river basins – The Limpopo, Orange, Juba and Shabelle basins. *Universal Journal of Geoscience*, **2** (7), 200–211.

MONDLANE, A. V., HANSSON, K., AND POPOV, O. (2013). Analysis of mathematical models and their application to extreme events. *World Academy of Science, Engineering and Technology; International Journal of Mathematical, Computational Science and Engineering*, **7** (4), 378–387.

MUDAVANHU, C. (2014). The impact of flood disaster on children education in Muzarabani District, Zimbabwe. *Jàmbá: Journal of Disaster Risk Studies*, **6** (1), 1–8.

MUJERE, N. (2011). Flood frequency analysis using the gumbel distribution. *International Journal on Computer Science and Engineering*, **3**, 2774–2778.

MUNICHRE (2011). *Topics in geo-natural catastrophes 2010: Analyses, assessments, positions*. Munich Reinsurance, Munich.

MUNICHRE (2013). *Floods dominate natural catastrophe statistics in first half of 2013*. Press Release, 9 July 2013. Munich Reinsurance, Munich.

MUSIYA, T. (2013). *UN: Mozambique floods displace 150k, leave 38 dead*. Associated Press CHOKWE, Mozambique (AP).
URL: *http://article.wn.com/view/2013/01/29/UN*

NEYKOV, N. M., NEYTCHEV, P. N., AND ZUCCHINI, W. (2014). Stochastic daily precipitation model with a heavy-tailed component. *Natural Hazards and Earth System Sciences*, **14**, 2321–2335.

NGUYEN, V. T. V. (2009). Recent advances in statistical modelling of extreme rainfalls and floods. *International journal on Hydropower and Dams*, **16** (2), 65–70.

NORTJE, J. (2010). Estimation of extreme flood peaks by selective statistical analyses of relevant flood peak data within similar hydrological regions. *Journal of South African Institute of Civil Engineers*, **52** (2), 48–57.

OCHA-ROSA (2011). *Southern Africa floods and cyclones update 2010-2011*. UN Organisation for the Coordination of Humanitarian Affairs - Regional Office for Southern Africa.
URL: *http://unocha.org/rosa*

OCHA-ROSA (2012). *Southern Africa floods and cyclones update. Situation Report Number 3, 05 April 2012*. UN Organisation for the Coordination of Humanitarian Affairs - Regional Office for Southern Africa.
URL: *http://unocha.org/rosa*

OCHA-ROSA (2013). *Southern Africa floods. Situation Report Number 4, 01 February 2013*. UN Organisation for the Coordination of Humanitarian Affairs - Regional Office for Southern Africa.
URL: *http://unocha.org/rosa*

OLOFINTOYE, O. O., SULE, B. F., AND SALAMI, A. W. (2009). Best-fit probability distribution model for peak daily rainfall of selected cities in Nigeria. *New York Science Journal*, **2** (3), 1–12.

OTTO, F. E. L., ROSIER, S. M., ALLEN, M. R., MASSEY, N. R., RYE, C. J., AND QUINTANA, J. I. (2014). Attribution analysis of high precipitation events in summer in england and wales over the last decade. *Climate Change*, doi 10.1007/s10584–014–1095–2.

PICKANDS, J. (1975). Statistical inference using extreme order statistics. *Annals of Statistics*, **3**, 119–131.

PRETORIUS, T. B. (2007). *Inferential data analysis: Hypothesis testing and decision-making*. Reach, Wandsbeck, South Africa.

RAJARAM, L. (2006). *Statistical models in environmental and life sciences*. PhD Thesis, USF Graduate School Theses and Dissertations, University of South Florida, paper 2668.
URL: *http://scholarcommons.usf.edu/etd/2668*

REDVERS, L. (2009). *Southern Africa: Floods-breaking the cycle*. Inter Press Service (IPS) News Agency.
URL: *http://ipsnews.net/africa/nota.asp?idnews=47579*

REID, P. (2000). *SOI/ENSO and their influence*. Climate Research Unit, Copyright 2000.

REIS, D. S. AND STEDINGER, J. R. (2005). Bayesian MCMC flood frequency analysis with historical information. *Journal of Hydrology*, **313**, 97–116.

REISS, R. D. AND THOMAS, M. (2007). *Statistical analysis of extreme values with application to insurance, finance, hydrology and other fields. 3rd Edition*. Birkhäuser Verlag, Basel.

RENARD, B., GARRETA, V., AND LANG, M. (2006). An application of Bayesian analysis and Markov chain Monte Carlo methods to the estiamtion of a regional trend in annual maxima. *Water Resources Research*, **42**, W12422.

RIBATET, M. A. (2006). *A user's guide to the POT package*. Version 1.4.
URL: *http://www.cran.r-project.org/*

RIBEREAU, P., GUILLOU, A., AND NAVEAU, P. (2008). Estimating return levels from maxima of non-stationary random sequences using the generalised pwm method. *Nonlinear Processes in Geophysics*, **15**, 1033–1039.

RICCI, V. (2005). *Fitting distributions with R*.
URL: *http://www.researchgate.net/publication/228791072/*

ROSTAMI, R. (2013). Regional flood frequency analysis based on L-moment approach (Case study West Azarbayjan basins). *Journal of Civil Engineering and Urbanisation*, **3** (3), 107–113.

ROWINSKI, P. M. AND STRUPCZEWSKI, W. G. (2001). A note on the applicability of log-Gumbel and log-logistic probability distributions in hydrological analysis: I. Known pdf. *Hydrological Sciences-Journal-des Sciences Hydrologique*, **47** (1), 107–122.

SAIDI, H., CIAMPITTIELLO, M., DRESTI, C., AND GHIGLIERI, G. (2013). Observed variability and trends in extreme rainfall indices and peaks-over-threshold series. *Hydrology and Earth System Sciences*, **10**, 6049–6079.

SCARROTT, C. AND MACDONALD, A. (2012). A review of extreme value threshold estimation and uncertainty quantification. *Statistical Journal*, **10** (1), 33–60.

SHABRI, A. AND ARIFF, N. M. (2009). Frequency analysis of maximum daily rainfalls via l-moment approach. *Sains Malaysiana*, **38** (2), 149–158.

SHO, K., IWASAKI, S., AND TOMINAGA, A. (2000). Effect of introducing uncertain historical hydrologic data on quantiles estimation accuracy. *Journal of Hydroscience and Hydraulic Engineering*, **18** (1), 45–52.

SIGAUKE, C. (2014). *Modelling electricity demand in South Africa*. PhD Thesis, University of Free State.

SINGO, L. R., KUNDU, P. M., ODIYO, J. O., MATHIVHA, F. I., AND NKUNA, T. R. (2012). *Flood frequency analysis of annual maximum stream flows for Luvuvhu River catchment, Limpopo Province, South Africa*. 16^{th} SANCIAHS National Hydrology Symposium, 1-3 October 2012, University of Pretoria, Pretoria. Last accessed: 2016-02-03.
URL: *http://www.ru.ac.za/static/institutes/SANCIAHS/2012/*

SINGPURWALLA, N. AND SMITH, R. (2006). A conversation with B. V. Gnedenko. *Reliability: Theory and Applications*, **1** (January 2006).

SMAKHTIN, V. (2014). *Managing floods and droughts through innovative water storage solutions*. In: M. Stal, S. Good and W. Ammann (Eds.) (2014): IDRC Davos 2014 programme and short abstracts, 81-251. Davos, Switzerland. Last accessed: 2015-03-20.
URL: *http://idrc.info/outcomes/conference-proceedings/*

SMITH, R. (1987). Estimating tails of probability distributions. *The Annals of Statistics*, **15** (3), 1174–1207.

SMITHERS, J. C. (2012). Methods for design flood estimation in South Africa. *Water SA*, **38** (4), 633–646.

SMITHERS, J. C., SCHULZE, R. E., PIKE, A., AND JEWITT, G. P. W. (2001).

A hydrological perspective of the February 2000 floods: A case study in the Sabie River catchment. *Water SA*, **27** (3), 325–332.

SOARES, C. G. AND SCOTTO, M. G. (2004). Application of the r largest order statistics for long-term predictions of significant wave height. *Coastal Engineering*, **51**, 387–394.

SOUTHWORTH, H. AND HEFFERNAN, J. E. (2013). *texmex: Statistical modelling of extreme values*. R package version 2.1.
URL: *http://cran.r-project.org/*

SPALIVEIRO, M., DE DAPPER, M., AND MALO, S. (2014). Flood analysis of the Limpopo River basin through past evolution reconstruction and a geomorphological approach. *Natural Hazards and Earth System Sciences*, **14**, 2027–2039.

STAL, M., GOOD, S., AND AMMANN, W. (2014). *International disaster and risk conference Davos 2014 programme and short abstracts collection: Integrative risk management-The role of science, technology and practice*. In: IDRC Davos 2014 programme and short abtracts, 81-251. Davos, Switzerland. Last accessed: 2016-03-15.
URL: *http://idrc.info/outcomes/conference-proceedings/*

STEPHENSON, A. G. AND RIBATET, M. A. (2006). *A user's guide to the evdbayes package*. Version 1.1.
URL: *http://www.cran.r-project.org/*

SUKLA, M. K., MUNGARAJ, A. K., AND SAHOO, L. N. (2014). An investigation on the stochastic modelling of daily rainfall amount in the Mahanadi Delta region, India. *Research Journal of Mathematical and Statistical Sciences*, **2** (9), 1–8.

TAWN, J. A. (1988). An extreme value theory model for dependent observations. *Journal of Hydrology*, **101**, 227–250.

THOMAS, V., ALBERT, J. R. G., AND HEPBURN, C. (2014). Contributors to the frequency of intense climate disaters in asia-pacific countries. *Climate Change*, **126**, 381–398.

TIERNEY, L. (1994). Markov chains for exploring posterior distributions. *The Annals of Statistics*, **22** (4), 1701–1728.

TOWLER, E., RAJAGOPALAN, B., GILLELAND, E., SUMMERS, R. S., YATES, D., AND KATZ, R. W. (2010). Modelling hydrologic and water quality extremes in a changing climate: A statistical approach based on extreme value theory. *Water Resources Research*, **46**, W11504.

TRAMBLAY, Y., AMOUSSOU, E., DORIGO, W., AND MAHE, G. (2014). Flood risk under future climate in data sparse regions: Linking extreme value models and flood generating processes. *Journal of Hydrology*, **519**, 549–558.

UN (2005). *United Nations World Conference on disaster reduction, 18-22 January 2005*. Report on World Conference on Disaster Reduction, A/CONF.206/6, 16 March 2005, Kobe, Japan.
URL: *http://www.unisdr.org/2005/wcdr/*

UNDP (2010). *Report on the Millennium Development Goals-Mozambique 2010*. UNDP Mozambique.

UNDP (2011). *Workshop on national flood risk assessment and mapping in Mozambique (Final Report)*. UNDP Mozambique, INGC, GRIP, Hepia/PMSA. 31 October – 4 November 2011. Maputo, Mozambique.

UNDP (2015). *About Mozambique*. UNDP Mozambique.
URL: *http://www.mz.undp.org/content/mozambique/en/home/countryinfo/*

UNEP-UNCHS (Sa). *Mozambique 2000 Floods*. UNEP/UNCHS Habitat Joint Mission.

URL: *http://reliefweb.int/report/mozambique/joint-unephabitat-mission-contributes-work-un-disaster-assessment-and-coordination*

VAN DEN BRINK, H. W. AND KÖNNEN, G. P. (2009). Estimating 10000-year return values from short time series. *International Journal of Climatology*, **31** (1), 115–126.

VAN OGTROP, F. F., HOEKSTRA, A. J., AND VAN DER MEULEN, F. (2005). Flood management in the lower Incomati River basin: Two alternatives. *Journal of the American Water Resources Association*, **41** (3), 607–619.

VANEM, E. (2015a). Non-stationary extreme value models to account for trends and shifts in the extreme wave climate due to climate change. *Applied Ocean Research*, **52**, 201–211.

VANEM, E. (2015b). Uncertainties in extreme value modelling of wave data in a climate change perspective. *Journal of Ocean Engineering and Marine Energy*, **1**, 339–359.

VASILIADES, L., GALIATSATOU, P., AND LOUKAS, A. (2013). Modelling hydrological extremes under non-stationary conditions using climate covariates. *Geophysical Research Abstracts*, **25**, EGU2013–12034.

VASILIADES, L., GALIATSATOU, P., AND LOUKAS, A. (2015). Nonstationary frequency analysis of annual maximum rainfall using climate covariates. *Water Resources Management*, **29** (2), 339–358.

VELASCO, M., VERSINI, P. A., CABELLO, A., AND BARRERA-ESCOLA, A. (2013). Assessment of flash floods taking into account climate change scenarios in the Liobregat River basin. *Natural Hazards and Earth System Sciences*, **13**, 3145–3156.

VERDON-KIDD, D. C. AND KIEM, A. S. (2015). Non-stationary in annual max-

ima rainfall across australia - implications for intensity-frequency-duration (ifd) relationships. *Statistical Journal*, **12**, 3449–3475.

VIDAL, I. (2014). A Bayesian analysis of the gumbel distribution: An application to extreme rainfall data. *Stochastic Environmental Research and Risk Assessment*, **28** (3), 571–582.

VIGLIONE, A., MERZ, R., SALINAS, J. L., AND BLÖSCHL, G. (2013). Flood frequency hydrology: 3. A Bayesian analysis. *Water Resources Research*, **49**, 675–692.

VOGEL, R. M., JR THOMAS, W. O., AND MCMAHON, T. A. (1993a). Flood-flow frequency model selection in Southwestern United States. *Journal of Water Resources Planning and Management*, **119** (3), May/June.

VOGEL, R. M., MCMAHON, T. A., AND CHIEW, F. H. S. (1993b). Flood-flow model selection in australia. *Journal of Hydrology*, **146**, 421–449.

VON BORTKIEWICZ, L. (1922). Variationsbreite und mittlerer fehler. *Sitzungsberichte Berliner Math. Gesellschaft*, **21**, 3–11.

VON MISES, R. (1923). Uber die variationsbreite einer Beobachtungsreihe. *Sitzungsberichte Berliner Math. Gesellschaft*, **22**, 3–8.

VON MISES, R. (1954). La distribution de la plus grande de n valeurs. *American Mathematical Society, Providence, RI, Selected Papers*, **2**, 271–294. This paper was the reproduction of Mises, R. von's original paper published in Revue Mathimatique de 1'Union Interbalkanique, 1, 141-160 (1936).

WIKIPEDIA (2015). *Nepal earthquake*. Wikipedia.
URL: *http://en.wikipedia.org/wiki/2015*

WISNER, B., BLAIKIE, P., CANNON, T., AND DAVKIES, I. (2004). *At risk-natural hazards, people's vulnerability and disasters*. Routledge, Wiltshire.

WMO (1999). *WMO statement on the status of the global climate in 1998*. WMO no. 896. World Meteorological Organisation, Geneva.

WMO (2009). *Guide to hydrological practices, Volume II: Management of water resources and application of hydrological practices*. WMO no. 168. World Meteorological Organisation, Geneva.

WMO (2012). *Limpopo River basin: A proposal to improve the flood forecasting and early warning system*. World Meteorological Organization (WMO) with the support of Limpopo Water Course Secretariat and the riparian states of Botswana, Mozambique, South Africa and Zimbabwe.

WMO (2013). *Global climate 2001-2010: A decade of climate extremes*. Summary Report, WMO no. 119. World Meteorological Organisation, Geneva.

YILMAZ, A. G., HOSSAIN, I., AND PERERA, B. J. C. (2014). Effect of climate change and variability on extreme rainfall intensity-frequency-duration relationships: a case study of Melbourne. *Hydrology and Earth System Sciences*, **18**, 4065–4076.

ZHAO, P. AND BALAKRISHNAN, N. (2015). Comparisons of large order statistics from multiple-outlier gamma models. *Methodology and Computing in Applied Probability*, **17** (3), 617–645.