

https://dspace.org/

Modelling heavy rainfall over time and space

Khuluse, Sibusisiwe Audrey

2011-06-06
http://hdl.handle.net/10539/10005

Downloaded from a DSpace repository. https://dspace.org/



Faculty of Science
School of Statistics and Actuarial Science

Modelling Heavy Rainfall Over Time and Space

A research report submitted in partial fulfilment of the requirements of the Degree for Master of Science in Mathematical Statistics at the University of the Witwatersrand by

Sibusisiwe Audrey Khuluse

Supervisor: Mr Mark Dowdeswell
Co-supervisor: Dr Pravesh Debba

2010

Acknowledgements

The task of starting and eventually completing a research project is never easy. One benefits largely from interactions with senior colleagues and from the insights of supervisors. I want to express my sincerest gratitude to my supervisors, Mark Dowdeswell and Dr Pravesh Debba. You gave me the freedom to explore my ideas using your insight and experience to help shape my ideas into a structured scientific argument. I remain grateful for the opportunities and relentless support that they afforded me.

In the quest to complete this research, I received an opportunity to be a visiting student at the University of Twente – Faculty of Geoinformation Science and Earth Observation (ITC) in the Netherlands and the Department of Mathematics and Computer Science at the University of Southern Denmark (USD). I express my sincerest gratitude to Professor Alfred Stein ITC for his supervision during the period I spent in the Earth Observation Science department. I benefited from his knowledge of geostatistics and statistics in general. Dr David Rossiter (ITC) made it simple to learn geostatistical analysis in R, which was crucial for the second stage of the project. His comments and guidance with regards to the project were beneficial. I am grateful to Professor Yuri Goegebuer of USD for being a hospitable host and an insightful guide. His comments on the project were valuable and the discussion on ideas for further research was revitalizing.

This research project would have been difficult to complete had it not been for the resources which were made available to me. Firstly I want to acknowledge the South African Weather Service for the data. I acknowledge the Council for Scientific and Industrial Research (CSIR) for the resources spent towards my studies. I acknowledge TATA Africa for the scholarship which partially financed the three months I spent as a visiting student at ITC and the University of Southern Denmark. I want to thank ITC, in particular the Earth Observation Science department for financing my stay in Enschede, Netherlands.

I want to thank my colleagues in the Logistics and Quantitative Methods competence area within the Built Environment unit of the CSIR, for their encouragement and ideas during our 'morning coffee gatherings'. To my friends and family, your enthusiasm and confidence in my abilities kept me going at times when I felt despondent. To my mother, you saw me start this journey, but the Lord had a better place for you to go before I reached the end. In the toughest times, thoughts of you kept me going. Ngiyabonga!

Abstract

Extreme Value Theory finds application in problems concerning low probability but high consequence events. In hydrology the study of heavy rainfall is important in regional flood risk assessment. In particular, the N-year return level is a key output of an extreme value analysis, hence care needs to be taken to ensure that the model is accurate and that the level of imprecision in the parameter estimates is made explicit.

Rainfall is a process that evolves over time and space. Therefore, it is anticipated that at extreme levels the process would continue to show temporal and spatial correlation. In this study interest is in whether any trends in heavy rainfall can be detected for the Western Cape. The focus is on obtaining the 50-year daily winter rainfall return level and investigating whether this quantity is homogenous over the study area. The study is carried out in two stages.

In the first stage, the point process approach to extreme value theory is applied to arrive at the return level estimates at each of the fifteen sites. Stationarity is assumed for the series at each station, thus an issue to deal with is that of short-range temporal correlation of threshold exceedances. The proportion of exceedances is found to be smaller (approximately 0.01) for stations towards the east such as Jonkersberg, Plettenbergbay and Tygerhoek. This can be attributed to rainfall values being mostly low, with few instances where large amounts of rainfall were observed. Looking at the parameters of the point process extreme value model, the location parameter estimate appears stable over the region in contrast to the scale parameter estimate which shows an increase towards in a south easterly direction. While the model is shown to fit exceedances at each station adequately, the degree of uncertainty is large for stations such as Tygerhoek, where the maximum observed rainfall value is approximately twice as large as the high rainfall values. This situation was also observed at other stations and in such cases removal of these high rainfall values was avoided to minimize the risk of obtaining inaccurate return level estimates. The key result is an N-year rainfall return level estimate at each site. Interest is in mapping an estimate of the 50-year daily winter rainfall return level, however to evaluate the adequacy of the model at each site the 25-year return level is considered since a 25 year return period is well within the range of the observed data. The 25-year daily winter rainfall return level estimate for Ladismith is the smallest at 22.42 mm. This can be attributed to the station's generally low observed winter rainfall values. In contrast, the return level estimate for Tygerhoek is high, almost six times larger than that of Ladismith at 119.16 mm. Visually design values show differences between sites, therefore it is of interest to investigate whether these differences can be modelled.

The second stage is the geostatistical analysis of the 50-year 24-hour rainfall return level.

The aim here is to quantify the degree of spatial variation in the 50-year 24-hour rainfall return level estimates and to use that association to predict values at unobserved sites within the study region. A tool for quantifying spatial variation is the variogram model. Estimation of the parameters of this model require a sufficiently large sample, which is a challenge in this study since there is only fifteen stations and therefore only fifteen observations for the geostatistical analysis. To address this challenge, observations are expanded in space and time and then standardized and to create a larger pool of data from which the variogram is estimated. The obtained estimates are used in ordinary and universal kriging to derive the 50-year 24-hour winter rainfall return level maps. It is shown that 50-year daily winter design rainfall over most of the Western Cape lies between 40 mm and 80 mm, but rises sharply as one moves towards the east coast of the region. This is largely due to the influence of large design values obtained for Tygerhoek. In ordinary kriging prediction uncertainty is lowest around observed values and is large if the distance from these points increases. Overall, prediction uncertainty maps show that ordinary kriging performs better than universal kriging where a linear regional trend in design values is included.

Keywords: Extreme value theory, extreme value modelling, Poisson point processes, threshold exceedance models, return level estimates, return level maps, geostatistics, ordinary kriging, universal kriging, modelling heavy rainfall.

Contents

1 Introduction **1**

1.1 Rationale 1

1.2 Modelling Rare Events 3

1.3 Research Objective 4

1.4 Structure of the Research Report 5

2 Literature Survey **6**

2.1 Studies of Rainfall 6

2.2 Classical Extreme Value Theory 8

2.3 The Threshold Exceedance Approach 10

2.4 Quantile Estimation 12

2.5 Inference 13

2.5.1 Maximum Likelihood Inference 14

2.5.2 Alternatives to Maximum Likelihood Inference 15

2.6 Modelling Issues in Extreme Value Analysis 16

2.6.1 Extreme Value Analysis when Data are Incomplete 16

2.6.2 Considering Temporal Characteristics 17

2.6.3 Spatial Variation of Extreme Values 18

2.7 Computation 20

3 Methods **22**

3.1 Defining Point Processes 22

3.2 Important Concepts 24

3.2.1 Stationarity 24

3.2.2 The Intensity of a Point Process 25

3.3 Poisson Point Processes 25

3.3.1 Maximum Likelihood Estimation of the Poisson Process 27

3.4 Connection between the Poisson Process and Extreme Value Theory 29

3.4.1 The Poisson Approximation to Extremes 29

3.5 Inference for the Point Process Model 32

3.5.1 Selecting the Appropriate Threshold 33

3.5.2 Estimation in the Point Process Model 36

3.6 Geostatistical Model for the Return Levels 40

3.6.1 Estimating the Semivariance 40

3.6.2 Spatial Prediction 42

4 Applying the Point Process Extreme Value Approach **45**

4.1 Description of the Rainfall Data 45

4.2 Site-wise Threshold Selection 54

4.3 Site-wise Point Process Extreme Value Models 66

5 Geostatistical Model of Rainfall Return Levels **75**

5.1 Exploratory Spatial Analysis of the Return Levels 75

5.2 Model for Spatial Correlation of Rainfall Return Levels 81

5.3 The Return Level Maps 85

6 Concluding Remarks **87**

6.1 Recommendations 89

A Appendix **90**

A.1 Threshold Sensitivity Analysis 90

A.2 Ultimate Models 108

List of Tables

1.1 Summary on historical disasters for South Africa 1900 – 2010 3

4.1 Details on the rainfall data obtained from South African Weather Services . 47

4.2 Sensitivity analysis: Threshold ranges and initial values for the numerical optimization of the log-likelihood function for all sites 54

4.3 Results of the point process model fit for CPT Int.and Tygerhoek. 66

4.4 Return level estimates – CPT Int.and Tygerhoek, with 95% profile confidence intervals 67

4.5 Parameter estimates for the extreme value model fitted at each site. Bold values in the column of selected thresholds u indicates series that have been declustered. The maximum likelihood estimates for the location (μ), scale (σ), shape (ξ) and 25 year return level estimates (x_{25}) are accompanied by the standard errors and confidence intervals in brackets 70

5.1 Investigating the possibility of linear spatial trend in 25 year return level estimates 77

5.2 Replicates of the design values in pseudo-time 81

5.3 Average design values for each return period – 25 to 50 years 81

5.4 Variogram model parameter estimates for the pooled design values 83

5.5 Variogram model parameter estimates for the pooled residuals, after removal of regional trend 83

A.1 Results from fitting point process model 109

List of Figures

4.1 Location of the fifteen weather stations in the Western Cape region 46

4.2 Daily rainfall values (winter): Atlantis – Malmesbury 50

4.3 Daily rainfall values (winter): Molteno – Wellington 51

4.4 Comparison of monthly profiles of daily rainfall values from predominantly winter and all-season rainfall regions of the Western Cape 52

4.5 Descriptive plots for the series at each station 53

4.6 Threshold selection – mean residual life plots for Cape Town Int. and Tygerhoek 55

4.7 Analyzing the stability of the scale and shape parameters to changes in the threshold value 59

4.8 Sensitivity of the point process extreme value model characteristics to the threshold – CPT Int.. 60

4.9 Sensitivity of the point process extreme value model characteristics to the threshold for Tygerhoek 61

4.10 Sensitivity of the return level estimates to threshold level for CPT Int. . . . 62

4.11 Sensitivity of the return level estimates to threshold level for Tygerhoek . . . 63

4.12 Level of clustering of exceedances at each threshold level for the stations Atlantis to Malmesbury 64

4.13 Level of clustering of exceedances at each threshold level for the stations Molteno to Wellington 65

4.14 Q-Q and return level plots for models fitted to data from CPT Int. 68

4.15 Q-Q and return level plots for models fitted to data from Tygerhoek 69

4.16 Q-Q plots to show goodness-of-fit of the models for the stations Atlantis to Malmesbury 71

4.17 Q-Q plots to show goodness-of-fit of the models for the stations Molteno to Wellington 72

4.18 Return level plots to show goodness-of-fit of the models for the stations Atlantis to Malmesbury 73

4.19 Return level plots to show goodness-of-fit of the models for the stations Molteno to Wellington 74

5.1 The 25 year return level values, where symbol size shows relative magnitudes 76

5.2 Relation of the 25 year return level estimate to the coordinates 78

5.3 Goodness-of-fit of linear spatial trend surface to the 25-year return level estimates in the study area 79

5.4 The relation of the 25-year return level estimates to altitude 80

5.5 Pooled variogram cloud of the standardized 25 up to 50 year return level estimates 82

5.6 Empirical variogram and potential variogram model curves superimposed. In red is the exponential model, blue is the spherical and in green the pentaspherical model. 84

5.7 Comparison of maps derived by ordinary Kriging against universal kriging for the 50 year 24-hour return level estimate 85

5.8 Map of the uncertainty about the ordinary and universal kriging estimates for the 50 year 24-hour return level estimate 86

A.1 Mean residual life plots: Atlantis to Molteno 91

A.2 Mean residual life plots: Paarl – Wellington 92

A.3 Threshold stability plots: Atlantis – Jonkersberg 93

A.4 Threshold stability plots: Ladismith – Malmesbury 94

A.5 Threshold stability plots: Molteno – Porteville 95

A.6 Threshold stability plot: Wellington 96

A.7 Additional threshold sensitivity diagnostics: Atlantis – CPT Astro. 97

A.8 Additional threshold sensitivity diagnostics: Excelsior – Jonkersberg 98

A.9 Additional threshold sensitivity diagnostics: Ladismith – Langebaanweg . . . 99

A.10 Additional threshold sensitivity diagnostics: Langgewens – Malmesbury . . . 100

A.11 Additional threshold sensitivity diagnostics: Molteno – Paarl 101

A.12 Additional threshold sensitivity diagnostics: Plettenbergbaai – Porteville . . 102

A.13 Additional threshold sensitivity diagnostics: Wellington 103

A.14 Sensitivity of the return level estimates to the threshold values: Atlantis – CPT Astr. Obs. 104

A.15 Sensitivity of the return level estimates to the threshold values: Excelsior – Jonkersberg 105

A.16 Sensitivity of the return level estimates to the threshold values: Ladismith – Langebaanweg 106

A.17 Sensitivity of the return level estimates to the threshold values: Langgewens – Malmesbury 107

Chapter 1

Introduction

> *According to usual opinions, the extreme values, and especially those which occur in meteorology, are so irregular that no prediction can be made of the maximum precipitation, extreme temperature, maximum pressure, etc., that will occur in a given period of time.*
> – Emil J. Gumbel, 1942 (Katz and Naveau, 2010)

In contrast to the historic criticism of extreme value analysis, early developments in extreme value theory were motivated by problems arising in climatology and hydrology (Katz and Naveau, 2010). Today methods of extremes continue to thrive in areas such as hydrology and coastal engineering, where estimates of design values (or return level estimates as they are known in extreme value theory), provide important information in planning and design of engineering projects (Katz et al., 2002). The complexity of the climate system and the spatio-temporal nature of climate process, provides fertile ground for further developments in the statistical analysis and modelling extremes and possibilities for solutions which combine deterministic geophysical models and stochastic methods.

1.1 Rationale

In recent years the frequency of occurrence of natural disasters has risen with mounting scientific evidence pointing to a changing climate as the root cause (IPCC, 2007)[1]. It is anticipated that during the course of this century the occurrence of natural hazards will intensify, rendering regions such as Africa more vulnerable to the impacts of a changing climate (IPCC, 2007; ROA, 2007; IPCC, 2008)[2]. The need to quantify historical patterns

[1] IPCC – Intergovernmental Panel on Climate Change
[2] International Council for Science Regional Office for Africa (ICSU ROA)

of processes driving such events is urgent, however the challenge, especially in Africa, is the scarcity of quality data and skills (ROA, 2007).

Hydro-meteorological events account for the highest proportion of disasters in sub-Saharan Africa, with 59% of the natural disasters in the period 1975 – 2002 being of hydro-meteorological origin (ROA, 2007). It is anticipated that this region will experience more frequent extreme precipitation events. Droughts and floods, which often result from extreme precipitation events are a concern for socio-economic stability, public health and safety of vulnerable individuals.

Eastern and southern parts of Africa are expected to experience a decline in mean precipitation by the end of the 21st century (IPCC, 2008), with water availability in nine countries[3] projected to be less than 1000 m^3 per person per year by 2025. In the south-western Cape in South Africa, it is projected that an increase in water demand due to climate change associated with global warming could be as much as 0.6% per year by 2020, whilst supply could be reduced[4] by 0.32% per year (IPCC, 2008).

Inference drawn from rainfall studies pertaining to South Africa, especially the south western part of South Africa, is that on average the number of rainy days is decreasing. However, on days when it does rain, it is more likely to rain heavily (Mason and Joubert, 1997; Mason et al., 1999; Fauchereau et al., 2003). This is a challenge for flood disaster assessment and management, especially since historical records of natural hazards from the Emergency Events Database[5] show that floods are most frequent and result in more damages in comparison to other types of natural disasters in the country. Table 1.1 provides a summary of hazard occurrence and damages for the period 1900 – 2010. While droughts have occurred nearly 3 times less than floods and disrupted the lives of nearly 18 million people, floods have affected far fewer people than droughts, but have claimed more lives and resulted in more financial losses for the damages. In fact, Table 1.1 re-iterates the finding that hydro-meteorological disasters in Sub-Saharan Africa occur more frequently than other types of natural disasters, as nearly 70% of historical disasters recorded on EM-DAT for South Africa are of hydro-meteorological type. In the 2006/07 financial year nearly 780 million Rands in financial resources were mobilized by the government to rehabilitate roads and infrastructure, as well as housing as a result of the Western/Eastern Cape floods, Northern Cape floods and Taung floods (DPLG, 2007). Therefore, careful study of heavy rainfall, the historic pattern and future possibilities is important in particular for the public, property and animals that may be at risk.

[3]Djibouti, Cape Verde, Kenya, Burundi, Rwanda, Malawi, Somalia, Egypt and South Africa

[4]A reduction in water supply capacity could either be a result of decreasing precipitation or an increase in evaporation

[5]EM-DAT: `www.em-dat.net` on 14 June 2010. The database is maintained by the World Health Organization Collaborating Centre for Research on the Epidemiology of Disasters (CRED) in collaboration with the Office of Foreign Disaster Assistance (OFDA-USAID). It is located at the Universitè Catholique de Louvain in Brussels, Belgium

Table 1.1: Summary on historical disasters for South Africa 1900 – 2010

Hazard Type	Event Count	People Killed	People Affected	Est. Damages (000 USD)
Drought	8	–	17475000	1000000
Earthquakes	8	70	1448	20000
Epidemic	7	325	112385	–
Extreme Temp.	2	52	–	–
Floods	30	1174	240150	1210029
Landslides	1	34	–	–
Storms	21	226	627472	754041
Wildfires	9	128	7380	440000

Water availability is critical for plant growth, however floods can have negative impacts on agricultural yields. The negative effect on agricultural production caused by such extremes is exacerbated by the farmers' lack of capacity to mitigate and adapt to these unfavourable conditions (Leichenko and O'Brien, 2002). In South Africa it is estimated that 20.7% of households are involved in agricultural production, with a large proportion in subsistence farming (StatsSA, 2010a). The Western Cape is one of three provinces in South Africa with a significant proportion (estimated at 23.5%) of agricultural producers who sell their produce. The annual contribution of agriculture, forestry and fisheries to gross domestic product (GDP) at constant 2005 prices was estimated at 2.5% in 2009 (StatsSA, 2010b). The commercial agriculture sector had an estimated 800 000 employees with a 54.2% and 45.8% split between full-time and casual employees (StatsSA, 2009). The contribution of agriculture to GDP in comparison with other economic sectors is small. However, for approximately one million employees in commercial agriculture and the predominantly rural populace dependent on subsistence farming, substantial reduction in yield as a result of increased frequency of heavy rainfall events may add to the strain of food security, unemployment and consequent poverty (IPCC, 2008).

> However big floods get, there will always be a bigger one coming, so says one theory of extremes, and experience suggests it is true.
> – President's Water Resources Commission, 1950 (chaired by Morris Cooke)

1.2 Modelling Rare Events

It is rare to observe the extremes of a process, presenting a challenge in the quantification of the pattern of such events. Much of the statistical theory and methods focus on quantifying the behaviour in the centre of the marginal distribution of the process. As a consequence, the tails of the marginal distribution of the process are often poorly estimated. Extreme

Value Theory (EVT) is attractive in that the distribution of the entire process need not be known, but the tail behaviour can be characterized, through modelling the extremal values of the process or those values beyond a particular threshold (Coles, 2001).

Rainfall, like most meteorological processes, exhibits spatial variation (Casson and Coles, 1999). Further, heavy rainfall tends to persist over several days (Mason and Joubert, 1997), exhibiting the tendency of extreme events to be correlated in time. Rainfall is therefore a spatially and temporally dependent process, therefore in modelling the extremes of this process one needs to account for both these dependencies.

The incorporation of temporal dependence and spatial variation is not a simple task. The main issue is that of high dimensionality. Further, rainfall recording relies on the availability of instrumentation, which may be broken during severe storms, resulting in missing data during periods of time where it is out of service. Further, stations within a particular region are not necessarily established at the same time, which results in site-wise time series of differing lengths. Characteristics such as short range correlation, seasonality, missing data and long-term trend are prevalent in the extremes of time series. Also, the ground-based weather observation network is often sparse resulting in large extents of the region not being monitored, which compromises the quality of the inference spatially. Some of these issues have been considered at length (Smith, 1989a; Coles, 1994; Morton et al., 1997). It is anticipated that these traits are also present in the data that will be used for this study, and that some of these existing methods will contribute to the analysis.

1.3 Research Objective

The aim of this study is to present a methodology to model heavy rainfall observed in the Western Cape over a period of time, by employing the theory of extreme values. Heavy rainfall is defined in the study as rainfall amounts above a specified threshold. Focus is on quantifying uncertainty due to temporal dependence at high levels of the process and spatial variation.

It is anticipated that models will be used to answer the following questions:

- Is there statistical evidence of a trend in the pattern of heavy rainfall in the Western Cape?
  - Can any trend be detected in the frequency of threshold exceedance?
  - How does modelling temporal dependence in the exceedance process affect the return level estimates?
- Is the fifty year return level estimate homogenous over the Western Cape region?

A substantial part of this study is to construct site-level models, with uncertainty about the resultant return level estimates explicitly quantified. In the latter part of the study, the geostatistical modelling of return level estimates, the degree of spatial variation in the estimates is quantified.

1.4 Structure of the Research Report

In this section the structure of the research report is described. The report is made up of six chapters with supplementary material which is located at the end as an appendix and a list of bibliographical material. A description of the chapters is as follows:

INTRODUCTION Highlights importance of studies of rainfall extremes by reflecting on impacts on climatology, occurrence of natural disasters and the economy. It provides background to the modelling framework and issues that need consideration, accompanied by specific questions that the research aims to address.

LITERATURE SURVEY Various studies of rainfall for southern Africa are reviewed. It provides an overview of extreme value theory and methods for incorporating temporal and spatial dependence.

METHODS Describes the point process approach of extreme value theory. Concepts of geospatial inference and prediction through kriging are introduced.

APPLYING THE POINT PROCESS EXTREME VALUES APPROACH The study area and data are described. A detailed discussion of the extreme value analysis at each site is given.

GEOSTATISTICAL MODEL OF RAINFALL RETURN LEVELS This chapter describes in detail the results of the regional analysis of daily rainfall return levels.

CONCLUDING REMARKS Formulates concluding remarks of this research, highlighting limitations of the study and avenues for further research.

Chapter 2

Literature Survey

This section reviews the theory supporting the statistical modelling of extreme values. Initially a discussion on rainfall and studies of extremes of this process in particular for South Africa is given. We discuss the founding theory of extreme values, termed the block maxima approach, which describes the limiting behaviour of the largest (or smallest) value in a random sample. Then threshold exceedance methods are reviewed. Lastly, a review of techniques used to incorporate uncertainties due to temporal and spatial dependence of rainfall extremes is given.

2.1 Studies of Rainfall

South Africa is considered to be a relatively dry country, with rainfall that is highly variable both in time and space. The driest parts of the country are in the west in places close to Namibia[1]. In contrast, the wetter areas, mainly on the mountainous areas of the south-western Cape and the eastern escarpment, receive average annual rainfall that is estimated to be more than 2000 mm (Kruger, 2007). In the south-western Cape in particular, annual rainfall varies from approximately 400 mm on the coastal plateau, to approximately 2000 mm at higher elevations.

The most important rain-producing system is the tropical-temperate trough (TTT) which links a surface heat low and easterly waves from central to western southern Africa to the westerly waves (or midlatitude disturbance) and low from the South Atlantic, south of Africa (Reason, 1998; Williams et al., 2007). This generally forms a band of cloud with a north-west to south-east orientation. The position of this cloud band, which is affected by the location of the surface heat low, is an important determinant of whether rain falls over

[1]The Namib and the Kalahari deserts are found in this region

the country. In the north, the low level flow is in an easterly direction advecting in moisture from the south western Indian Ocean, whilst in the south; strong westerly waves advect moisture from the South Atlantic. Uplift as a result of convergence of these two sources of moisture transport, causes widespread rains over the country.

This TTT system is responsible for the summer rainfall over the region, affecting central and eastern parts of South Africa (Preston-Whyte and Tyson, 1988; Reason, 1998). Insufficient coupling of the two sources of moisture advection, resulting in cloud bands truncating at relatively low latitudes is responsible for passage of midlatitude cold fronts, bringing rainfall to the south-western Cape and south coast regions of South Africa. The amplitude of the disturbance in the westerly waves is greatest in winter, coinciding with the south-western Cape receiving most of its annual rainfall (Reason, 1998). Thus, the main drivers of rainfall in the southwest of South Africa (largely the Western Cape Province) which is classified as a winter rainfall region are orography[2] and frontal activity (Preston-Whyte and Tyson, 1988; Reason, 1998). Recently a link has been found with the Antarctic Oscillation and the rainfall in this winter rainfall region (Reason et al., 2006; Kruger, 2007). The south coast, consisting of areas east of the Western Cape and west of the Eastern Cape provinces, is classified as an all season rainfall region. This region is influenced by the tropical-temperate trough in the summer and frontal activity in winter.

Atmospheric circulations as noted above play an important role in the occurrence and diversity of rainfall in South Africa. Furthermore, this is unique due to the location of South Africa between the Indian, Atlantic and Ant-arctic Oceans. Therefore, in an attempt to accurately model rainfall extremes, it is plausible to take atmospheric and oceanic associations to extreme rainfall into account (Williams et al., 2007). Possible sources of this information include indices obtained from dynamic climate models of sea-surface temperatures, near-surface geopotential height and soil moisture indices (Reason, 1998; Williams et al., 2007; Shongwe et al., 2009). The inclusion of additional information is not considered in this study as this could not be obtained during the period when the study was undertaken.

There have been few studies focussing on modelling rainfall extremes for southern Africa (New et al., 2006). Partly, this is attributable to the paucity of data, for example, low coverage of the surface observing network (consisting of conventional and automatic weather stations on land), in relation to the desired minimum level as specified by the Regional Basic Synoptic Network (RBSN) (Schulze, 2007). In 2004, the global average of 77% coverage was declared deficient, meaning that Africa at coverage of only 53% is not doing too well according to RBSN standards (Schulze, 2007). The South African Weather Service (SAWS) has approximately two thousand three hundred weather stations scattered throughout the country, as well as in the Marion, Gough and the Vesleskarvet islands. Compared to the countries in

[2]Orography refers to the region's elevated terrain. Orographic rainfall occurs on the windward side of the elevated terrain, caused by either upward deflection of large scale horizontal wind flow or upward propagation of moist air on the orographic slope resulting from daytime heating of the mountain terrain. As the air cools, water vapour condenses to form clouds, which cause rain if the water droplets grow large enough.

southern Africa, South Africa has a dense network of stations which enables enrichment of knowledge regarding precipitation extremes in this region (Shongwe et al., 2009).

Approaches to modelling extreme rainfall in South Africa have been based on extreme rainfall indices and to a lesser extent extreme value theory. Examples of observed extreme daily rainfall indices that have been analyzed include daily rainfall totals greater than 10 mm for heavy rainfall, 20 mm for very heavy rainfall and the upper percentiles (95^{th} and 98^{th}) as indices for extreme rainfall (Kruger, 2006; New et al., 2006). Using data from one hundred and thirty eight stations, no significant changes (or trends) in extreme daily precipitation could be found over most of the country for the period 1910 – 2004 (Kruger, 2006). New et al. (2006) analyzed trends in daily temperature and rainfall extremes for Southern and West Africa. The analysis was for a total of sixty three locations, of which only eight were South African, covering a thirty year period 1961 – 2000. Spatial patterns of heavy and extreme rainfall were found to be inconsistent over the study region. Three out of the ten indices for extreme precipitation showed statistically significant increasing trends. The average wet day precipitation and the annual daily precipitation maximum showed increasing trends for stations on the east coast of Southern Africa and West Africa. Indices of heavy and extreme precipitation showed an increasing, but non-significant trend for Cape Town.

The Regional Frequency Approach (RFA) was used to estimate both short and long duration design rainfall events and design floods for South Africa (Smithers and Schulze, 2000, 2002). In particular for design rainfall estimation a Scale Invariance approach to the Regional L-Moment Algorithm was developed as a result of observed changes in scaling for certain durations (Smithers and Schulze, 2002). Simulation as well as historical data studies have been done to investigate possible changes in the rainfall regime as a result of climate change (Mason and Joubert, 1997; Mason et al., 1999; Fauchereau et al., 2003). The extreme value approach considered was the block maxima approach, fitting the GEV model to annual daily maxima and annual pentad (five days) maxima (Mason and Joubert, 1997; Mason et al., 1999). The point process approach has not been widely applied for studying rainfall extremes in South Africa.

In the next section, the classical method of extreme value analysis is briefly reviewed.

2.2 Classical Extreme Value Theory

The foundation of EVT is due to Fisher and Tippett (1928) who described the limiting behaviour of the smallest (or largest) member of a sample, in what has become known as the "three types theorem". Their aim was to find the exact distribution of the extreme member of the sample, however, they found this to be a numerically difficult task. Instead of the exact distribution, they found the possible limiting forms, noting that in all cases they explored, a particular group of distributions appeared. They considered a

sample of size $m \times n$, where both n and m are large. Let the random variable $M_{m,j} = \max(X_{1,j}, X_{2,j}, \ldots, X_{m,j})$, $j = 1, 2, \ldots, n$ be the maximum of a sequence of m random variables , such that $\{M_{m,1}, M_{m,2}, \ldots, M_{m,n}\}$ is a sequence of n independent and identically distributed maxima. Fisher and Tippett stated that the extreme member of the sample $m \times n$, $M_n = \max(M_{m,1}, M_{m,2}, \ldots, M_{m,n})$, has the same limiting distribution as the individual $M_{m,j}$ $j = 1, 2, \ldots, n$ as $m, n \to \infty$. Practically this means that the largest observation in a sample has the same distribution as that of maxima taken over sufficiently large and fixed subsets of the sample.

To derive the limiting distribution, consider the maximum from a sequence of m random variables. For now we ignore the index n which is the total number of such subsets. The maximum of a sequence of m random variables is

$$M_m = \max(X_1, X_2, \ldots, X_m) \; . \tag{2.2.1}$$

Suppose that the probability of an observation being at most x is denoted by $F(x) = P(X_i \leq x), i = 1, 2, \ldots, m$. Using the assumption of independence and identically distribution of random variables, the probability that the largest of the sequence is at most x is,

$$\begin{aligned} P(M_m \leq x) &= P(X_1 \leq x, X_2 \leq x, \ldots, X_m \leq x) \\ &= [P(X \leq x)]^m \\ &= F^m(x) \; . \end{aligned}$$

To ensure that the distribution of M_m does not degenerate to a point mass as $m \to \infty$, Fisher and Tippett (1928) constructed the functional equation

$$F^m(x) = F(a_m x + b_m) \; , \tag{2.2.2}$$

where the positive $\{a_m, m \geq 1\}$ and $\{b_m, m \geq 1\}$ are sequences of constants. They showed that the solutions of the functional Equation 2.2.2, give all possible limiting forms of distributions. These fall into three classes, resulting in only three possible limiting curves, that is:

$$F(a_m x + b_m) \to = \begin{cases} \exp(-\exp(-x)) \, , & -\infty < x < \infty \\ \exp(-x^{-k}) \, , & k > 0 \\ \exp(-(-x)^k) \, , & k < 0 \end{cases}$$

Therefore the limiting distribution of the largest of $m \times n$ random variables, must satisfy a functional equation which limits its form to one of only two main types – the limit as $1/k \to 0$ and the limit when $1/k$ is a single non-zero value.

Consider the limiting distribution of the largest of $m \times n$ random variables M_n, which was shown to be that same as that of the subset maxima M_m by Fisher and Tippett (1928). Assuming sequences of normalizing constants $\{a_n > 0\}$ and $\{b_n\}$ exist, then $\left(\dfrac{M_n - b_n}{a_n} \right)$ also converges in distribution to a random variable belonging to one of only three types of

distributions. These three types are called the Fréchet, Gumbel and Weibull distributions and they are unified under the generalized extreme value (GEV) family of distributions. The distribution of M_n can therefore be approximated by the generalized extreme value distribution (GEV), that is:

$$P\left[\frac{M_n - b_n}{a_n} \leq x\right] \longrightarrow G(x;\mu,\sigma,\xi) = \exp\left[-\left\{1+\xi\left(\frac{x-\mu}{\sigma}\right)\right\}_+^{-\frac{1}{\xi}}\right] \tag{2.2.3}$$

where $1+\xi\left(\frac{x-\mu}{\sigma}\right) > 0$, $\mu \in (-\infty;\infty)$, $\sigma > 0$ and $\xi \in (-\infty;\infty)$. The notation y_+ is used extensively in this report. It is defined as follows:

$$y_+ = \max(y,0) \tag{2.2.4}$$

$$= \begin{cases} y & \text{if} \quad y > 0 \\ 0 & \text{otherwise} \end{cases} \tag{2.2.5}$$

The theoretical requirement is that the normalizing sequences $\{a_n > 0\}$ and $\{b_n\}$ must exist, however practically these need not be known. They can be approximated by the location (μ) and scale (σ) parameters of the GEV distribution. The parameter ξ describes the shape of the tail of the population's marginal distribution. Mathematical justification of the theory by Fisher and Tippett (1928) was given by Gnedenko (1943). Further contributions to the statistical methodology for a series maxima (or minima) of some random process, observed over a fixed time interval, was given by Gumbel (1958).

Characterizing the behaviour of the largest rainfall event taken over some fixed period of time is important in flood design and other hydrological studies. However, in the presence of more information, the block maxima approach is considered wasteful of data since in creating blocks, all observations are discarded, except the largest value in each block. This flaw is the main reason for popularity of threshold exceedance methods, especially in environmental applications where the number of years of observation is often small, resulting in small samples when annual maxima are considered.

2.3 The Threshold Exceedance Approach

The method based on threshold exceedances is a common alternative to the block maxima approach. It is based on the distribution of exceedances beyond a suitably *high* threshold u. Given that an observation exceeds u, the probability that the excess is at least y can be approximated by the Generalized Pareto Distribution (GPD), that is for $Y = X - u > 0$:

$$P(X \geq u + y | X > u) \longrightarrow H(y;\tilde{\sigma},\xi) = 1 - \left(1 - \frac{\xi y}{\tilde{\sigma}}\right)_+^{\frac{1}{\xi}} \tag{2.3.1}$$

where $\tilde{\sigma} = \sigma + \xi(u - \mu) > 0$. James Pickands III formulated the idea of a Generalized Pareto upper tail of a continuous distribution function. He made precise the connection that, if the approximating distribution of block maxima lies in the domain of attraction of $G(x; \mu, \sigma, \xi)$ (Equation 2.2.3), then for sufficiently large thresholds u, the corresponding excesses can be approximated by a distribution within the Generalized Pareto family. Further, he showed that the shape parameters of the two limiting distributions are equivalent. Some statistical properties of the GPD were investigated by Davison (1984); Smith (1984) and Davison and Smith (1990), including the development of inferential methods for incorporating of covariate information.

The point process characterization was developed by Pickands (1971). It was given thorough mathematical justification, in the case of stationary stochastic processes, by Leadbetter (1983); Leadbetter et al. (1983). The use of the point process characterization as a modelling tool for peaks over threshold was advocated by Smith (1989b), showing the ease with which non-stationarity can be accounted for in the model. To define the point process approach, consider a sequence of n random variables and a large threshold u. The bivariate point process is defined by

$$N_n = \left\{ \left(\frac{j}{n+1}, X_j \right), \qquad j = 1, 2, \ldots, n \right\}$$

restricted to the region $A_u = (0, 1) \times (u, \infty)$, converges in distribution to the non-homogeneous Poisson process with intensity measure in the parametric family

$$\Lambda((a, b) \times (x, \infty)) = (b - a) \left[1 + \frac{\xi}{\sigma}(x - \mu) \right]_+^{-\frac{1}{\xi}} \tag{2.3.2}$$

where $\mu \in (-\infty, \infty)$, $\sigma \in (0, \infty)$ and $\xi \in (-\infty, \infty)$.

From a practical perspective Casson and Coles (1999) suggest multiplying the right hand side of Equation 2.3.2 by an arbitrary scaling coefficient m. This scaling coefficient determines the time scale of the parameters. It is common in practice, especially for environmental processes to be interested in yearly time scales. In such cases m is the number of years of observation $m = n/n_y$, where n is the sample size and n_y is the number of observations in a year. By standard properties of a Poisson process and Equation 2.3.2, it follows that the annual maximum distribution will be given by the equation

$$P(M_{n_y} \leq x) = \exp \left\{ -\Lambda \left(\left(0, \frac{n_y}{n} \right) \times (x, \infty) \right) \right\} \tag{2.3.3}$$

$$= \exp \left[- \left\{ 1 + \frac{\xi}{\sigma}(x - \mu) \right\}_+^{-\frac{1}{\xi}} \right] \tag{2.3.4}$$

which corresponds to the GEV distribution.

Theoretically derivation of a distribution for threshold excesses also follows from properties of Poisson processes and Equation 2.3.2. Consider a bivariate point process of times of exceedance and excess values (T_i, X_i), for which $X_i > u$. Provided convergence to the Poisson process can be assumed,

$$P(X_i > x | X_i > u) = \left[1 + \frac{\xi}{\tilde{\sigma}}(x - \mu)\right]_+^{-\frac{1}{\xi}} \tag{2.3.5}$$

which is the generalized Pareto distribution (GPD) as given by Davison and Smith (1990), with $\tilde{\sigma} = \sigma + \xi(u - \mu) > 0$ as before.

Essentially the classical and the threshold exceedance method based on the GPD, can be thought of as being unified within the point process characterization. The advantages of the point process characterization are: the invariance of the parameter estimates to threshold choice and the incorporation of the exceedance rate into the analysis. The likelihood function, which is censored at the threshold, can be maximized numerically, following methods developed by Davison and Smith (1990). Upon maximization of the likelihood function, the parameter estimates obtained can be used to calculate the r-year return level estimate x_r by substitution and solving for the unknown quantile.

The point process approach to modelling the extremes has been popular in the hydrological domain with studies dating back to 1975 (such as the flood studies report (NERC, 1975)). Recently, the popularity in employing this approach can be attributed to the ease with which temporal and spatial features can be incorporated into the model. This is important in environmental studies as most processes inherently have these features.

2.4 Quantile Estimation

The results of an extreme value analysis are easier to interpret in terms of return levels as opposed to the individual model parameters. The N-year return level z_N is the $(1 - 1/N)$ quantile of the annual maximum distribution. Using the GEV, the return level can be obtained by solving

$$1 - \frac{1}{N} = \begin{cases} \exp\left[-\left(1 + \frac{\xi}{\sigma}(z_N - \mu)\right)^{-\frac{1}{\xi}}\right], & \xi \neq 0 \\ \exp\left[-\exp\left(-\sigma^{-1}(z_N - \mu)\right)\right], & \xi = 0 \end{cases}$$

Exploiting the approximation $1 - \frac{1}{n} \approx \exp\left(-\frac{1}{n}\right)$, this simplifies to

$$z_N = \begin{cases} \mu - \frac{\sigma}{\xi}\left[1 - (-\log(1-p))^{-\xi}\right], & \xi \neq 0 \\ \mu - \sigma \log\left(-\log(1-p)\right), & \xi = 0 \end{cases} \tag{2.4.1}$$

For the threshold exceedances, estimation of the return level involves estimating the probability of an individual observation exceeding the threshold u. Suppose there is m exceedances during a period of n years. Then, the exceedance process is assumed to be Poisson with rate λ (per year) estimated by $\hat{\lambda} = m/n$. From Section 2.3, for an arbitrary $x > u$

$$P(X > x|X > u) = \left[1 + \xi\left(\frac{x-u}{\sigma_u}\right)\right]^{-\frac{1}{\xi}}$$

so that the mean crossing rate (per year) of level x is

$$\lambda\left[1 + \xi\left(\frac{x-u}{\sigma_u}\right)\right]^{-\frac{1}{\xi}} .$$

Following from the definition of the return level estimate as the level which is exceeded on average once every N years, we set

$$\frac{1}{N} = \lambda\left[1 + \xi\left(\frac{x-u}{\sigma_u}\right)\right]^{-\frac{1}{\xi}}$$

to obtain the return level

$$q_N = \begin{cases} u + \dfrac{\sigma_u}{\xi}\left[(\lambda N)^{\xi} - 1\right], & \xi \neq 0 \\ u + \sigma_u \log(\lambda N), & \xi = 0 \end{cases} \tag{2.4.2}$$

The estimate of the return levels $\hat{z}_N$ and $\hat{q}_N$ can be obtained by substituting the respective MLEs of the parameters and $\hat{\lambda} = m/n$ respectively for the GEV and GP distributions (Davison and Smith, 1990; Coles, 2001). The corresponding Wald-type confidence intervals can be calculated once the variance has been estimated using the delta function approximation. In practice gain in precision is attained when inference is based on profile likelihood (Coles, 2001).

2.5 Inference

Theoretically, maximum likelihood approximation results are valid for large sample sizes and in practice these estimators are still applicable when sample size is small. Simplicity of maximum likelihood inference even in cases where the model may be structurally complicated, as is the case when covariate information is included, make it popular in extreme value modelling (Coles, 2001; Katz et al., 2002). Alternatives have been suggested mainly due to concerns over the reliability of maximum likelihood inference when sample size is small (Katz et al., 2002), which is a common problem in extreme value modelling. In the next section maximum likelihood inference in extreme value analysis is discussed using the Poisson-Generalized Pareto Distribution P-GPD for illustration. While the P-GPD is not formulated explicitly through a point process approach, the parameters of this model are theoretically equivalent to those derived of the point process extreme value model.

2.5.1 Maximum Likelihood Inference

Consider random variables $X_1, X_2, \ldots, X_m$ having marginal distribution $F(x)$. Denoting the unknown parameters of F as θ, the probability of an observed data as a function of the parameters θ, is the likelihood function. Assuming independence of the observations, the likelihood function is

$$L(\theta) = f(\mathbf{x}|\theta) = \prod_{i=1}^{m} f(x_i;\theta) \ .$$

Maximizing the logarithm of the likelihood with respect to each parameter yields maximum likelihood estimates (MLE) of the unknown parameters. A desirable property of the maximum likelihood estimator is that it is approximately Gaussian distributed with the variance-covariance matrix approximated by the inverse of the *observed information matrix*[3] of $-\log L$. The square root of the i^{th} diagonal element of this matrix corresponds to the standard error of the i^{th} parameter estimate. Wald-type confidence intervals can be constructed for each estimate using the normality property of the MLEs.

For derivation of the log-likelihood for threshold excesses in a P-GPD model, consider a process $\{X_1, \ldots, X_n\}$, and a threshold u, beyond which a Poisson rate for the frequency of threshold exceedance is assumed. The exceedance probability is,

$$\tau = P(X_i > u) \approx \frac{1}{n}\left[1 + \xi\left(\frac{u-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}} \ . \tag{2.5.1}$$

For the exceedances, the likelihood contribution is

$$P(X_i = x) = \tau P(X_i = x \mid X_i > u) \ .$$

In contrast to the likelihood for the GPD distribution, in the P-GPD model the probability of non-exceedance is also considered. The likelihood function is,

$$\begin{aligned} L(\tau, \tilde{\sigma}, \xi) &= (1-\tau)^{n-n_u}\tau^{n_u}\prod_{i=1}^{n_u}\frac{1}{\tilde{\sigma}}\left[1 + \frac{\xi}{\tilde{\sigma}}(x_i - u)\right]^{-\frac{1}{\xi}-1} \\ &= \exp(-n\tau)\left(\frac{\tau}{\tilde{\sigma}}\right)^{n_u}\prod_{i=1}^{n_u}\left[1 + \frac{\xi}{\tilde{\sigma}}(x_i - u)\right]^{-\frac{1}{\xi}-1} \end{aligned}$$

where $\tilde{\sigma} = \sigma + \xi(u-\mu)$. The log-likelihood that is to be minimized to get parameter estimates $(\hat{\tau}, \hat{\tilde{\sigma}}, \hat{\xi})$ is,

$$\begin{aligned} -\,l(\tau, \tilde{\sigma}, \xi) &= n\tau - n_u \log\tau + n_u \log\sigma + \\ &\qquad \left(\frac{1}{\xi} + 1\right)\sum_{i=1}^{n_u}\log\left[1 + \frac{\xi}{\tilde{\sigma}}(x_i - u)\right] \end{aligned} \tag{2.5.2}$$

[3]The observed information matrix is also known as the Hessian matrix. The elements of this matrix are the second order partial derivatives evaluated at the maximum point of the likelihood surface of the estimator.

Suppose you want to infer about component θ_i of the parameter vector θ. Define the log-likelihood $l(\theta_i, \theta_{-i})$ where θ_{-i} denotes all the components of θ, excluding θ_i. In simple terms, re-express the parameter you wish to estimate (θ_i) as a function of the other parameters. Construct the re-parameterized log likelihood function. The profile likelihood for each θ_i is obtained by maximizing the log-likelihood with respect to all the other components $\theta_{-\mathbf{i}}$. That is,

$$l_p(\theta_i) = \max_{\theta_{-i}} l(\theta_i, \theta_{-i}) \ .$$

Under standard regularity conditions for maximum likelihood and the null hypothesis that θ_i is the true parameter, an approximate $100(1-\alpha)\%$ confidence interval for θ_i consists of all values for which

$$2\left[\log l_p(\hat{\theta}_i) - \log l_p(\theta_i)\right] \approx \chi^2_{1;1-\alpha}$$

where $\chi^2_{1;1-\alpha}$ is the $(1-\alpha)$ quantile of the χ^2_1 distribution. Profile likelihood confidence intervals are asymmetric, hence their suitability when the distribution is skew, which is typical of extreme value distributions.

2.5.2 Alternatives to Maximum Likelihood Inference

Various techniques have been proposed as an alternative to the MLE for parameter estimation in extreme value modelling. These include the method Probability-Weighted Moments (PWM) where the parameters are estimated as specified functions of ordered statistics which are empirical estimates of probability-weighted moments for the GEV (Hosking et al., 1985),

$$\beta_r = M_{1,r,0} = E[X\{F(X)\}^r], \quad r = 0, 1, 2, \ldots .$$

This method of estimation, especially in hydrology is considered to be the main contender of the Maximum Likelihood Estimator (MLE) (Katz et al., 2002). The motivation for the method of PWM is that the MLE performs well in large samples, but often fails in small samples. Using the aid of simulation it was shown that for very small samples, the inefficiency of the MLE was due to the non-convergence of the Newton-Raphson approximation and that in such cases PWM outperforms the MLE (Hosking et al., 1985). Although simple to compute and their importance for small sample estimation, the criticism against PWM is their lack of generality, especially when complex models are considered. Often there is not much difference in performance when the two methods are compared, even for small samples (Coles, 2001; Katz et al., 2002; Smith, 2003).

Bayesian techniques have gained popularity within the extreme value modelling community. Here, parameters are considered as random variables instead of constant values. Since parameters are unknown, we can only formulate a *belief* about their distribution, known as the *prior* distribution. This prior distribution, together with the likelihood function, form the *posterior* distribution of the parameters, defined as the conditional distribution of the

parameters given the observed data $f(\theta|\mathbf{x})$. Direct implementation of Bayes theorem is generally complicated, and standard numerical estimation techniques may be difficult to obtain, however the breakthrough solution to this problem is the use of the Markov Chain Monte Carlo (MCMC) approximation techniques (Hastings, 1970; Smith and Roberts, 1993).

The Bayesian framework is attractive in EVT because it provides a coherent framework for the incorporation of additional information (Coles et al., 2003), the analysis in cases where the maximum likelihood function is non-regular, as well as the ease with which predictive uncertainty is incorporated through the predictive density (Davison, 1986; Engelund and Rackwitz, 1992; Coles and Powell, 1996). The term used in hydrology for the predictive density of the return level is *design flood distribution*. Although there is strong evidence in support of the value of the Bayesian approach to EVT, a crucial consideration is the choice of an appropriate prior distribution. The choice of an appropriate prior has substantial bearing on the resulting posterior distribution of the estimator, which forms the basis of one's conclusions.

There is a variety of other estimation techniques in addition the method of PWM and the Bayesian approach, however, likelihood based techniques remain attractive due to their all-round utility and adaptability to complex models (Coles, 2001). The MLE of the GEV is obtained through iterative numerical procedure, with the conditions for regular estimation being satisfied when $\xi > -0.5$ (Prescott and Walden, 1980). Similarly the MLE for the GPD is an iterative solution of the likelihood function derived by Davison and Smith (1990). Particularly important in this result is that it is also obtained in the presence of covariate information.

2.6 Modelling Issues in Extreme Value Analysis

In practice there are many issues to consider when modelling extremes, as in most cases the data violates the assumptions of the model. This section outlines some of these issues, mainly the research that has been done, providing a mix of solutions that modellers can use when facing similar problems with the data at hand. The discussion is not exhaustive and does not aim to be, because the extreme value analysis is applicable in many fields and therefore giving rise to a wide array of problems and ways in which they can be solved.

2.6.1 Extreme Value Analysis when Data are Incomplete

This issue of missing observations is generally contentious in extreme value analysis, because using imputation methods of missing data based on expected values may be appropriate since the characteristics of expected values are different from those of extreme observations (Smith,

1989b). If the level of incompleteness is small, it may be justifiable to ignore the missing data as the effect can be thought to be minimal, however in the presence of additional data there are alternatives. One option is to incorporate into the model the proportion of incompleteness per fixed interval of analysis as a measure that adjusts the model output for the uncertainty arising from having incomplete data (Smith, 1989a; Smith and Shively, 1995). Expert knowledge, related records which are available in the public domain and other related information can be incorporated into the analysis through a Bayesian analysis framework (Coles and Tawn, 1996a; Sisson et al., 2006), provided that caution is exercised in incorporating such information as priors (Coles and Powell, 1996) to avoid getting erroneous results.

In cases where related data series is available, for example a data series available in a neighbouring site or when a series of a variable that is highly correlated with the one of interest, regression methods can be used. The idea is that given high correlation between the two series, the unknown values can be predicted using a regression model (Draghicescu and Ignaccolo, 2009; Zhang et al., 2009).

2.6.2 Considering Temporal Characteristics

The main assumption for the limiting distributions of extreme values is that the random sequences are independent and identically distribution (i.i.d.). This assumption is often violated by most meteorological processes. In South Africa rainfall is seasonal, with the Western Cape receiving between 30–50% of the regions total annual rainfall in the 3-month period June–August (New et al., 2006). Further, it has been found that a heavy rainfall day is likely to be succeeded by another rainy day (Mason et al., 1999).

The tendency of rainy days occurring successively pertains to the issue of short-range temporal dependence. This is an important issue in the threshold exceedance model because high level exceedances tend to cluster (Coles, 2001; Bierlant et al., 2004), and the effect of ignoring short-range temporal dependence is over-estimation of the return level. Some of the pioneering work in the area of temporal dependence in random sequences, was by Leadbetter (1983) and Leadbetter and Rootzén (1988). They advocated that subject to a local mixing condition which limit long-range dependence, the Fisher-Tippett theorem still holds for stationary processes subject to an additional parameter θ (the extremal index) describing the extent of the clustering at extreme levels. Since then, different approaches have been proposed for modelling the extremal index (including Smith, 1984; Davison and Smith, 1990; Smith and Weissman, 1994; Ledford and Tawn, 1996, 1998; Ferro and Segers, 2003; Heffernan and Tawn, 2004).

Rainfall, like most environmental processes change systematically with time. Such changes in shorter time periods are due to seasonality and over longer time periods gradual changes (trends) or even shifts can be observed. Intuitively, it is anticipated that these changes will

also manifest in the extreme levels of the process. The incorporation of non-stationarity of extremes in the extreme value theory context has developed largely into two streams (Coles, 2001):

- The *separate seasons approach* which involves separating the year into definitive seasons, modelling each season separately and eventually forming a joint model.
- The *continuous-time model approach* assumes a functional form for seasonality which is incorporated as covariates of the model parameters.

The separate seasons approach was illustrated by Morton et al. (1997) in the study of extreme wave heights in North Cormorant in the North Sea. They constructed point process models for each season, then aggregated these models to arrive at a return level corresponding to the entire year. In a study of temporal characteristics of heavy rainfall in south-west of England, Coles (1994) compared the separate monthly GEV models with a model where the location was a sinusoidal function of time. Méndez et al. (2008) considered sinusoidal location and scale parameters using the point process approach to model extreme wave height for given storm durations.

An earlier application of the point process extreme value approach in the case of non-stationarity was by Smith (1989a), where interest was in statistically modelling tropospheric ozone above the threshold level of 12 parts per hundred million (pphm) in order to detect and measure trends in the data. Non-stationarity due to seasonality was eliminated by restricting the analysis to only the summer period. As a further development Smith and Shively (1995) still considered the non-homogeneous Poisson process model, but in addition to accounting for temporal trends, meteorological conditions affecting troposheric ozone levels were also included. Here the parameters were defined as regression models of time and meteorological covariates (Smith and Shively, 1995).

2.6.3 Spatial Variation of Extreme Values

Another feature prominent in environmental processes is that locations which are closer tend to experience similar conditions. Therefore, it is expected that the return level of an environmental phenomenon will show variation in space. There are several approaches for incorporating spatial variation of the extreme value process, and the choice of a particular approach depends on the purpose of the study, the availability of additional information and computational tools and the understanding of the method. In this study, our objective is to investigate whether there is spatial variation in the return level over the study area and not necessarily spatial dependence between extreme observations themselves.

Several methodologies have been developed for modelling spatial extremes, which can be narrowed to three general approaches. The first involves modelling stronger forms of spatial

dependence as multivariate extreme value distributions based on the theory of max-stable processes. The formulation of max-stable processes and its connection to higher dimension extremes was due to de Haan and Resnick (1977); de Haan (1984); de Haan and Pickands (1984, 1986). Adaptation of this theory in support of modelling of spatial extremes was performed by Smith (1991); Coles (1993). In the max-stable approach the aim is to obtain the joint distribution across sites of the extreme observations, similar to obtaining component-wise distribution of the standardized maxima in multivariate extreme value theory. Methods for max-stable process of threshold exceedances are less developed, and hence they are currently an area of active research. This approach was not considered for this study because of the insufficient literature on the methodology and our interest in finding evidence of spatial variation in the return level rather than a model for dependence between threshold exceedances across the sites.

The second approach is based on a latent spatial process (Coles and Tawn, 1996b; Coles, 2001). Assume a continuous set $\mathbf{S} \in \mathbb{R}^2$ which indexes a geographical area – where the two dimensions are geographic coordinates, for example longitude and latitude. Consider rainfall observed at a particular location in $s_i : i = 1, 2, \ldots, n$ at any time instance, denoted as $X(s_i)_{t_j}, s_i \in \mathbf{S}$ and $t_j \in \mathbf{T}$, where $j = 1, 2, \ldots, k$. At each discrete site $\{\mathbf{X}(s_i) - u_{s_i} | \mathbf{X}(s_i) > u_{s_i}\}$ can be approximated by the threshold exceedance distribution. For the point process approach, the resulting parameters at each site s_i can be expressed as $\mu(s_i)$, $\sigma(s_i)$, $\xi(s_i)$ and $\zeta_u(s_i)$. Spatial variation in extremes is modelled by requiring that the parameters $(\mu(s), \sigma(s), \xi(s))$ vary smoothly over $\mathbf{S} \in \mathbb{R}^2$. This can be achieved by requiring that

$$\mu(s) = h_\mu(s; \theta_\mu), \quad \sigma(s) = h_\sigma(s; \theta_\sigma), \quad \xi(s) = h_\xi(s; \theta_\xi) \tag{2.6.1}$$

where the parameter vectors $(\theta_\mu, \theta_\sigma, \theta_\xi)$ determine the extent of the spatial variation in each parameter. For example, for the scale parameter the function h can be defined as:

$$h_\sigma(s, \theta_\sigma) = \mathbf{X}^T \theta_\sigma + g(s)$$

where θ_σ is a vector of length p of coefficients of the corresponding predictors. Thus, the extreme value density is implicitly a function of the h functions which consists of a design matrix consisting of spatial co-ordinates s and other predictors which account for the overall spatial trend, as well as smooth polynomial functions of s for local spatial dependence. Penalized likelihood inference can be used to obtain the coefficients of the design matrix and of the interpolating polynomials (Chavez-Demoulin and Davison, 2005). An alternative to this approach is the local likelihood approach, also a non-parametric regression method where smoothing takes place within the local likelihood instead of the parameter space. Butler et al. (2007) applied local likelihood methods to simulated storm surge data for the North Sea. Their model aimed to capture changes in storm surges resulting from physical processes operating at different temporal scales and spatial variation by considering the parameters to be functions of location. Once the coefficients have been obtained, values of the parameters of the extreme value model can be predicted at sites within the region which have not been sampled. The return level is a quantile of the estimated distribution, hence

with parameters of the distribution estimated for all locations in the study region, the return level map is obtained by simple substitution.

The third approach, which is used in this study, involves applying classical geostatistics to rainfall return level estimates (or design values) (Szolgay et al., 2009). The semivariogram models the spatial variation in the design rainfall values (Prudhomme, 1999; Prudhomme and Reed, 1999). The aim of the geostatistical analysis is to obtain an estimate of the design rainfall surface. In this approach, once the semivariogram model is obtained, prediction of the design values at unsampled locations is achieved through kriging. The result is a N year design rainfall map for the study area.

2.7 Computation

In this study the computation was done using mainly R software[4]. The R computing environment provides a suite of integrated functionalities for data manipulation, processing and visualization (R Development Core Team, 2008). R is freely available under the terms of the Free Software Foundation's GNU General Public License in source code form. There are no constraint on the operating system platforms, including the Windows Operating System which was used in this study. R can be interfaced with procedures written in C, C++ and Fortran. R codes or output can also be integrated into other software and programs, such as RExcel in Microsoft Windows Excel, Python and IDL in ENVI and LaTex. This is useful especially in applications of extreme value models for environmental processes, because often the results of the analysis is input into deterministic models which may be programmed in an environment outside R.

The download contains default functionalities known as R-base. For extreme value functions additional packages have to be downloaded as they are not contained in base. These are downloaded from the CRAN site onto the resident R library, then once installed, they can be loaded at the user's discretion. The flexibility of the R environment, enables the user to also compute their own functions should they be not be available or easily found from the base and additional packages. For this study the `POT version 1.1.-0` (Ribatet, 2007) package, was used. Earlier versions of `POT` did not contain functions to fit the point process model, only the GPD could be fitted. This version contains basic functions to fit the point process model, for additional processing one has to either compute new or manipulate existing functions.

There are various other softwares available for such analysis and the list below is not meant to be exhaustive, but provides a glimpse of what is available. A comprehensive review of available software was done by Stephenson and Gilleland (2006). Examples include *Extreme*

[4]The software is downloaded from the Comprehensive R Archive Network (CRAN) site, `http://cran.r-project.org/`

Values In MATLAB (EVIM). S-Plus has a substantial suite dedicated to this type of analysis, namely *Extreme Values In S-Plus (EVIS)* which was developed at ETH in Zurich. S-Plus also has functions to implement Vector Generalized Additive Models (VGAMs), for which the extreme value models are a special class (Yee and Wild, 1996). In R, these functions are packaged as `VGAM`. There are also other packages in R that enable extreme value analysis for which more information can be found on the web-site in CRAN Task Views under Environmetrics. Xtremes is a package made available by Reiss and Thomas (2007) which also makes available estimation through methods other than maximum likelihood. This package is accompanied by StatPasc, a Pascal statistical programming package. Although it is an advantage to have a wide array of software at your disposal, it does present a challenge for the practitioner to chose one that is suitable for the application and whose output can be interpreted to enable identification of bugs and to prevent arriving at incorrect results and conclusions.

An overview of the theory of extreme values has been given, which highlights the research that has contributed to the growth of this field. Whilst the probabilistic theory, initially laid in 1928 has remained undisputed, the last two decades have seen a rapid growth in extreme value application. This may be encouraged by the large number of fields or disciplines where it is important to model and predict rare events. Additionally, growth may be due to the increase in computing power and the availability of algorithms for the models.

In the next chapter, the point process approach to threshold exceedances will be discussed. This method was adopted for studying rainfall extremes in the Western Cape.

Chapter 3

Methods

In this chapter the statistical approach used in this study for the analysis of extremes values of the winter rainfall process at fifteen sites is described. We define what is meant by point processes in Section 3.1, describing properties that are useful in latter sections. Thereafter, the Poisson point process is defined in Section 3.3. An extension of the point process framework to the theory of extreme value is discussed in Section 3.4. The chapter concludes with a discussion on the geostatistical technique used to produce a map of the 50-year 24-hour winter rainfall return values.

3.1 Defining Point Processes

Point processes are stochastic processes where the realizations are point events in a certain parameter space $t \in \mathbf{T}$, which is usually, but not restricted to, time. An example of point processes in time, is the occurrence of rain storms at a particular point location. An example of a point process in a domain other than time, is in pyrostatistics[1], where interest may be on the spatial distribution of lightning strikes in a particular region. Since the pattern of points in space is of concern, the spatial point process is of interest in this case, with the parameter space being in terms of geographic locations. Analyses of point processes in dimensions other than time are practically important, however, for the moment attention is restricted to point processes in time.

Definition 3.1.1. The *random point process* $\{T_1, T_2, \ldots\}$ is defined as a sequence of random variables

$$T_1 < T_2 < T_3 < \cdots$$

such that $P\left(\lim_{i\to\infty} T_i = +\infty\right) = 1$ (Beichelt, 2006).

[1]Pyrostatistics is a term used to describe statistical methods used in fire ecology.

In this study the term 'point process' used, refers to a 'random point process'.

The sequence of inter-event times is constructed from the occurrence process as $Y_i = T_i - T_{i-1}$. This is sometimes referred to as the interval process because it is the waiting time until the next event occurs. While it is important to know the arrival pattern of an event and the expected waiting time until the subsequent event, sometimes it may be more important to know about the distribution of the number of events that occurred in $(0, t]$.

Definition 3.1.2. For a point process $\{T_1, T_2, T_3, \ldots\}$ corresponding to the occurrence of a particular event, the counting measure $\{N(t),\, t \geq 0\}$ of the number of events occurring in $(0, t]$ is defined as:

$$N(t) = \max\{n,\, T_n \leq t\}$$

with state space $\mathbf{Z} = \{0, 1, 2, \ldots\}$.

Any counting point process $\{N(t),\, t \geq 0\}$ has the following properties:

1. $N(0) = 0$
2. $N(s) \leq N(t)$, where $s \leq t$
3. $N(s, t) = N(t) - N(s)$, for any s, t, where $0 \leq s < t$.

The third property means that the increment $N(s, t)$ is equal to the number of events that occur in the interval $(s, t]$. In the above definition, we did not specify the characteristics of the subsets over-which the counting measure is taken. To 'avoid consideration of meaningless events', the subsets of $\mathbb{R}$ over-which the random counting process is taken, are restricted to the *Borel sigma-field* $\mathcal{B}$[2] (Cox and Isham, 1980). An important property of $\mathcal{B}$ is that it is closed under the operations of complements and countable union of it's members, that is measurable sets (Mukhopadhyay, 2000). Consequently, the counting measure N is finite, that is, a finite number of events are recorded in finite time[3].

Another important assumption for point processes is that of *simplicity*. This is imposed by requiring that the counting process be *orderly*,

$$P(N(t, t + h) > 1) = o(h), \quad t \in \mathbb{R} \,. \tag{3.1.1}$$

The *simplicity* of a counting process can be defined as,

$$P(N(\{t\}) > 1) = 0$$

[2]For a Borel sigma-field, $\mathcal{B} = \{A_i \subseteq \Omega,\, i \in \mathrm{I} = \{1, 2, 3, \ldots\}\}$, any subset of the sample space, $A \in \Omega$, is an 'event' if and only if it is an element of the Borel sigma-field.

[3]An 'explosion' occurs, when an infinite number of events are generated in finite time, which is not considered here.

where the set $\{t\}$ consists of the singleton $t \in \mathbb{R}$. Simplicity and orderliness are equivalent, however, for most processes, including the Poisson process, equivalence can be assumed (Cox and Isham, 1980).

The choice between the point, interval or counting process is trivial, because they are statistically equivalent (Cox and Isham, 1980; Beichelt, 2006),

$$\{T_1, T_2, \ldots\} \Longleftrightarrow \{Y_1, Y_2, \ldots\} \Longleftrightarrow \{N(t),\, t \geq 0\} \,.$$

The connection between the counting and the interval process can be made explicit by the relation,

$$P(N(t) > n) = P(T_{n+1} \leq t) \tag{3.1.2}$$

which means that more than n events in $(0, t]$ is possible if and only if the $(n+1)^{\text{st}}$ event from the origin occurs by time t. In this study the counting process representation is used.

Often interest is not only on whether the event occurs, but also the magnitude of the event. In making inference about the behaviour of heavy rainfall, both the frequency and the magnitude of rainfall events are important. If $\{T_1, T_2, \ldots\}$ is the point process, with the random marks M_i assigned to the event time T_i. Then the sequence $\{(T_1, M_1), (T_2, M_2), \ldots\}$ is the *marked point process*. This is a bivariate process containing information on both the frequency and size of a particular event.

In this section basic definitions relating to processes have been discussed. In the next section, an overview of concepts that are important in point process theory, and which are also relevant when Poisson process limit to extreme values is considered.

3.2 Important Concepts

Questions related to how an 'object' behaves on average are endemic in the field of statistics. Often, in quantifying this 'average' behaviour assumptions are made on the nature of the object or 'phenomenon' to be quantified. Often, the assumptions are about the *independence* of the phenomenon, in time, space or in relation, to other phenomena or objects. In this section, stationarity of point processes is discussed, followed by the definition of the intensity of a point process, which provides information on how the process behaves on average. Similar to other statistical methods, these are important concepts in the analysis of extreme values, because often in practice extreme events have been observed to persist.

3.2.1 Stationarity

Complete *stationarity* requires that the process be invariant to absolute shifts in time. Thus, the joint distribution of the inter-event process must be invariant to absolute time shifts.

Complete stationarity is often not practical and it suffices to investigate a particular property of a process to assume stationarity. Consider the counting process $\{N(t),\, t \geq 0\}$, for any $0 \leq s < t$ and $r = 0, 1, 2, \ldots$, the probability distribution of any increment

$$P[N(s,t) = r] = P[N(s, s+h) = r] = p_r(h)$$

depends only on the difference $h = t - s > 0$. Weak (or second order) stationarity is defined when the mean and variance of the counting process are invariant to absolute time shifts. If the interval sequence is strongly stationary, then the corresponding counting process has homogeneous increments. The converse is also true (Beichelt, 2006).

3.2.2 The Intensity of a Point Process

The average or mean, is an important quantity in the study of random variables. Similarly, for point processes, there is value in knowing the *average number of events* in a particular interval. For a counting process $\{N(t),\, t \geq 0\}$, the probability distribution of the increments is given by;

$$p_r(t) = p_r(0,t) = P[N(t) = r]$$

for $r = 0, 1, 2, \ldots$. The *intensity measure* is defined as,

$$\Lambda(t) = E(N(t)) = \sum_{r=0}^{\infty} r p_r(t), \qquad t \geq 0 \tag{3.2.1}$$

When the time axis is re-scaled, such that the occurrence of events is restricted to the interval $(0, 1]$, the intensity measure in $(0, t]$ can be re-express as

$$\Lambda(t) = \int_0^t \lambda(x) dx \; .$$

which is the intensity measure for non-stationary processes. The *intensity function* is $\lambda(\cdot)$. The mean for the stationary process is given by,

$$\Lambda(t) = \int_0^t \lambda dx = \lambda t \; .$$

The intensity measure is central in the point process approach to extreme value theory. Prior to discussing this link, the Poisson point process is defined.

3.3 Poisson Point Processes

The homogeneous Poisson process is the simplest of point processes, whose role in point process theory can be considered analogous to the normal distribution in the study of random variables (Cox and Isham, 1980).

Definition 3.3.1. A counting process $\{N(t),\, t \geq 0\}$ is defined as a homogeneous Poisson process with positive intensity function λ, if it has the properties:

1. $N(0) = 0$
2. $\{N(t),\, t \geq 0\}$ has independent increments
3. $N(s,t) = N(t) - N(s)$ is Poisson distributed, with intensity $\lambda(t-s)$, where $0 \leq s < t$,

The third property means that for any $k = 1, 2, \ldots$ and arbitrary disjoint Borel sets $\{A_1, A_2, \ldots, A_k\}$, the number of points $\{N(A_1), N(A_2), \ldots, N(A_k)\}$ are independently distributed as Poisson, with means $\{\lambda|A_1|, \lambda|A_2|, \ldots, \lambda|A_k|\}$. The mean $\lambda|A_k|$ is the product of the Poisson rate and the size of subset A_k of the sample space.

For the connection between the Poisson distribution and the Poisson process, consider very small sub-intervals $h \to 0$, where the probability of two or more events is zero. With λ being the rate of occurrence of independent events, a single or non-occurrence of an event in $(t, t+h]$ has probability λh or $(1 - \lambda h)$, respectively. If we consider n distinct positions on the real line, then the probability of observing r counts is considered as a BIN$(n, \lambda h)$ event. For large samples and $\lambda h \to 0$ (for $h \to 0$),

$$\begin{aligned} P[N(t, t+h) = r] &= \frac{n!}{r!(n-r)!}\left(\frac{\lambda h}{n}\right)^r \left(1 - \frac{\lambda h}{n}\right)^{n-r} \\ &= \frac{n \times (n-1) \times \cdots \times (n-r+1) \times (n-r)!}{r!(n-r)!} \times \\ &\quad \left(\frac{\lambda h}{n}\right)^r \left(\frac{n}{n - \lambda h}\right)^r \left(1 - \frac{\lambda h}{n}\right)^n \\ &= \frac{(\lambda h)^r}{r!} \prod_{i=0}^{r-1} \left(1 - \frac{i}{n}\right) \left(1 - \frac{\lambda h}{n}\right)^{-r} \left(1 - \frac{\lambda h}{n}\right)^n \\ &\approx \exp(-\lambda h) \frac{(\lambda h)^r}{r!} \end{aligned} \tag{3.3.1}$$

Taking the limit $n \to \infty$, for all fixed $i = 1, 2, \ldots, r-1$, $\lambda > 0$ and non-limit h, such that $\left(1 - \frac{i}{n}\right) \to 1$, $\left(1 - \frac{\lambda h}{n}\right) \to 1$ and $\left(1 - \frac{\lambda h}{n}\right)^n \to e^{-\lambda h}$. This leads to the poisson distribution (Cameron and Trivedi, 1998; Mukhopadhyay, 2000). Hence, the simple counting process with stationary, independent increments is approximately Poisson distributed. Normalizing the exposure interval to unity, i.e. $h = 1$, leads to the usual Poisson density for a random variable, say X. A further remark on the homogeneous Poisson process, is that the intervals are independent and exponentially distributed with parameter λ^{-1}. The homogeneous Poisson process is known as the model for 'complete random scatter' due to the constant intensity function irrespective of the history of the process, and due to the combination of

both the homogeneity and independence of increments, the points are uniformly distributed over the interval $(0, t]$.

In practice, events are often observed where the rate of occurrence changes in time. A process which satisfies all the conditions listed in Definition 3.3.1, except the homogeneity of the increments, is defined as a *non-homogeneous Poisson process* with $\lambda(t)$ as the intensity density function. In the interval $(0, t]$, the intensity measure of the non-homogeneous Poisson process in continuous time is,

$$\Lambda(t) = \int_0^t \lambda(x)dx \ . \tag{3.3.2}$$

If the process is in discrete time, then the integral is replaced by the summation. Essentially a non-homogeneous Poisson process satisfies the properties listed in Definition 3.3.1, with λ replaced by the time varying $\lambda(t)$.

The fundamental property of Poisson processes is that events in disjoint intervals occur independently. Systematic variations (trends, etc.) are admitted through the non-constant intensity function. Independence is the key assumption for Poisson processes, hence for physical processes where there is natural clustering (which is often an indication of a certain degree of dependence in the process) or natural spacing, poisson models may perform poorly and hence methods for compensating for these effects need to be pursued if it is considered impractical to switch to other models.

3.3.1 Maximum Likelihood Estimation of the Poisson Process

When investigating characteristics of any real-world phenomena, unless it is a census, the practitioner has a sample of data from the population of interest. The objective is to draw conclusions about the characteristics of the population by learning information contained in the sample. Hence, the use of models, which are hoped to be simple representations of much more complex behaviour which is unknown. Restricting attention to modelling the non-homogeneous Poisson process, the objective is to estimate model parameters given the observed sample points in an interval $\mathcal{A}$ which is a subset of the real line[4]. Assuming a parametric family $\lambda(\cdot;\theta)$ for the intensity function – concern is on estimating the unknown vector of parameters θ.

Using the maximum likelihood estimation framework, let $I_i = [t_i, t_i + h_i]$ for $i = 1, 2, \ldots, n$ be small intervals based around each observation. Defining $\mathcal{I} = \mathcal{A} - \bigcup_{i=1}^{n} I_i$, thus reducing the complete (unknown) intensity function of the process to the Poisson process intensity. Note

[4]A one-dimensional subset is assumed for simplicity, however, the argument is similar for the non-homogeneous Poisson process in higher dimensions

that the intervals are constructed such that multiple occurrences in an interval are avoided, and interval $\mathcal{I}$ is such that it contains points that are only within region $\mathcal{A}$, ignoring those that fall outside. By the Poisson property, the probability of at least one count in each interval I_i is

$$P[N(I_i) = 1] = \Lambda(I_i; \theta) \exp\{-\Lambda(I_i; \theta)\}$$

where

$$\Lambda(I_i; \theta) = \int_{t_i}^{t_i + h_i} \lambda(u) du \approx \lambda(t_i) h_i \ .$$

Combining the above results

$$P[N(I_i) = 1] = \lambda(t_i) h_i \exp\{-\lambda(t_i) h_i\} \approx \lambda(t_i) h_i$$

for $h_i \to 0$. In the case of no counts of the event in the observation interval, then

$$P[N(\mathcal{I}) = 0] = \exp\{-\Lambda(\mathcal{I})\} \approx \exp\{-\Lambda(\mathcal{A})\}$$

since the increments h_i are small. The likelihood is,

$$\begin{aligned} L(\theta; t_1, t_2, \ldots, t_n) &= P[N(\mathcal{I}) = 0, N(I_1) = 1, N(I_2) = 1, \ldots, N(I_n) = 1] \\ &= P[N(\mathcal{I}) = 0] \prod_{i=1}^{n} P[N(I_i) = 1] \\ &= \exp\{-\Lambda(\mathcal{A}; \theta)\} \prod_{i=1}^{n} \lambda(t_i; \theta) h_i \end{aligned}$$

which results from the independence of counts on disjoint intervals (or Borel sets). Dividing by h_i for the density, the resulting likelihood function is

$$L(\theta; t_1, t_2, \ldots, t_n) = \exp\{-\Lambda(\mathcal{A}; \theta)\} \prod_{i=1}^{n} \lambda(t_i; \theta) \tag{3.3.3}$$

where

$$\Lambda(\mathcal{A}; \theta) = \int_{\mathcal{A}} \lambda(u; \theta) du \ .$$

The log-likelihood is given by

$$l(\lambda(\theta, \mathbf{t})) = \Lambda(\mathcal{A}; \theta) + \sum_{i=1}^{n} \log \lambda(t_i; \theta) \ .$$

The maximization of the log-likelihood requires numerical techniques (Coles, 2001).

3.4 Connection between the Poisson Process and Extreme Value Theory

In practice, extreme values, given that they are not measurement errors, correspond to a class of events that are less likely to occur. Even though the chance that an extreme event will happen is low, the fact is they are possible events, and therefore there is merit in studying their characteristics as a step towards formulating plans to reduce the negative impacts.

> *The law of rare events states that the total number of events will follow, approximately, the Poisson distribution if any event may occur in any of a large number of trials but the probability of occurrence in any given trial is small.*
> – Cameron and Trivedi (1998).

For any event, with potentially negative consequences, the main concerns are in finding out, "when will an event of similar nature re-occur?" and "how large is this likely to be?" Therefore, the two aspects of interest are; the frequency with which the event occurs and the size of the anticipated event. The point process approach to extreme value analysis is based on this concept.

In Section 3.3, the Poisson point process, which is a 'simple' formulation from which more complex models are built was defined. The use of this point process model as a framework for representing extreme events is the purpose of this section, showing how this representation unifies the generalized extreme value (GEV) model for block maxima and the generalized Pareto distribution for threshold excesses. For statistical application, maximum likelihood inference is discussed along the issue of threshold selection. Lastly, manipulations which are necessary to consider as far as autocorrelation at extreme levels of a process is concerned are discussed.

3.4.1 The Poisson Approximation to Extremes

Prior to stating the Poisson limit for extremes, the concept of convergence in distribution for the counting process, needs to be defined.

Definition 3.4.1. Consider a sequence of point processes $N_1, N_2, \ldots$ on an interval $\mathcal{A}$. The sequence $\{N_n\}$ converges in distribution to $\{N\}$ if, for each choice of m and for all bounded Borel sets $\{A_1, \ldots, A_m\}$, the occurrence of points on the boundary (denoted by ∂A) of each set is restricted to zero. The joint distribution of $\{N_n(A_1), N_n(A_2), \ldots, N_n(A_m)\}$ converges to that of the point process $\{N(A_1), N(A_2), \ldots, N(A_m)\}$.

Consider the series of independent and identically distributed random variables $\{X_i;\, i = 1, 2, \ldots\}$, with unknown distribution function F. The Fisher-Tippett theorem (in chapter 2) states that for 'well-behaved[5]' $\{X_i\}$, there exist appropriate sequences of constants such that the normalized sample maximum taken over an appropriate length can be approximated by the generalized extreme value distribution.

The series can be reformulated in 2-dimensions as the bivariate point process $\{(i, X_i);\, i = 1, 2, \ldots, n\}$, where the first dimension refers to the position of X_i in the sequence, and the second dimension is the value attained. As a consequence of the Fisher-Tippett theorem, the behaviour at extreme levels of this bivariate point process can be characterized on regions of the form $A = [t_1, t_2] \times [u, \infty)$. This description is formally expressed as the Poisson process limit theorem, with proof given elaborately by Leadbetter et al. (1983).

Theorem 3.4.1. *Let $\{X_1, X_2, \ldots\}$ be a series of independent and identically distributed random variables for which there are sequences of constants $\{a_n > 0\}$ and $\{b_n\}$ such that*

$$P\left[\frac{M_n - b_n}{a_n} \leq x\right] \to G(x)$$

where

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

with the lower and upper endpoints of G denoted as x_- and x_+ respectively. Then for any $u > x_-$, the sequence of point processes

$$N_n = \left\{\left(\frac{i}{n+1}; \frac{X_i - b_n}{a_n}\right) :\ i = 1, 2, \ldots, n\right\} \tag{3.4.1}$$

converges on regions of the form $\mathcal{A} = (0, 1) \times [u, \infty)$ to a Poisson process with intensity measure on $A = [t_1, t_2] \times [x, x_+)$ given by

$$\Lambda(A) = (t_2 - t_1)\left[1 + \xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}. \tag{3.4.2}$$

As an illustration, consider the operation of *thinning* on a point process. Thinning (sometimes called *splitting*) refers to the removal of points from the original process according to some probabilistic mechanism. In the simple case, the removal is according to a binomial mechanism, where a point is retained with probability p or removed with probability $(1 - p)$, independently of all other points (Cox and Isham, 1980; Leadbetter et al., 1983). In this study, interest is on days with heavy rainfall, meaning daily rainfall levels which are

[5]In the extreme value context, this refers to the sequences being independent and identically distributed.

above the level that is 'normally' received at that particular location specifically in winter. Considering these as a marked point process,

$$N_n = \left\{ \left(\frac{i}{n+1}, \frac{X_i - b_n}{a_n} \right) ;\ i = 1, 2, \ldots, n \right\} .$$

The scaling in the first and second dimension is necessary for the Poisson process presentation and the extreme value limit respectively.

Assuming the marks are mutually independent and independent against the occurrence process – in our case, that is the size of the rainfall events are independent and they do not have influence on the frequency of days with rain. Suppose a threshold u is chosen, high enough to extract events that are extreme, such that the extreme value limit is applicable, but low enough so that model parameters can be attained with acceptable precision (Coles, 2001). Defining mark space,

$$M = \begin{cases} \dfrac{X_i - b_n}{a_n} > u, & p \\ \dfrac{X_i - b_n}{a_n} \leq u, & 1 - p \end{cases} \tag{3.4.3}$$

For a large sample, $n \longrightarrow \infty$ and on bounded region $\mathcal{A} = (0, 1) \times [u, \infty)$, the probability of no exceedance is,

$$\begin{aligned} P(\text{no points in } \mathcal{A}) &= P(N_n(\cdot) = 0) \\ &\approx \exp\{-\Lambda(\mathcal{A})\} . \end{aligned}$$

This is a result of the convergence in distribution of point processes and the Poisson property. Considering the maximum of the sample, $M_n = \max\{X_1, X_2, \ldots, X_n\}$,

$$\begin{aligned} P(\text{no points in } \mathcal{A}) &= P\left(\frac{M_n - b_n}{a_n} \leq u \right) \\ &\approx \exp\left\{ - \left[1 + \xi \left(\frac{u - \mu}{\sigma} \right) \right]^{-\frac{1}{\xi}} \right\} . \end{aligned}$$

Therefore, it can be deduced that,

$$\Lambda(\mathcal{A}) = \left[1 + \xi \left(\frac{u - \mu}{\sigma} \right) \right]^{-\frac{1}{\xi}} .$$

Since the X_i are mutually independent, by the thinning operation given in Equation 3.4.3, is by a binomial mechanism with probability,

$$p \approx \frac{1}{n} \left[1 + \xi \left(\frac{u - \mu}{\sigma} \right) \right]^{-\frac{1}{\xi}} .$$

By the time homogeneity of the Poisson process, for $A = [t_1, t_2] \times [x, \infty)$ and $x > u$, the limiting distribution of $N_n(A)$ is also POI($\Lambda(A)$), with

$$\Lambda(A) = (t_2 - t_1)\left[1 + \xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}} \tag{3.4.4}$$

provided $1 + \xi(x - \mu)/\sigma > 0$.

The connection of the point process model to the classic GEV is derived from the above as follows: for regions $A_x = (0, 1) \times [x, \infty)$,

$$P\left(\frac{M_n - b_n}{a_n} \leq x\right) = e^{-\Lambda(A_x)} \longrightarrow \exp\left\{-\left[1 + \xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\} . \tag{3.4.5}$$

For the distribution of threshold excesses, consider factorizing the region $\Lambda(A_x)$ in Equation 3.4.4 as,

$$\Lambda(A_x) = \Lambda_1([t_1, t_2]) \times \Lambda_2([x, \infty)) .$$

This results from the assumption that event sizes are mutually independent and that they do not influence whether or not an event occurs. Then,

$$\begin{aligned} P\left[\frac{X_i - b_n}{a_n} > x \middle| \frac{X_i - b_n}{a_n} > u\right] &= \frac{\Lambda_2[x, \infty)}{\Lambda_2[u, \infty)} \\ &= \frac{n^{-1}[1 + \xi(x-\mu)/\sigma]^{-\frac{1}{\xi}}}{n^{-1}[1 + \xi(u-\mu)/\sigma]^{-\frac{1}{\xi}}} \\ &= \left[1 + \frac{\xi(x-\mu)/\sigma}{1 + \xi(u-\mu)/\sigma}\right]^{-\frac{1}{\xi}} \\ &= \left[1 + \xi\left(\frac{x-u}{\sigma_*}\right)\right]^{-\frac{1}{\xi}} \end{aligned} \tag{3.4.6}$$

with $\sigma_* = \sigma + \xi(u - \mu)$.

The GEV and the GPD are special cases of the point process model, thus, depending on the problem, one of these models instead of the point process model could be used. The choice of point process representation in this study is based on its flexibility – allowing inferences on both the annual maximum distribution and the threshold excess distribution, to be drawn from the same model. Further, the parameters are in terms of the GEV, hence scale parameter of the model is invariant to the threshold (Coles, 2001).

3.5 Inference for the Point Process Model

Statistical modelling of extremes using the point process approach, entails the selection of a threshold, such that the data above the chosen threshold can be approximated by point

process model. The selection of the threshold is not trivial as this has implications for bias of the model since samples over which the models are fitted are finite. Threshold choice also affects precision of the estimator since a threshold that is too high results in few excesses and therefore large deviations in the estimates. These are issues that were taken into account in applying the model in this study, as discussed in Section 3.5.1. The maximum likelihood estimation technique was used for inference. Discussion on the form of the likelihood function, resulting parameter estimates together with the precision estimates and estimation of the return levels is discussed in Section 3.3.1.

3.5.1 Selecting the Appropriate Threshold

Threshold selection in extreme value analysis often involves subjectivity, hence it is appropriate that the sensitivity of the model parameters at a range of thresholds is evaluated (Coles, 2001; Bierlant et al., 2004). The mean residual life plot and the 'threshold stability' plot that were developed by Smith et al. (1990) have been widely used in applications of threshold exceedance models (Cooley et al., 2007), as an alternative to fixing the threshold at a specific high percentile.

In Section 3.4.1, the connection of the point process model with the GPD was discussed. Using this result, if the approximation by the GPD is appropriate for the excesses of the series $\{X_1, X_2, \ldots, X_n\}$, the mean excess at threshold u_0 is:

$$E(\mathbf{X} - u_0 \mid \mathbf{X} > u_0) = \frac{\sigma_*}{1-\xi}$$

provided that $\xi < 1$ and $u_0 > 0$. The subscript on the scale parameter denote correspondence to the generalized Pareto distribution. As consequence of the stability property of extreme value distributions, the validity of the GPD approximation at u_0, means validity of the approximation at any higher thresholds $u > u_0$, subject to appropriate changes in the scale parameter, as indicated by the prime. Note that the shape parameter should stay the same. Therefore,

$$E(\mathbf{X} - u|\mathbf{X} > u) = \frac{1}{1-\xi}(\sigma'_* + \xi u) \tag{3.5.1}$$

provided that $\xi < 1$ and $(\sigma'_* + \xi u) > 0$. The implication of the relation given in Equation 3.5.1, is that the mean excess is a linear function of the threshold. This idea leads to the following procedure: consider $0 < u < x_{\max}$, the mean excess in Equation 3.5.1 can be approximated by the sample mean excess $\overline{Y}_u$ denoted as,

$$\overline{Y}_u = \frac{1}{n_u}\sum_{i=1}^{n_u}(x_{(i)} - u) \tag{3.5.2}$$

where $\{x_{(1)}, \ldots, x_{(n_u)}\}$ denotes the ordered observations that exceed the threshold. With reference to the theoretical mean of the GPD, the plot of points $\left(u, \overline{Y}_u;\, u < x_{\max}\right)$ is expected to be linear above the threshold where the GPD provides a valid approximation.

This is the *mean residual life plot* (MRL plot), also referred to as the mean excess plot. The intercept and slope of the MRL plot estimate those of the theoretical mean excess function, $\sigma_*^{'}/(1-\xi)$ and slope $\xi/(1-\xi)$ respectively (Davison and Smith, 1990; Coles, 2001; Ribatet, 2007). Confidence intervals of the plot are obtained using the approximate normality of the sample means. The linearity property of the mean residual life plot for suitably high threshold provided basis for using this plot as a graphical tool for threshold selection.

Consider the rainfall series at each station to be $\{X_1, X_2, \ldots, X_n\}$, where n in this study varied because of the varying lengths of data that was available for each station. The MRL plots were drawn for each station, deriving from possible points from which linearity in the plot was observed. This was a subjective choice, hence for each station a range of threshold values were chosen, and the next step was to look at the sensitivity of the model parameters to the threshold. Special attention was paid to the shape parameter, because it is expected to remain the same once the asymptotic distribution has been reached.

The sensitivity of the model parameters to the threshold is observed using the 'threshold stability plot' (or threshold choice plot (Ribatet, 2007)). Reconsider the threshold $u > u_0$, where the scale parameter $\sigma_* = \sigma + \xi(u - \mu)$, which is the scale of the point process model corresponding to threshold excesses, that is, the GPD approximation. This parameter is constant with respect to u, hence, if u_0 is a valid threshold for excesses to follow a GPD, then both σ_* and ξ should be constant above u_0. For finite samples, as in our rainfall study, the two parameters are estimated, and due to sampling variability the property of invariance above a suitable threshold is not possible, however it is expected that they should be stable after allowance for sampling errors. The plots $\{(u, \hat{\sigma}_*); u < x_{\max}\}$ and $\{(u, \hat{\xi}); u < x_{\max}\}$ are used to identify the lowest value u for which the parameter estimates remain near constant (Coles, 2001), using the Fisher Information matrix to obtain confidence intervals.

In practice, there may still be uncertainties regarding the choice of threshold even after using the MRL plot and the threshold stability plot. Shorter ranges can be created, and the sample mean excess and specifically the shape parameter can be plotted against the threshold. Again interest is linearity for the sample mean excess plot, and stability in the estimated shape parameter plot. An additional sensitivity plot can be created using the idea of linearity of the MRL plot. The idea is based on finding a line, such that the deviations of the observed sample mean excesses from this line are small. This line is the estimated theoretical mean function,

$$M_u = \frac{\hat{\sigma}'_*}{1-\hat{\xi}} + \frac{\hat{\xi}}{1-\hat{\xi}}u$$

where the Poisson process limit to extreme values is assumed to be valid from threshold u. The prime is again used here to differentiate the GPD scale parameter obtained for $u > u_0$ from that obtained for u_0. Consider the threshold u_0 and the corresponding parameter estimates, $(\hat{\tau}_0, \hat{\mu}_0, \hat{\sigma}_*, \hat{\xi}_0)$, where τ_0 is the exceedance proportion. Given these parameter estimates, the estimated mean excess values for thresholds $u > u_0$, can be derived from M_u.

Taking the sample mean excess for each threshold u_i as the observed values, the residuals can be calculated as

$$e_i = \overline{Y}_{u_i} - \widehat{M}_{u_i}$$

where $i = 1, 2, \ldots, k$, $\overline{Y}_{u_i}$ is the i^{th} sample mean excess of threshold u_i. The endpoint of the threshold range is k. The root mean squared error (RMSE) can be calculated to give an indication of how close the observed mean excesses are to the line given by the estimated mean function. The same procedure can be done for each threshold, treating it as the "suitable one", fitting the point process model and using the resulting estimates to create the mean excess line, and finally evaluating the closeness of the line to the observed sample means. Interest will be in that threshold for which the RMSE is small.

The Poisson process limit advocates that the number of exceedances over a fixed interval is a Poisson distributed random variate, with the average number of exceedances given by λ for an independent process. Generating counts of excesses over each interval in the study period, these should be distributed as Poisson, hence the mean of the sample of counts should be equivalent to the variance. The *dispersion index* (DI), could therefore be used as a measure for evaluating deviations from the Poisson approximation assumption. This motivates plotting the dispersion index against the threshold as a diagnostic for threshold selection. The dispersion index for each threshold u_i within a range is given by,

$$\text{DI} = \frac{s^2}{\lambda} \, . \tag{3.5.3}$$

The dispersion index ideally should be one, but in practice over- or under-dispersion occurs. To test whether the deviation from unit dispersion is by chance or due to the inappropriateness of the Poisson distribution assumption, confidence intervals are calculated using the chi-square test (Cunnane, 1979; Méndez et al., 2008). That is $P[\text{DI} \in I_\alpha] = \alpha$, where

$$I_\alpha = \left[\frac{\chi^2_{(1-\alpha)/2;(n_{u_i}-1)}}{n_{u_i}-1}, \frac{\chi^2_{1-(1-\alpha)/2;(n_{u_i}-1)}}{n_{u_i}-1} \right] \, .$$

Significant over-dispersion implies the process is more clustered than the Poisson, and significant under-dispersion implies a process that is more regular than the Poisson process. In the plot of DI the aim is to find threshold value/s that are close to one, implying the appropriateness of the Poisson process limit assumption.

The deviance function can be used as a diagnostic. The deviance function,

$$D(\theta) = 2\{l(\hat{\theta}_0) - l(\theta)\} \tag{3.5.4}$$

is a measure for quantifying the uncertainty about the maximum likelihood estimator (Coles, 2001), where $l(\hat{\theta}_0)$ is the maximized log-likelihood. Small deviance for a fitted model, corresponds to high likelihood. In threshold selection, for each threshold in the range the point process model was fitted, and scaled by the number of exceedances of the threshold. This

scaled deviance was plotted against the threshold range. Generally the deviance would decrease as the threshold increases, but due to the rationing by the number of exceedances, it is expected that the scaled deviance would increase as the higher thresholds are considered because there are fewer exceedances at higher thresholds. Therefore the objective in the scaled deviance plot is to look for threshold values where the increase is at a constant rate and where there are dips.

The main aim in fitting extreme value models is to estimate the return level (or design values as it is termed in engineering). The sensitivity of the return level estimates to the choice threshold value needs to be investigated, as important decisions are made based on this value, and it is important that a stable estimate is obtained. A diagnostic plot of the N-year return level against a range of threshold, can be done to evaluate the stability.

All these plots are diagnostic measures and are used precursor to the actual model fitting. Once the threshold has been chosen, the model is fitted through maximum likelihood techniques as discussed in the following section.

3.5.2 Estimation in the Point Process Model

One of the usefulness of the point process model for extreme values is that non-stationarity can be incorporated with ease, and the maximum likelihood estimator is most suitable when structural models are to be fit (Katz et al., 2002). Consider a region $\mathcal{A} = (0,1) \times [u, \infty)$, where u is the threshold that has been selected using methods described in Section 3.5.1. The observed exceedances can be relabeled as $\{(t_1, x_1), (t_2, x_2), \ldots, (t_{N(A)}, x_{N(A)})\}$. To express the extreme value limits in annual terms, that is annual maxima, then an adjustment is made by multiplying $\Lambda(A)$ (with functional form given in Equation 3.4.4) by a factor n_y giving the number of years of observation. Assuming the Poisson process is an acceptable approximation, following the arguments presented in Section 3.3.1, results in the likelihood function,

$$\begin{aligned} L(A; \mu, \sigma, \xi) &= \exp\{-\Lambda(A)\} \prod_{i=1}^{N(A)} \lambda(t_i, x_i) \\ &= \exp\left\{-n_y \left(1 + \xi \frac{u-\mu}{\sigma}\right)^{-\frac{1}{\xi}}\right\} \times \prod_{i=1}^{N(A)} \frac{1}{\sigma} \left(1 + \xi \frac{x_i - \mu}{\sigma}\right)^{-\frac{1}{\xi}-1} \end{aligned}$$

and taking the logarithm, the negative log-likelihood function is

$$\begin{aligned} -\, l(A; \mu, \sigma, \xi) &= n_y \left(1 + \xi \left(\frac{u-\mu}{\sigma}\right)\right)^{-\frac{1}{\xi}} + n_u \log \sigma + \\ &\quad \sum_{i=1}^{n_u} \left(\frac{1}{\xi} + 1\right) \log\left(1 + \xi \frac{x_i - \mu}{\sigma}\right) . \end{aligned} \tag{3.5.5}$$

The parameter estimates are obtained by minimizing this negative log-likelihood. The parameters correspond to the distribution of the maxima, namely, the GEV. The parameters for the GPD, in particular the scale parameter is obtained through the relation $\hat{\sigma}_* = \hat{\sigma} + \hat{\xi}(u - \hat{\mu})$. The shape parameter and the threshold exceedance proportion are the same for the GPD and GEV. The maximum likelihood estimator of the threshold exceedance proportion is,

$$\hat{\tau} = \frac{n_u}{n} \ .$$

If interest is directly in the threshold excess distribution, the log-likelihood can be written from the result of Equation 3.4.6, to get the GPD parameter estimates directly. The log-likelihood function is given in Section 2.5.1.

In practice, EVT is often used to obtain an estimate of the design level or *the return level*. This measure enables the quantification of return periods of events that may be of interest, for example, in the construction of a dam, the magnitude of the design flood and the maximum possible rainfall of different durations are important quantities. The 1-in-N year return level corresponds to the $(1 - 1/N)$ quantile of the fitted distribution. Denoting this quantile by x_N, in the case of threshold exceedances,

$$P(X > x_N) = P(X > u)P(X > x_N \mid X > u) = \frac{1}{N} \ .$$

Hence, the N-year return level is,

$$x_N = u + \frac{\sigma_*}{\xi}\left[(\tau n_y N)^{\xi} - 1\right] \ . \tag{3.5.6}$$

Estimation of this quantity is by substitution of the MLE's $(\hat{\tau}, \hat{\sigma}_*, \hat{\xi})$. Since the return level is a function of the GPD parameters, the delta method can be used to obtain the confidence interval, however in this study profile likelihood confidence intervals were calculated. They have been shown in practice to be suitable for extreme value distributions because they are asymmetrical and therefore take into account the skewness of the distribution (Coles, 2001; Smith, 2003). To obtain the profile likelihood, x_N, Equation 3.5.6 was re-expressed as

$$\sigma_* = \frac{\xi(x_N - u)}{(\tau n_y N)^{\xi} - 1} \tag{3.5.7}$$

so that the log-likelihood 2.5.2 is written as a function of (x_N, τ, ξ), the scale parameter being replaced by the function defined in Equation 3.5.7. Maximization of the log-likelihood with respect to τ and ξ leads to a function of x_N, namely the profile likelihood function $l_{\mathbf{X}}(x_N)$. Details on the profile likelihood are found in appendix 2.5.1. Provided that the regularity conditions hold, i.e. $\xi > -0.5$, the $100(1-\alpha)\%$ profile confidence interval for x_N consists of all values for which

$$l_X(\hat{x}_N) - l_X(x_N) \leq \frac{\chi^2_{1;1-\alpha}}{2} \ .$$

The same idea of the profile likelihood function can be used for estimating the confidence intervals for the parameters.

The Poisson process limit to extreme values is based on the assumption of independent and identically distributed sequence of random variables. This is easily violated by time series rainfall values because rainfall is correlated in time, as it can be seen from the tendency of rainfall to occur on successive days, the seasonal behaviour and the long-term cyclical behaviour as an influence of oceanic convection. The latter two characteristics are the result of differences in systematic behaviour in time, namely, the sequence of random variables not having the same marginal distribution. As discussed in chapter 2, there are methods of dealing with non-stationarity specifically for extreme values. In this study only the winter season is considered, hence assuming that the process is stationary, it is the effect of short-range temporal correlations that needs to be examined. The effect of ignoring temporal dependence when modelling extreme values is the risk of incorrect estimation of the return levels (Bierlant et al., 2004, for details).

To account for short-range temporal dependence, the extremal types theorem is modified, by assuming the existence of a condition that limits the extent of long-range dependence at extreme levels of a process. This is termed the $D(u_n)$ condition, formulated by Leadbetter (1983); Leadbetter et al. (1983) and defined below.

Definition 3.5.1. A stationary series $\{X_1, X_2, \ldots\}$ is said to satisfy the $D(u_n)$ condition if, for all $i_1 < \cdots < i_p < j_1 < \cdots < j_q$ with $j_1 - i_p > l$,

$$\begin{aligned} &| P\{X_{i_1} \leq u_n, \ldots, X_{i_p} \leq u_n, X_{j_1} \leq u_n, \ldots, X_{j_q} \leq u_n\} \\ &-P\{X_{i_1} \leq u_n, \ldots, X_{i_p} \leq u_n\} P\{X_{j_1} \leq u_n, \ldots, X_{j_q} \leq u_n\} \mid \leq \alpha(n, l) \end{aligned} \tag{3.5.8}$$

where $\alpha(n, l_n) \to 0$ for some sequences $l_n = o(n)$.

While dependent sequences are qualitatively different from independent sequences, the $D(u_n)$ conditions allows that the extreme value limit laws to still hold, for sets that are sufficiently apart. Since the limit laws are not affected, the shape parameter remains the same, but the other parameters change when accounting for the dependence. As a result the relationship between the extreme value model of the dependent and the independent sequence is as follows:

$$G_{\text{dep}}(z) = \left(G_{\text{indep}}(z)\right)^{\theta} \tag{3.5.9}$$

where $0 < \theta \leq 1$ is termed the extremal index.

The extremal index is a measure of the tendency of the process to cluster at extreme levels, where the reciprocal is the limiting mean cluster size. For independent sequences $\theta = 1$, but the converse is not necessarily true.

As consequent of the $D(u_n)$ condition, the Poisson process limit remains valid for exceedances

of the threshold. However, the tendency of dependent observations to cluster, means that the exceedance probability needs to be reduced by a factor θ. There are various ways in which the extremal index is estimated based on declustering techniques, the simplest being the *runs* declustering method. In this method, clusters are formed by arbitrarily specifying run length r, such that a cluster is considered active until r consecutive values fall below the threshold u. Then, instead of modelling the sequence of all threshold excesses with the GPD, only the sequence of maxima taken from each cluster is modelled. The idea is that the cluster maxima are sufficiently far apart to warrant the assumption that they are independent, hence the applicability of the extreme value model. The extremal index is then estimated as the quotient of the number of clusters over the number of exceedances of the threshold u.

The choice of run-length r affects the bias-variance trade-off. A value of r that is too small raises concerns over the validity of the assumption of independence of cluster maxima. Conversely, large values of r could result in too few cluster maxima, hence, raising concern over the precision of the GP distribution's parameter estimates. Therefore an arbitrary choice of the run-length, may introduce uncertainty regarding the quality of the model's outputs. The method of Ferro and Segers (2003) aims to reduce this uncertainty by an optimal estimation of run length and the extremal index, based on the distribution of the inter-exceedance times of the process.

Once short-range temporal dependence has been accounted for by considering the point process extreme value model for just the cluster maxima, return level estimates are obtained. The process for this study therefore involves obtaining the point process extreme value model at each site, firstly assuming that the daily rainfall values are independent. Thereafter, those exceedance series which show evidence of temporal correlation are subjected to declustering and then application of the model to the cluster maxima. This leads to the distribution of rainfall extrema and therefore return level estimates for each site. However, to gain insight about the dependence of the N-year return levels over the study regions, the methods have to be taken a step further. Some of the common ways to model extreme values spatially include: spatial interpolation of the return levels, interpolation of the parameters of the extreme value model, considering the observations in space as a max-stable process and using techniques of multivariate extreme value theory to obtain a measure of spatial dependence and the latent process approach, which is a hierarchical approach where marginal extreme value models are obtained at each site, then spatial dependence is induced by considering the obtained parameter estimates as a realization of a latent spatial process. In this study the purpose is to investigate whether the return level estimates shows evidence of spatial variation. Therefore, geostatistical methods on the return level estimates are considered.

3.6 Geostatistical Model for the Return Levels

In the previous section univariate extreme value methods leading to derivation of return level estimates at each site were discussed. Another interest in this study is to quantify spatial variation in return level values. Site-wise N-year 24-hour rainfall return level (or design rainfall) estimates form sample points over the study region which are used in the spatial analysis. These site-wise return level estimates are assumed to be a single realisation of an unknown underlying spatial process which is continuous over the study region. The intention is to construct the N-year return level surface over the entire region, that is, predict the N-year 24-hour design rainfall at sites that were not sampled.

3.6.1 Estimating the Semivariance

Design values at each site are assumed to be a single realisation from a random process which is continuous over the study area. This continuous surface or spatial random field is denoted as $\{Y(\mathbf{s}, \mathbf{t}) : \mathbf{s} \in \mathbf{D} \subset \mathbb{R}^2, \mathbf{t} \in \mathbf{T}\}$, where $\mathbf{D}$ is a fixed, continuous subset of a two-dimensional plane and $\mathbf{t} \in \mathbf{T}$ is a vector that represents the temporal component. The primary objective is to make inference about the unknown underlying design rainfall process from which we assume our sample to be generated. An important property in spatial analysis is that the strength of the association in attribute values decreases as the distance between measurement locations increases, i.e. spatial correlation.

> *. . . the sensitivity of geographic and other phenomena to local interactions implies that we should carefully measure and analyze relations among near things*
> – (Miller, 2004)

The covariance statistic is useful in quantifying spatial correlation, however, it is common in geostatistics to use the variogram $2\gamma(\mathbf{h})$ which is also a measure for the second-order property of the underlying process. The autocovariance and the semivariogram are related as follows:

$$\gamma(\mathbf{h}) = C(\mathbf{0}) - C(\mathbf{h})$$

where $\mathbf{h}$ is a vector of distances between locations.

In geostatistics, it is assumed that spatial correlation can be represented by some parametric model. Estimation of the parameters of the variogram model when the sample size is small may lead to unreliable results (Cressie, 1993). With design rainfall values at just fifteen sites, estimation and prediction uncertainty is likely to be high. The method of regionalisation as described in (Stein and Sterk, 1999; Sterk et al., 2004) is used to extend the observations in space and time, thereby increasing the size of the sample to reduce

the amount of uncertainty in variogram model estimation. Design values are considered as space-time observations, with space relating to their position on the earth's surface and time represented as a sequence of return periods.

A common assumption in statistics is that of 'independent and identically distributed' random variables. Its counterpart in geostatistics is *strict stationarity*, where if any set of locations are shifted spatially by $\mathbf{h}$, then observations from the two sets of locations will have the same probability density. Strict stationarity can be relaxed by requiring the mean and covariance to be stationary, that is a constant mean and covariance which only depends on the distance between any two locations. A weaker assumption, termed *intrinsic stationarity*, requires that a process have a constant mean and a stationary variogram. Another property that is required in order to be able to estimate the parameters of the model for the underlying unknown spatial process is *ergodicity*. This means that the single realization of the process must be able to wander through all possible values that the process can take. Hence, in spatial statistics stationarity and ergodicity are necessary assumptions.

In this study, the space-time field from which the design rainfall values are sampled is assumed to be intrinsically stationary and ergodic (Stein and Sterk, 1999). That is

$$\begin{aligned} 0 &= E_{\mathbf{s},\mathbf{t}}[Y(\mathbf{s}+\mathbf{h_s},\mathbf{t}+\mathbf{h_t}) - Y(\mathbf{s},\mathbf{t})] \\ 2\gamma(\mathbf{h_s},\mathbf{h_t}) &= \text{var}(Y(\mathbf{s},\mathbf{t}) - Y(\mathbf{s}+\mathbf{h_s},\mathbf{t}+\mathbf{h_t})) \end{aligned}$$

To tackle the challenge of having a small sample size for variogram estimation, the approach of Stein and Sterk (1999); Sterk et al. (2004) is followed. Expansion of the sample spatially is achieved by considering replication of the design values in pseudo-time, where $Y(s_i, t_j)$ is a return level for the $t_j : j = 1, 2, \ldots, p$ return period at site $s_i : i = 1, 2, \ldots, n$. That is, at each site there is a sequence of return level estimates indexed by $\mathbf{t_j} : j = 1, 2, \ldots, p$ equally spaced return periods used as pseudo-time. This is equivalent to having p realisations of the underlying spatial process (or p return level surfaces). It is assumed that the return periods satisfy the requirement that $t_k \bigcap t_j = \emptyset$ for $k \neq j$. To satisfy the requirement of a constant mean for intrinsic stationarity, the data is standardized by the ratio of the overall average design rainfall (m_0) and the average design rainfall for that specific period (m_{t_j}), $j = 1, 2, \ldots, p$. The standardized design values at site $s_i : i = 1, 2, \ldots, n$ are denoted as

$$\tilde{Y}(s_i, t_j) = Y(s_i, t_j) \times \frac{m_0}{m_{t_j}} \tag{3.6.1}$$

This removes the effect of temporal replication, which is a 'step-like' effect on return level estimates at each site caused by the fact that as higher return periods are considered, higher return levels are obtained. Upon removal of this effect through standardization by the ratio of the means as given in Equation 3.6.1, it becomes plausible to assume that temporal lag-effects are negligible, i.e. $\mathbf{h_t} = 0$. That is for the standardized data the mean return level surface is invariant to changes in return periods. This reduces the space-time variogram

into a spatial variogram. At each location there are p standardized observations. These can be extended in space, by displacing the set of locations by a fixed distance c repetitively for $p-1$ instances. The displacements are as follows $\{cs_i, 2cs_i, \ldots, (p-1)cs_i\}$ where s_i is the index for the original coordinates. The dispersal in space is justified by the property of spatial autocorrelation, where spatial variation is a function of distances between paired locations rather than exact locations. In this way a larger data set is created for the purpose of choosing a variogram model and estimating its parameters. Three models are considered in this study: the exponential, the spherical and the penta-spherical models. The parametric representations of these models can be found in Cressie (1993). The penta-spherical model is given by

$$\gamma(\mathbf{h}) = \begin{cases} C_0 + C\left(\frac{15}{8}\frac{\mathbf{h}}{a} - \frac{5}{4}\left(\frac{\mathbf{h}}{a}\right)^3 + \frac{3}{8}\left(\frac{\mathbf{h}}{a}\right)^5\right), & 0 \leq \mathbf{h} < a \\ C_0 + C, & \mathbf{h} \geq a \end{cases} \tag{3.6.2}$$

where C_0 is the nugget effect, composed of micro-scale variation and measurement error. The partial sill is C and the range is a. The range is the distance beyond which there is lack of spatial correlations between values. Estimation in this study is by weighted least squares (WLS) (Cressie, 1993). It is important to note that in WLS, the weights increase with the number of pairs in each lag class and that the weights increase near the origin. Hence it is important for the variogram to fit well near the origin. The weighting factor used in this study is $N_j/[\gamma(h_j)]^2$, the ratio of the number of point pairs at distance lag h_j to the square of the semivariance at that lag.

Estimates of the partial sill and nugget for a particular return period are obtained by multiplying estimates of the variogram model for the pooled data (that is the extended sample) with the square of the reciprocal of the standardization factor. The re-scaled parameters are input in the kriging procedure to obtain the desired design rainfall map. Kriging is a spatial prediction technique and details on its application in this study are discussed in the next section.

3.6.2 Spatial Prediction

Kriging is a well-known generalized least-squares technique that allows one to account for spatial dependence in the observations as given by the variogram model (Goovaerts, 2000). There are several texts detailing this theory (Isaaks and Srivastava, 1989; Cressie, 1993). Basically in kriging, design values at unsampled locations will be predicted by the predictor

$$Y(s_0) = \sum_{i=1}^{p} \lambda_i y_i, \quad \text{with a constraint on weights } \sum_{i=1}^{p} \lambda_i = 1 \tag{3.6.3}$$

Kriging is known as the Best Linear Unbiased Predictor (BLUP), because the kriging weights are chosen so as to minimize the estimation variance $\mathrm{Var}(Y(s_0) - Y(s))$, whilst ensuring that the estimator is unbiased, i.e. $\mathrm{E}(Y(s_0) - Y(s)) = 0$. *Linearity* is a result of the predictor

being a linear combination of attribute values from un-sampled neighbouring locations or sites. Weights are obtained by solving a kriging system of equations as detailed in Cressie (1993).

In the case of ordinary kriging as described above the spatial process is assumed constant over the study region. If there is evidence of a global spatial trend, the random process at location s can be defined as

$$U(s) = X^{\mathrm{T}}\beta + \epsilon(s) \tag{3.6.4}$$

The model for spatial variation due to a non-constant mean of the process is given by

$$\mathrm{E}(U(s)) = \sum_{k=1}^{q} \beta_k f(s) \tag{3.6.5}$$

where q is the number of beta coefficients. For variogram estimation, initially parameters of the trend surface model given in Equation 3.6.5 are estimated. Once the global trend has been removed any remaining spatial correlation is detected as variation in residuals and modelled through a residual variogram model. Estimation of the parameters of that residual variogram continues similarly to the constant mean case.

For the universal kriging predictor, the value to be predicted can be expressed as the linear combination of measured values

$$U(s_0) = \sum_{i=1}^{p} \lambda_i u(s_i)\ . \tag{3.6.6}$$

The kriging weights are obtained by minimizing

$$\left(U(s_0) - \sum_{i=1}^{p} \lambda_i U(s_i)\right)^2 = \left(\mathbf{X}^{\mathrm{T}}\beta + \epsilon(s_0) - \lambda^{\mathrm{T}}\mathbf{X}\beta - \sum_{i=1}^{p} \lambda_i \epsilon(s_i)\right)^2 . \tag{3.6.7}$$

Universal kriging is also considered in this case study with the results compared to the ordinary kriging case. To derive the ordinary kriging map for the 50-year 24-hour winter rainfall return level, the parameters are re-adjusted by the reciprocal of the standardizing ratio to derive the corresponding variogram model parameters. Kriging is applied to the fifteen 50 year return level estimates using these parameter to obtain the 50-year 24-hour winter rainfall return level surface.

In this chapter we discussed the extreme value approach that is used to analyse heavy rainfall at each site. The output of that exercise is an estimate of the rainfall return level at each site. This is enough information if the objective is to describe the behaviour of heavy rainfall at each site in isolation. However, for the purpose of this study this is not enough as one of the key research objective was to investigate whether the fifty year rainfall return level is homogeneous over the study region. To achieve this objective, spatial correlation of

return level estimates is modelled with resulting parameter estimates used to derive a kriging map for the 50-year 24-hour winter rainfall return level over the study area. The case study and the results obtained are explained in more detail in the next chapter.

Chapter 4

Applying the Point Process Extreme Value Approach

The Western Cape is climatologically diverse with distinct micro- and macro climates caused by the varied topography and the influence of the South West Indian and South Atlantic Oceans' circulation. The province is classified as a Mediterranean-type climate with cold and wet winters, and warm and dry summers. Thunderstorms are occasional with most of the precipitation being of frontal and orographic[1] origin. Frontal activity is intense during the winter season, hence the contribution of total winter rainfall to total annual rainfall is high for this province.

In this chapter the data are described, giving insight into the degree of incompleteness, a summary of the observed data at each site as well as commonalities present across sites within the region. The application of the point process extreme value method requires the selection of a suitable threshold. This is not a trivial task as current methods involve some degree of uncertainty. The suite of methods used in choosing a plausible threshold are discussed. Finally fitting of the point process extreme value model at each site is discussed.

4.1 Description of the Rainfall Data

Observed daily rainfall data (in mm) were obtained from the South African Weather Service (SAWS)[2]. SAWS has a ground-based precipitation network where rainfall is recorded mainly with two instruments. The first is the conventional rain gauge, which involves manual recording of the rainfall amount by the observer. The second method is through the Automatic

[1] Orographic precipitation, also known as relief precipitation, is precipitation caused by a forced upward movement of air upon encountering elevated terrain (e.g. hills, mountains).

[2] www.weathersa.co.za

Weather Stations (AWS) which became operational from the late 1980's (Kruger, 2007). A day runs from 08:00 am to 08:00 am the following day, instead of midnight to midnight. This is to accommodate the working hours of the observers. The spanning across two calendar days, the possibility of errors when the observers manually record the rainfall values and the possibility of error when quality control checks are done on the AWS record, are a concern with regards to the quality of the data. However, measures are undertaken by SAWS to ensure that the quality of the data is of acceptable level (Kruger, 2007). It will be assumed that the data used here are of acceptable quality.

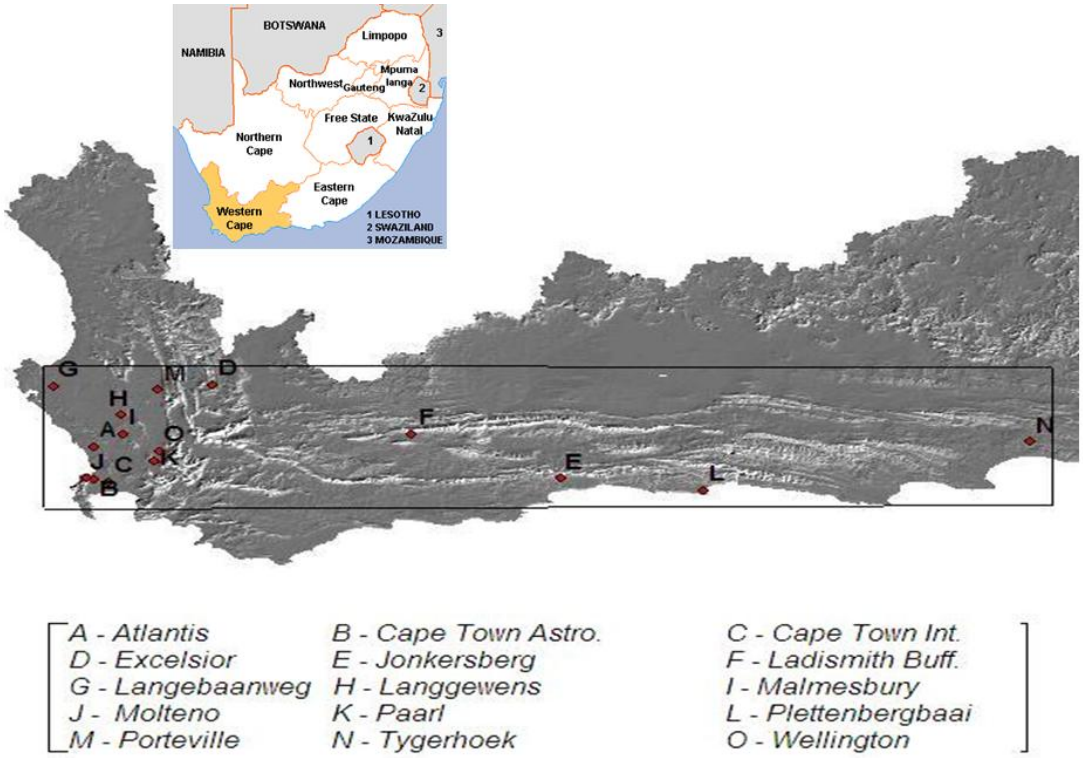


Figure 4.1: Location of the fifteen weather stations in the Western Cape region

Fifteen rainfall data series were obtained from the weather stations in the Western Cape. The selection of these stations was based on the criteria that the data should be of good quality as determined by SAWS, with low level of incompleteness. Hence, only seven stations had 50 years of daily rainfall observations, the other eight stations had shorter series.

Non-zero rainfall values were recorded if the amount that fell was at least 0.1 mm. Due to quality issues already stated, daily accumulated rainfall values were recorded, where it could not be ascertained whether rainfall fell on prior days (usually 1 to five days), these

were recorded as '0.0 A'. The days where 'A' was recorded were usually followed by 'C' which meant that the record for that day could be an accumulated amount for the previous 'A' daily recordings. On more than 90% of the occasions, the accumulated values were found to be at most 5 mm. Therefore, the suspect zero recordings were left unchanged and the 'potentially' accumulated values were taken to be rainfall amounts recorded for that day. This shortcoming with respect to the data set is not expected to have an impact on the inference based extreme value models, as these values are expected to fall below the threshold in almost all the sites because their size is small relative to data that is considered extreme.

A description of the data which also highlights the quality issues of incompleteness and suspect observations is included in Table 4.1. In the table 'Period' corresponds to start and end dates of the series for each station. The column 'Miss' gives the percentage of the data that was missing for each site within the period of observation for each site. The last two columns give the percentages of the data that was considered suspect and accumulated as explained earlier in this section.

Table 4.1: Details on the rainfall data obtained from South African Weather Services

Site	Long.	Lat.	Altitude	Period	Sample Size	Miss (%)	Susp. (%)	Accum. (%)
Atlantis	18.483	-33.607	121	10/1979 - 12/2007	2576	3.45	0.19	0.04
Cape Town Astro.	18.477	-33.935	15	01/1958 - 12/2007	4543	1.24	7.40	3.61
Cape Town Int.	18.597	-33.969	44	01/1958 - 12/2007	4600	0.00	0.00	0.00
Excelsior	19.43	-32.963	958	03/1993 - 12/2007	1378	6.39	0.00	0.00
Jonkersberg	22.227	-33.934	325	01/1958 - 12/2007	4600	0.00	0.59	0.30
Ladismith	21.035	-33.476	400	01/1985 - 12/2007	2116	0.00	0.00	0.00
Langebaanweg	18.157	-32.972	31	03/1973 - 12/2007	3091	4.01	0.13	0.19
Langgewens	18.706	-33.276	179	01/1958 - 12/2007	4600	0.00	0.02	0.02
Malmesbury	18.718	-33.472	108	03/1993 - 12/2007	1374	6.66	0.00	0.00
Molteno	18.417	-33.933	93	01/1958 - 12/2007	4600	0.00	0.00	0.02
Paarl	18.967	-33.75	145	01/1958 - 10/1998	3681	2.41	0.00	0.00
Plettenbergbaai	23.372	-34.058	73	01/1958 - 12/2007	4598	0.04	1.20	0.33
Porterville	18.994	-33.012	142	01/1959 - 10/2007	4508	0.00	3.79	1.66
Tygerhoek	25.993	-33.553	457	01/1958 - 12/2007	4599	0.02	0.11	0.04
Wellington	19.006	-33.651	176	04/1988 - 12/2007	1836	0.22	0.05	0.11

The analysis is restricted to the winter season, hence the data description in Table 4.1 corresponds to data for the months June, July and August of each year. The percentage of missing observations is generally low at below 3%, with four of the stations (Atlantis, Excelsior, Langebaanweg and Malmesbury) having higher proportions of missing values, but even in these cases, the proportion is less than 0.1. Therefore, the effect of the incompleteness was assumed to be minimal for the analysis.

To better understand the data, scatter-plots are drawn (Fig 4.2 and 4.3) for each station. Some of the plots appear truncated because these are for the sites that had shorter series as given in Table 4.1. No clear trends are evident from these plots, which concurs with previous studies (Mason et al., 1999; Kruger, 2007) that there is no strong evidence in support of trends in daily rainfall observations. Rainfall values above 60 mm occur infrequently and values above 100 mm are rare. The exception is Tygerhoek (Fig 4.2), where values in excess of 100 mm have occurred frequently during the study period. Tygerhoek is situated in the south-west Cape region, which is the transition zone from the predominantly winter rainfall, to the predominantly summer rainfall region of South Africa. This region is classified as an all-season rainfall region (Preston-Whyte and Tyson, 1988). The highest observed value is 261.1 mm for this station. The data reveals that 89.2 mm, 261.1 mm and 61.7 mm of rain fell on 20, 21 and 22 August in 1971, respectively. Ladismith appears to have predominantly low rainfall values in comparison to the other stations.

Seasonality is characteristic of rainfall in Western Cape. Most of the province receives a large proportion of it's rainfall in the winter months of June, July and August. The typical monthly profile for most of the province is similar to that of Cape Town in Fig 4.4(a). There is a clear peak in the winter months. High daily rainfall values are also observed during early spring and late autumn. In moving east along the southern Cape coast, to Tygerhoek, the profile changes. This region is classified as an all rainfall season, as is evident from Fig 4.4(b), which shows a monthly profile which lacks a clear peak. The three largest observed rainfall values for this station were in August. In this study stationarity is assumed and since interest is in heavy rainfall events, only the winter season, for which most of the province receives a large portion of the total annual rainfall, is considered.

Information on the distribution of daily rainfall at each station is obtained from the box-plots for each station in Fig 4.5. This corresponds to non-zero daily rainfall series at each station. The varying thickness of the boxes corresponds to the varying lengths of the series of non-zero observations at the different locations. The distribution of data at each station shows positive skewness, typical of heavy-tails and is often observed for hydrological processes (Katz et al., 2002). Looking at the quartiles and the variability at each station, there appears to be two groups as described below:

- Relatively low values for Ladismith, Langebaanweg, Langgewens and Malmesbury. The median is low (less than 10 mm), compared to the rest of the stations. There is very little variation and the outlying observations are low in comparison to the other stations.
- The variation is higher for the remaining stations, with unusually large rainfall values occurring frequently, especially for Paarl and Tygerhoek.

The descriptive plots discussed above highlight the variability of rainfall over the study area. Large extremes are expected at stations like Tygerhoek and Paarl, while stations such as

Ladismith are expected to have comparatively smaller extremes. These characteristics will affect the choice of threshold. Data-driven methods will be used in the selection of the threshold, thus it is expected that the resulting thresholds will reflect the characteristics shown in Fig 4.5. We shall assume stationarity for the subset of data used in this study.

In the next section methods that are used to find the appropriate threshold are discussed.

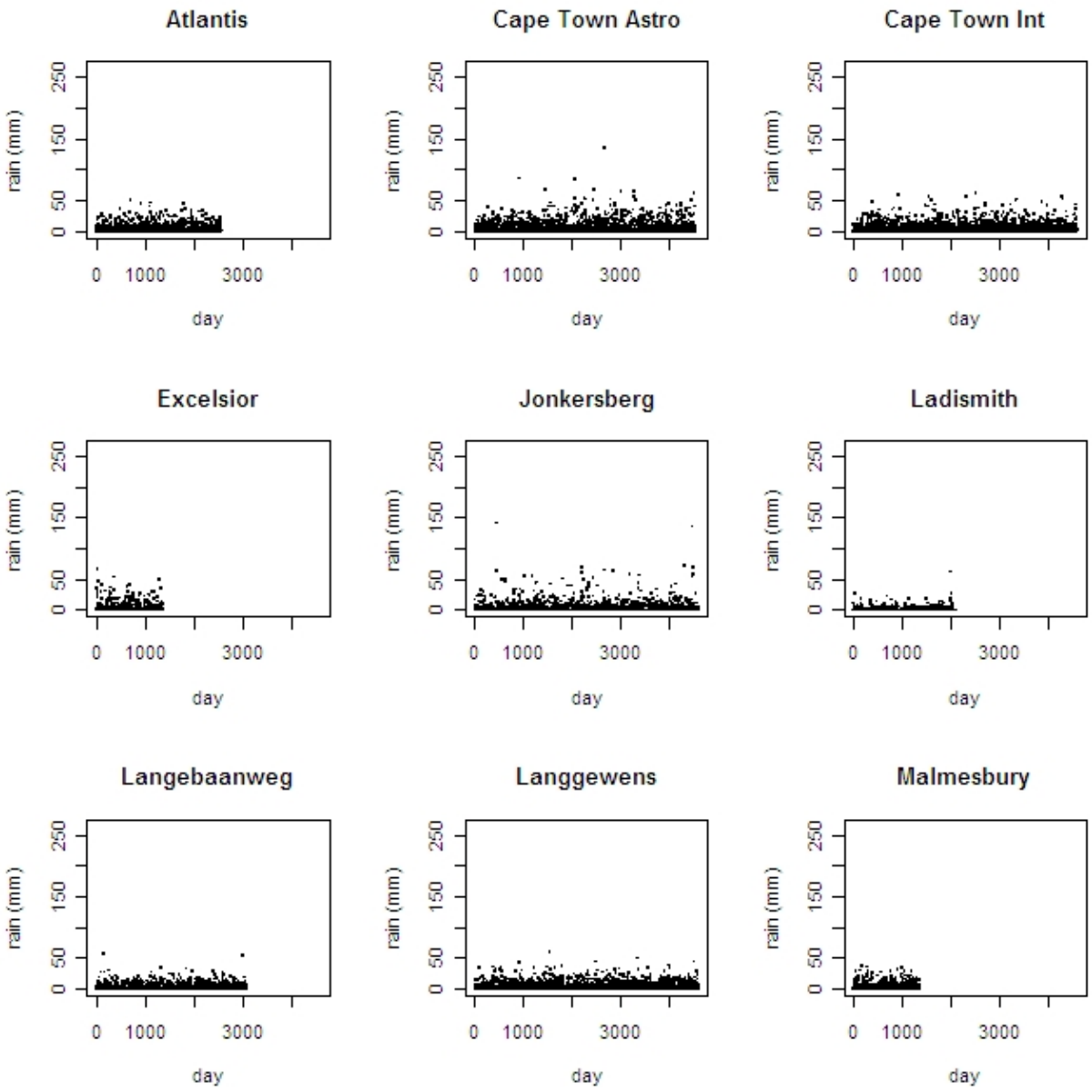


Figure 4.2: Daily rainfall values (winter): Atlantis – Malmesbury

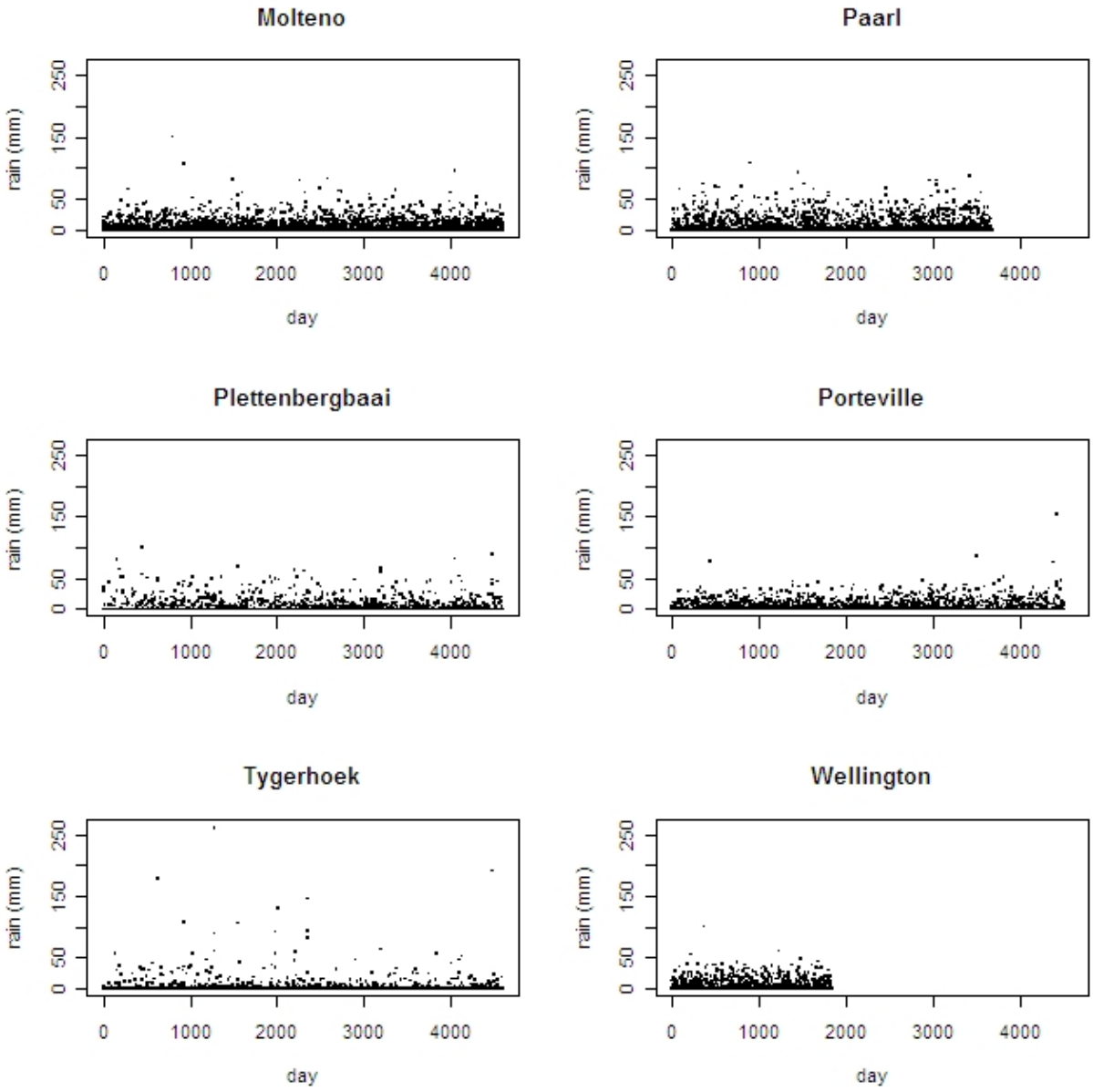
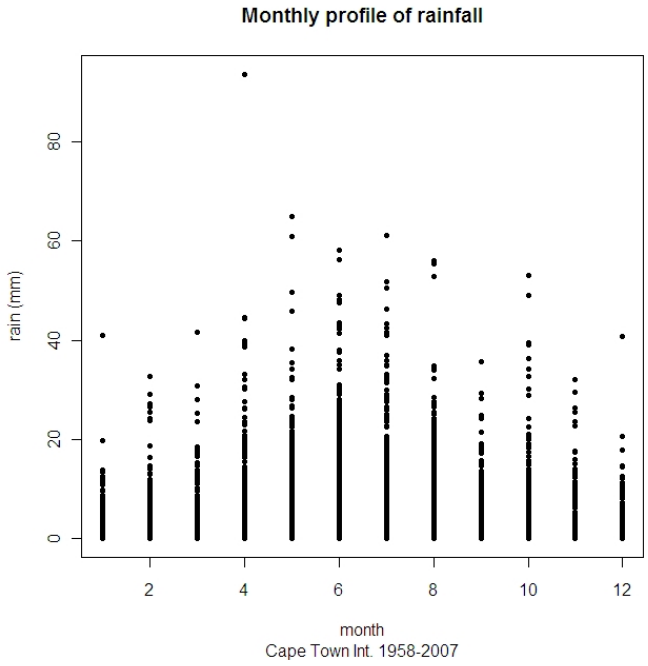
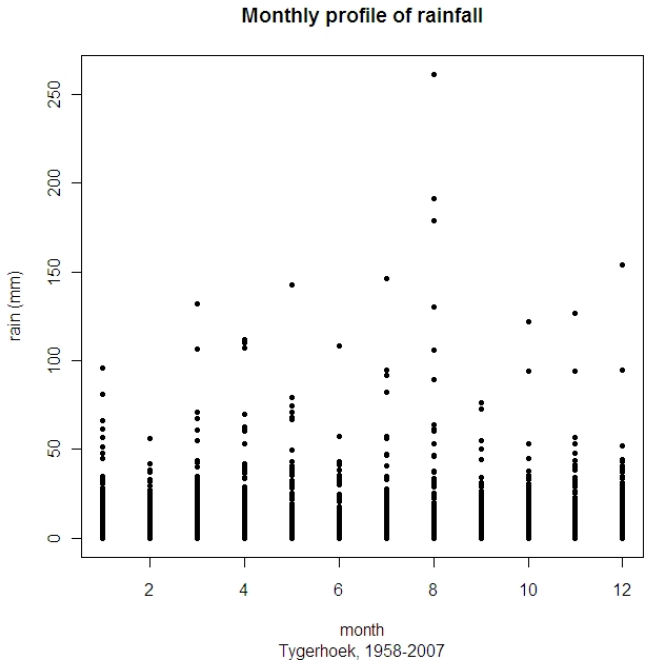


Figure 4.3: Daily rainfall values (winter): Molteno – Wellington



(a)



(b)

Figure 4.4: Comparison of monthly profiles of daily rainfall values from predominantly winter and all-season rainfall regions of the Western Cape

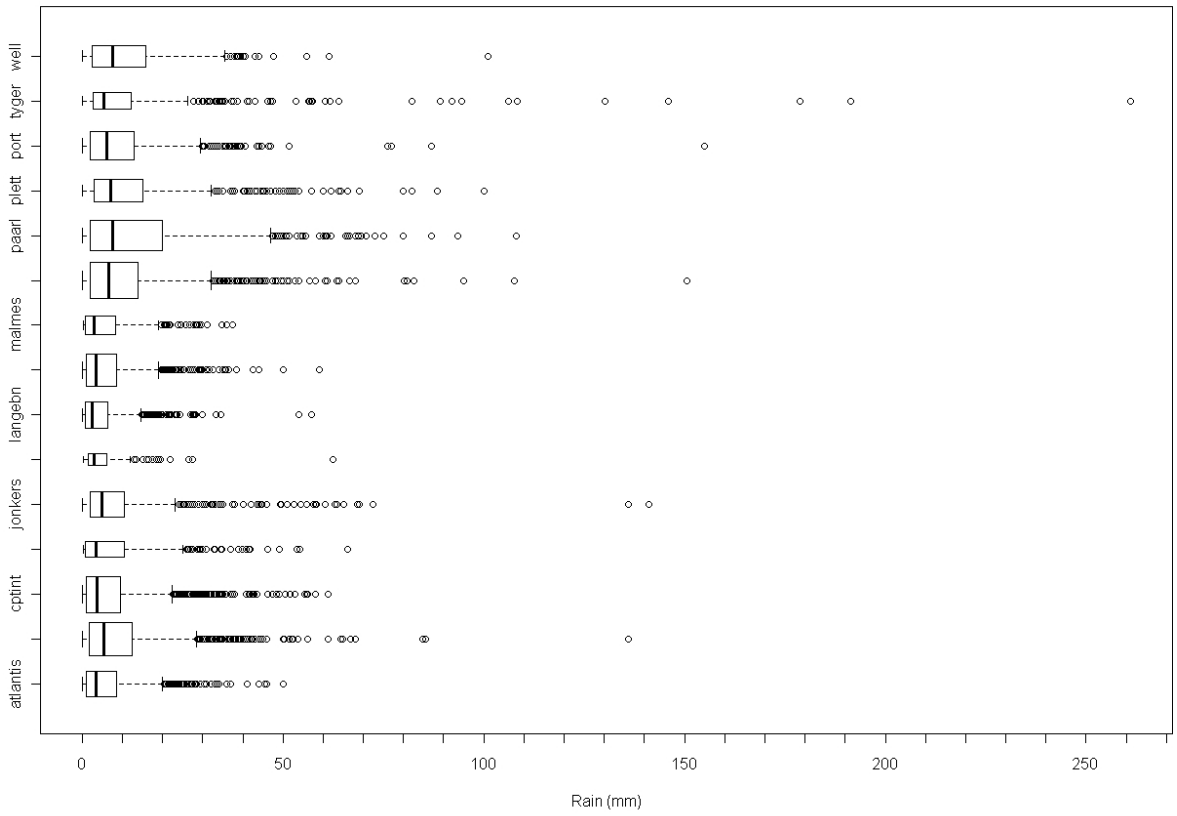


Figure 4.5: Descriptive plots for the series at each station

4.2 Site-wise Threshold Selection

The first step in fitting the point process model is the selection of a suitable threshold (u), as discussed in Section 3.5.1. For practical reasons, the sensitivity analysis for two stations is discussed in detail in this section, with graphical output for the remaining stations attached in the appendix. The two stations, Cape Town International airport and Tygerhoek, are chosen because of their location and the completeness of the data (refer to Table 4.1).

Table 4.2: Sensitivity analysis: Threshold ranges and initial values for the numerical optimization of the log-likelihood function for all sites

Site	Mean (Rainy days)	90^{th} Percentile	Threshold Range	Initial Values (Optimization)
Atlantis	6.40	8.05	8 - 32	(6, 11, -0.05)
Cape Town Astro.	9.12	10.6	11 - 54	(9, 10, -0.05)
Cape Town Int.	6.73	9.71	10 - 43	(7, 11, -0.05)
Excelsior	7.94	6.46	6 - 41	(8, 11, -0.05)
Jonkersberg	8.78	5.00	5 - 58	(9, 6, 0.05)
Ladismith	4.84	0.00	0 - 19	(5, 4, 0.05)
Langebaanweg	4.61	5.20	5 - 28	(5, 5, 0.05)
Langgewens	6.02	7.20	7 - 35	(6, 8, 0.005)
Malmesbury	5.79	7.20	7 - 21	(6, 11, -0.05)
Molteno	10.14	14.00	14 - 63	(10, 10, 0.05)
Paarl	13.33	18.00	18 - 72	(13, 20, -0.05)
Plettenbergbaai	11.78	5.00	5 - 59	(12, 15, -0.05)
Porteville	9.35	9.00	9 - 44	(9, 10, 0.05)
Tygerhoek	13.10	0.00	0 - 88	(13, 20, 0.05)
Wellington	10.59	14.85	15 - 45	(11, 10, 0.05)

Diagnostics that are often used for threshold selection are the mean residual life plot and the threshold stability plot (Coles, 2001). The same convention is followed in this study, but to reduce the uncertainty regarding the choice of threshold an additional sensitivity study is done. For this threshold ranges given in Table 4.2 are created for each site, such that the lower bound of the range corresponds to the 90^{th} percentile of the rainfall series at that site. The upper bounds are chosen close to the 99.8^{th} percentile. These ranges are also used to create the threshold stability plots. To assist in the convergence of the 'Nelder-Mead' optimization routine for the rest of the threshold diagnostics, starting values as given in Table 4.2 were chosen as follows:

- the mean of the non-zero rainfall series is used as the starting value for the location parameter

- the value of the scale parameter at lower thresholds is determined from the threshold stability plot and used as an initial value for the scale parameter
- the direction of the slope of the mean excess is determined from the MRL plot and depending on whether it is decreasing, increasing or constant. The starting values for the shape parameter are -0.05, 0.05 and 0.005, respectively.

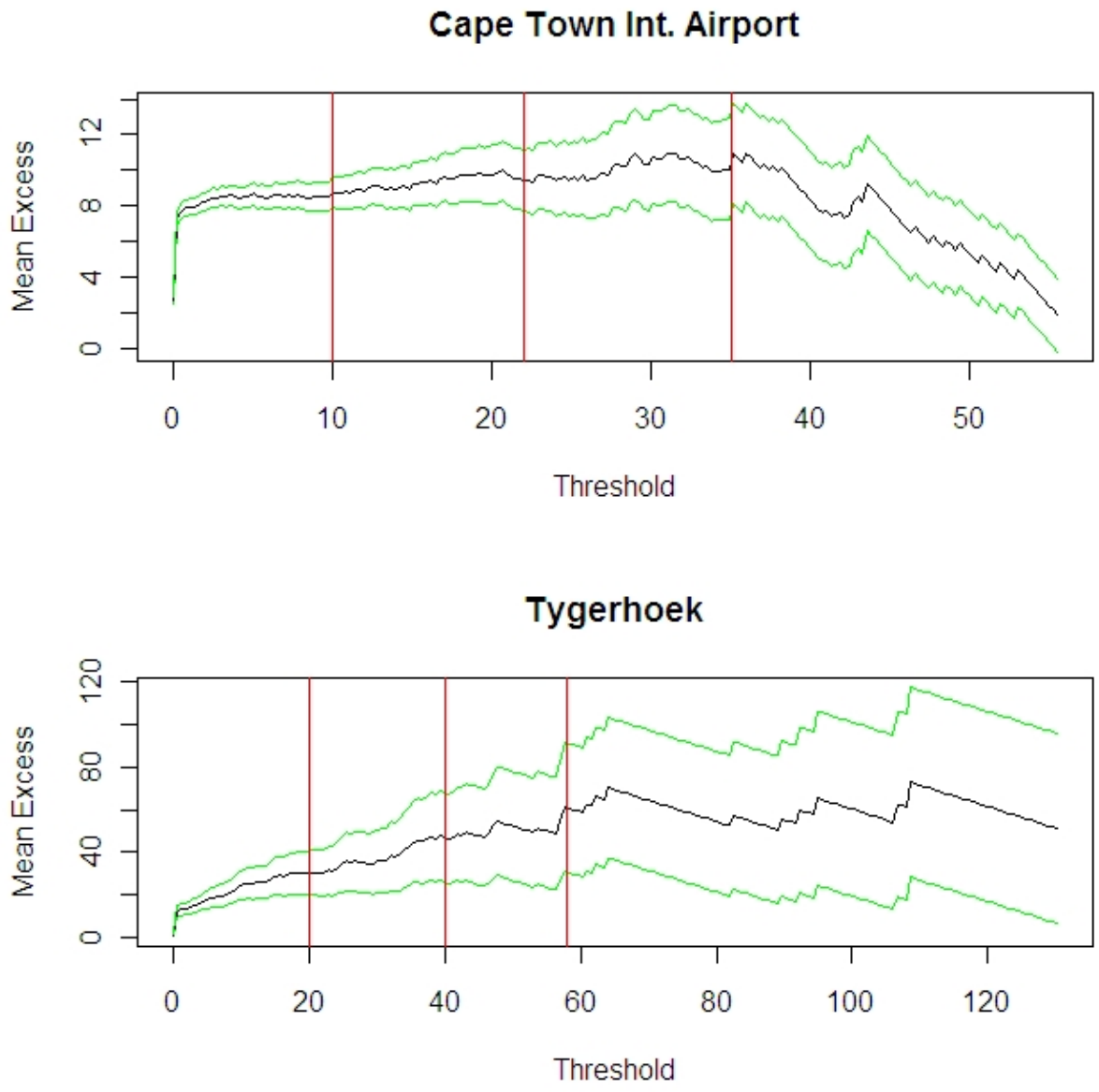


Figure 4.6: Threshold selection – mean residual life plots for Cape Town Int. and Tygerhoek

Firstly the search for linearity in the mean residual life (MRL) plot and stability in the plot

of the threshold against parameter estimates is discussed. From Fig 4.6, the MRL plot for Cape Town International (CPT Int.) is non-linear. A decreasing linear trend, indicative of a non-positive shape parameter is observed for values beyond 35 mm. This seems to be an appropriate value with respect to the assumptions regarding the asymptotic distribution, however such a high threshold may result in few exceedances, hence reduced precisions about the parameter estimates. Therefore lower values 10 mm and 22 mm are also considered. The MRL plot for Tygerhoek is positively linear until $u = 60$ mm, thereafter it appears constant. While this value can be considered a suitable threshold in terms of the model unbiasedness, the wide confidence bands indicate that there are few exceedances of this value. Therefore, candidate thresholds are also considered at 20 mm and 40 mm. In each case there is more than one potential threshold identified from the MRL plot. This diagnostic alone does not provide enough evidence to justify the choice of one value over another. Therefore, further diagnostics are considered.

In the point process limit to extreme values, the distribution of the excesses of a suitable threshold can be approximated by the GPD. Therefore, we anticipate that the parameters of the GPD will be stable for suitable threshold values. For the rainfall series at CPT Int. in Fig 4.7(a), the shape parameter is stable until threshold value of 20 mm. The decreasing trend thereafter indicates that convergence may not have been reached. It stabilizes at 35 mm indicating that it is plausible to assume convergence to the GPD from that point beyond. The concern here, as stated in the MRL plot, is that exceedances of 35 mm may be too few. The plot for Tygerhoek (Fig 4.7(b)) shows that the shape parameter is more unstable than the scale parameter. The confidence bands are high from very low threshold values indicating that the degree of uncertainty regarding the shape of the tail of the rainfall distribution for this station is high. This parameter can be considered stable for threshold values up to 20 mm, followed by a slight decrease towards negative shape parameter, stabilizing after $u = 60$ mm. This concurs with the information obtained from the MRL plot.

Additional diagnostics for threshold selection in this study are based on the plausibility of the Poisson process assumption for the point process approach to extreme value modelling and similar to the MRL plot, the relatioship between the sample mean excess and the threshold. The goodness of the model fit for the different threshold values is also used as a diagnostic. From the sensitivity plots for CPT Int.(Fig 4.8), the number of exceedances decreases exponentially as the threshold increases. A similar trend is also observed for the count of exceedances at other stations. There are less than fifty values above 35 mm, but for a threshold values below 22 mm, more than one hundred exceedances are observed. An exponential decrease in the number of exceedances with increasing threshold is also observed for Tygerhoek, but in this case it is more rapid. The number of exceedances drop below one hundred before the threshold 20 mm is reached and there are very few values above 60 mm.

In the point process approach, the assumption is that the distribution of the number of exceedances taken over a pre-specified block is approximately Poisson. In this case the

annual winter blocks are considered. For a suitable threshold the Poisson assumption should be valid. A characteristic of the Poisson distribution is the equality of the mean and the variance, hence to test whether the assumption is violated, the dispersion index is tested for statistically significant deviation from unity (refer to Section 3.5.1). The dispersion index plot shows the unit line in red, while the dotted lines are the bounds of the 90% confidence interval and the dashed lines for the 95% confidence interval. For CPT Int. there is significant over-dispersion for threshold values below 20 mm and the index is close to one between 24 and 26 mm and at 30 mm. For Tygerhoek there is significant over-dispersion for threshold values higher than 20 mm. Theoretically, the relationship between the mean excess and the threshold is linear, hence in plotting the sample mean excess against the threshold, the search is for the point from which a linear trend is observed. This plot can be thought of as zooming-in to our range of interest within the MRL plot. If values beyond 35 mm are ignored for CPT Int., a positive linear trend starting from 22 mm is observed. If the higher values are not ignored, then a downward linear trend is observed from the threshold value 26 mm. For Tygerhoek, a positive linear trend from 40 mm is observed.

In the plot of the root mean square error against the threshold, the objective is to look at values where the RMSE is minimum. This diagnostic measures the closeness of the predicted mean excess values from the estimated linear mean excess function to the observed mean excess values. This is minimum for 22 mm and 26 mm for CPT Int. and 42 mm for Tygerhoek. The shape parameter is appears stable until 26 mm for CPT Int. and between 40 mm and 60 mm for Tygerhoek. The last diagnostic is based on the model deviance statistic. This is a measure of the goodness of fit of the model with the specified parameters as compared to a saturated model. The deviance is scaled by the number of exceedances, because it tends to decrease as the sample size gets smaller. Lack of change in the deviance statistic as higher thresholds are considered indicates that little information is gained by fitting the extreme value model to exceedances of the higher threshold. Therefore in plotting the scaled deviance statistic against the threshold, the objective is look for thresholds for which the change in the scaled deviance is minimal. For CPT Int., the slope is gentle between 20 mm and 22 mm, and from 32 mm and beyond, while for Tygerhoek the slope is gently around 20 mm and 40 mm and is constant after 60 mm.

Lastly, the sensitivity of the return level estimates (Equation 3.5.6) to the threshold is examined, specifically the 5-, 25-, 50-, and 100- year return levels. This is the key output of the extreme value analysis, hence it is important to ensure that a reliable estimate of the return level is obtained. If u_0 is an appropriate threshold, then the return level estimate should be stable for threshold higher than this value as a result of stability in the model parameter estimates. From Fig 4.8 and 4.11 stability is reached beyond 26 mm for CPT Int. and from about 38 mm for Tygerhoek.

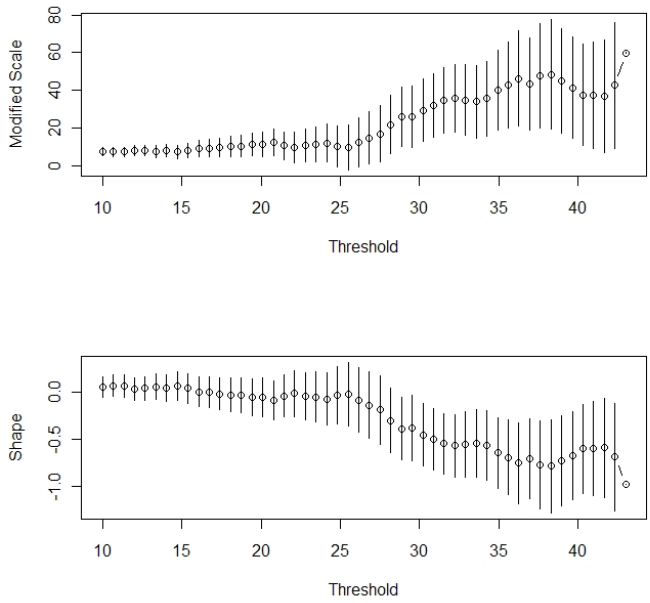
The high variability observed in the threshold diagnostic plots discussed above may be the result of short-range correlations present in the series at each site, which has not been accounted for. Hence, as a final diagnostic, changes of the extremal index as a measure of

the degree of clustering at high levels of the series, with threshold are explored. The aim is to identify threshold values where the degree of clustering is minimal. The presence of short-range temporal dependence violates the key assumption of extreme value models – the independence of the observations. Fig 4.12 and 4.13 show extremal index plots for all the sites, but as previously, the discussion is limited to CPT Int. and Tygerhoek. Weak levels of clustering occur for extremal index values beyond 0.9, hence exceedances of thresholds with corresponding indices which are at least 0.9 are assumed independent.

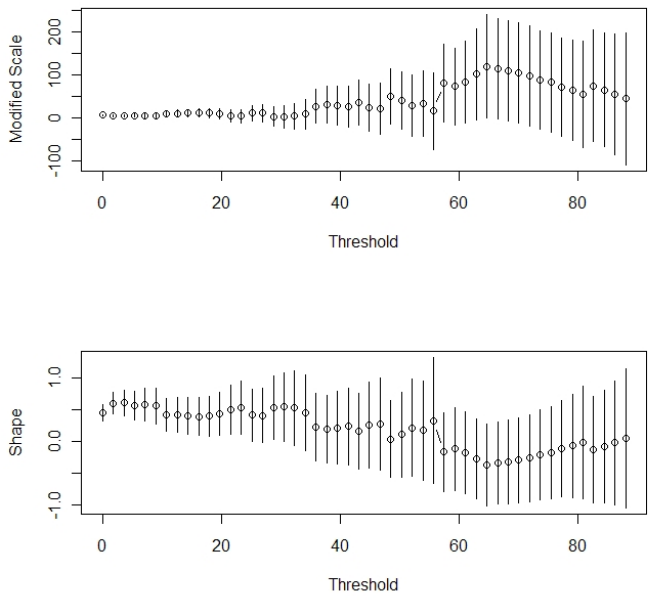
The extremal index plot against the threshold, reveals very weak levels of temporal dependence for thresholds higher than 25 mm for CPT Int.. Weak levels of dependence have little effect on the parameter estimates, hence fitting the model declustering will only be considered for those stations where the extremal indices for the considered thresholds are far below 0.9. In the case of Tygerhoek, the extemal index is stable below 0.80 for threshold higher than 10 mm. For this station declustering has to be considered.

The above threshold sensitivity study reveals that a suitable threshold for CPT Int. is 26 mm. For Tygerhoek a threshold close to 40 mm seems appropriate, but considering the presence of temporal dependence, a lower threshold whose exceedances will be declustered is also chosen. The parameter estimates are stable from these values. One could consider higher thresholds, but the number of exceedances drops exponentially when higher thresholds are considered. This is a problem because while it is important for bias to be kept minimal, it is also important to maintain the degree of precision in the parameter estimates at acceptable levels. From the diagnostics, threshold values around 20 mm appeared plausible, hence for comparison thresholds chosen for CPT Int. are 22 mm and 26 mm. It was a similar situation for Tygerhoek, hence the two thresholds for this station are 18 mm and 38 mm.

In the next section the point process extreme value model is fitted to two thresholds at each site. Further, declustering will be applied to rainfall data from those stations where temporal dependence needs to be accounted for. The result of this exercise will be a single model for each site from which the return level estimates will be obtained. Threshold diagnostic plots for the other stations can be found in the appendix A.1.



(a) GPD parameter stability against changing threshold - CPT Int.



(b) GPD parameter stability against changing threshold - Tygerhoek

Figure 4.7: Analyzing the stability of the scale and shape parameters to changes in the threshold value

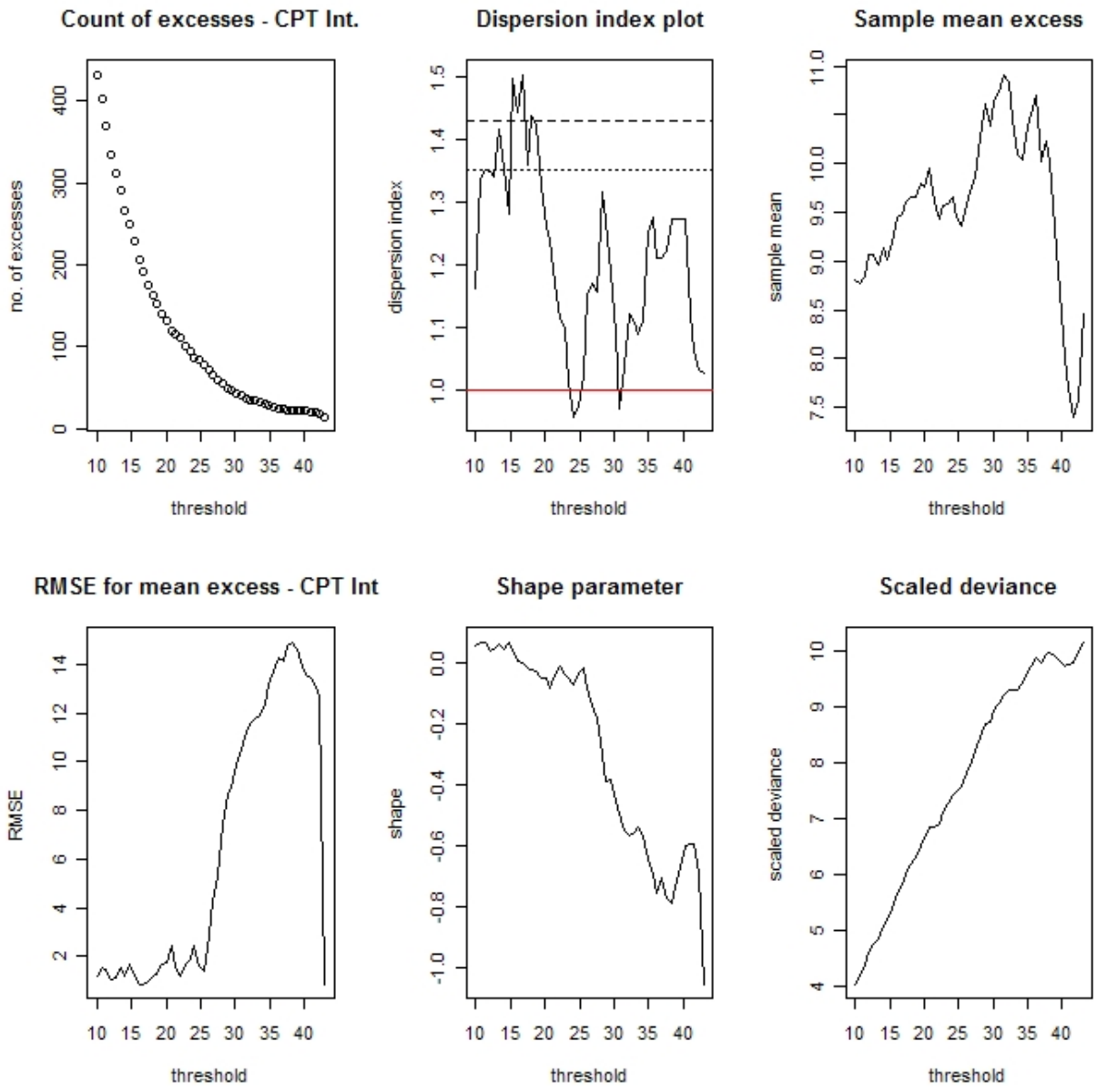


Figure 4.8: Sensitivity of the point process extreme value model characteristics to the threshold – CPT Int..

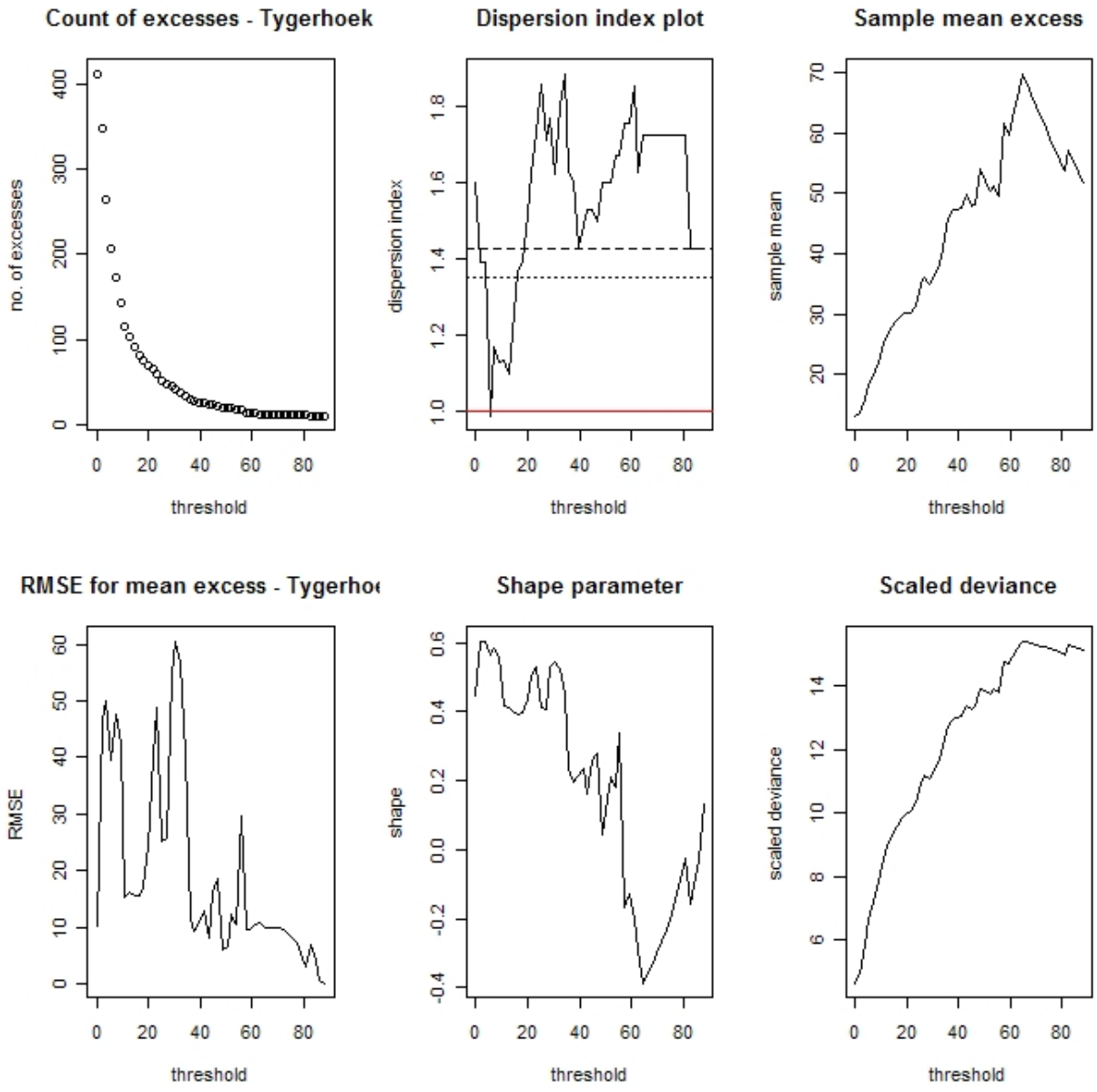


Figure 4.9: Sensitivity of the point process extreme value model characteristics to the threshold for Tygerhoek

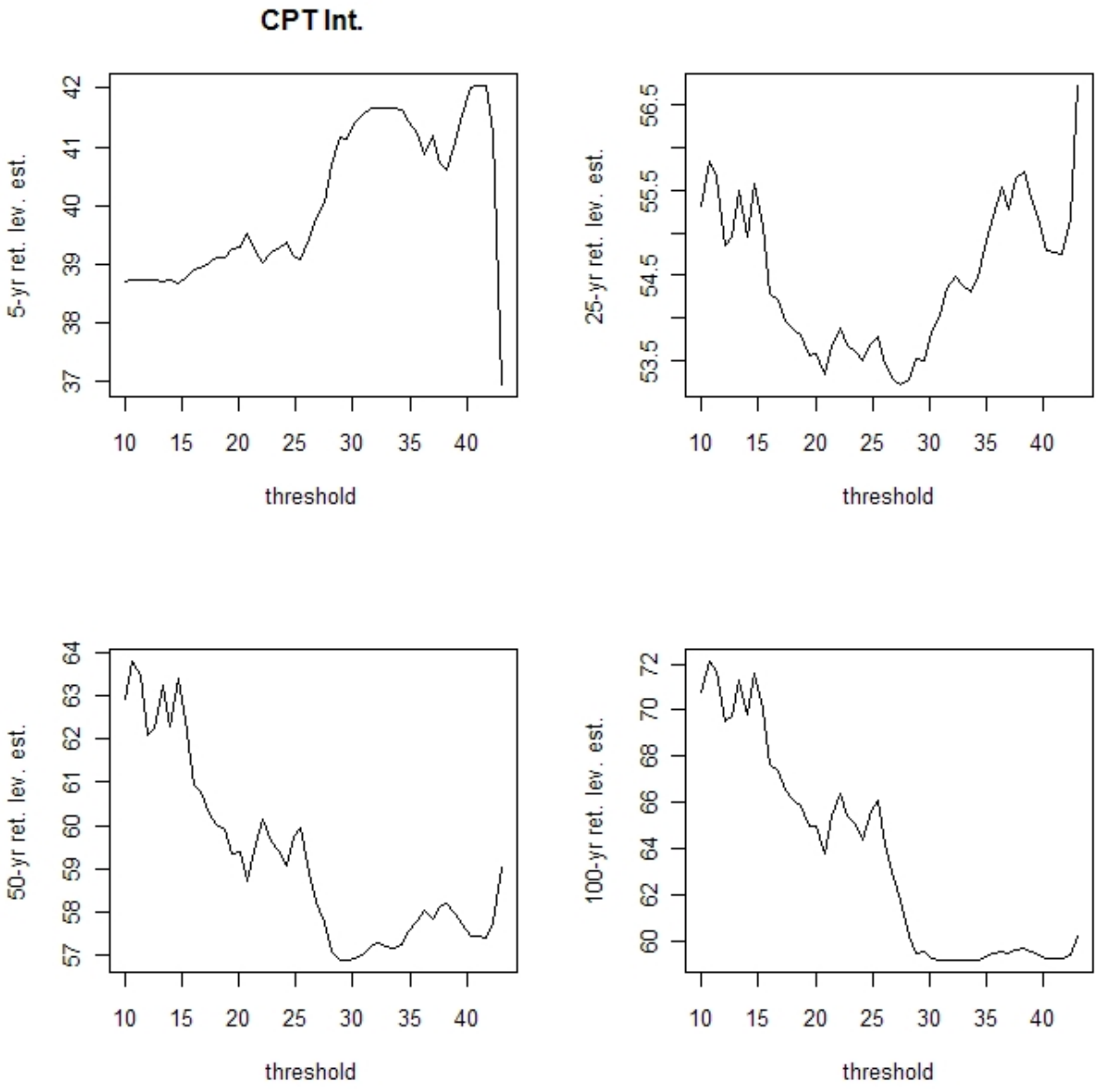


Figure 4.10: Sensitivity of the return level estimates to threshold level for CPT Int.

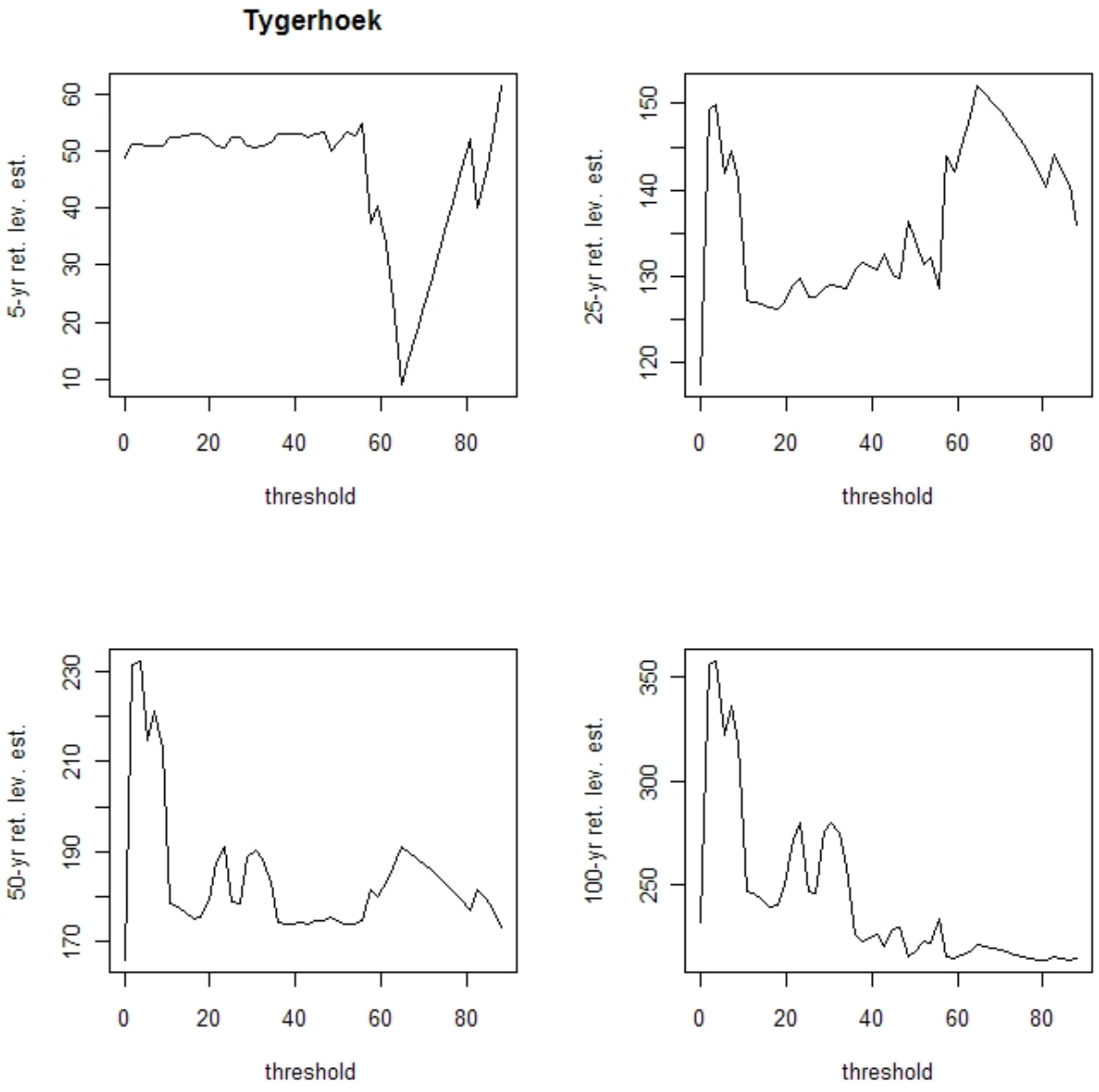


Figure 4.11: Sensitivity of the return level estimates to threshold level for Tygerhoek

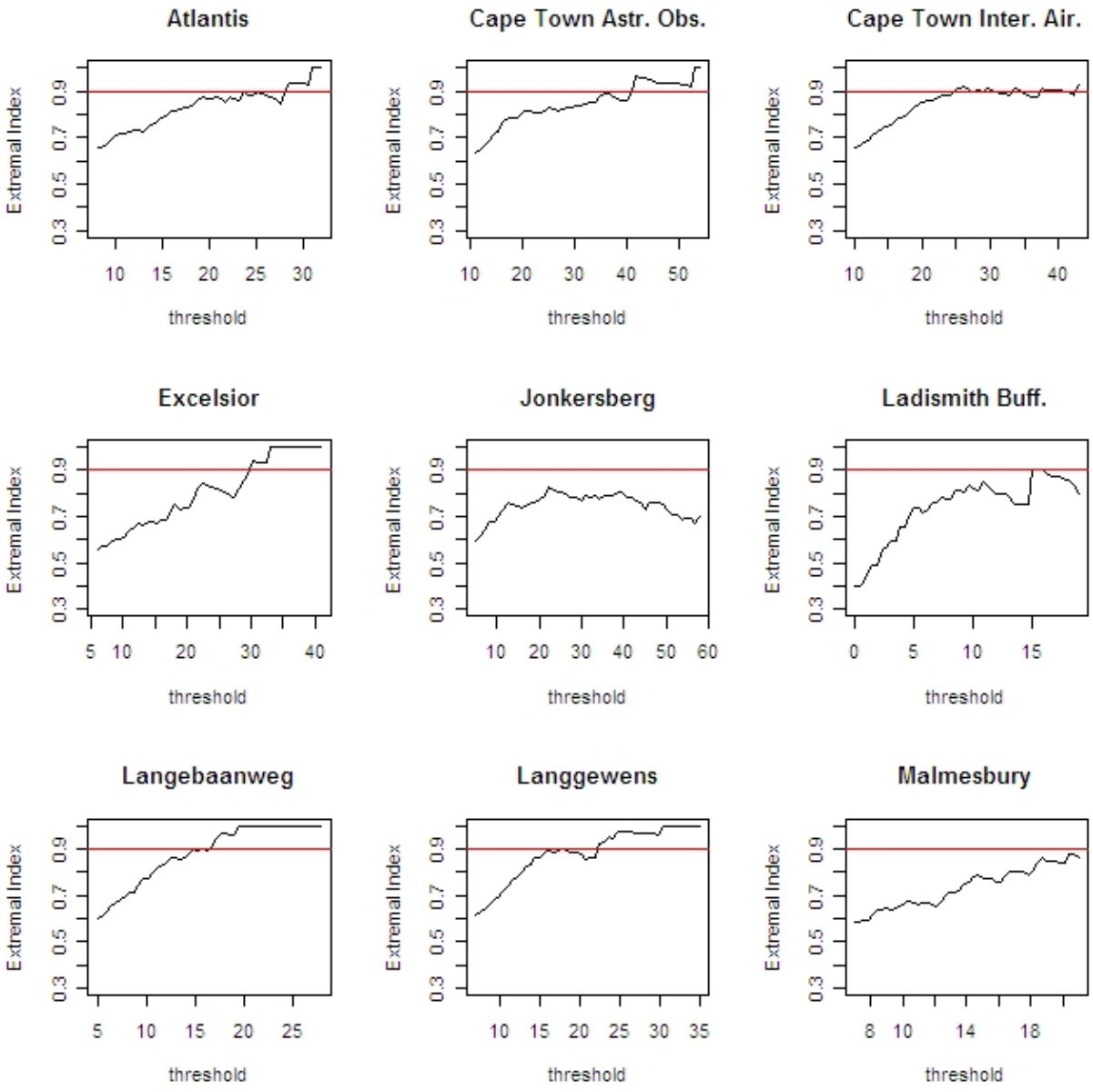


Figure 4.12: Level of clustering of exceedances at each threshold level for the stations Atlantis to Malmesbury

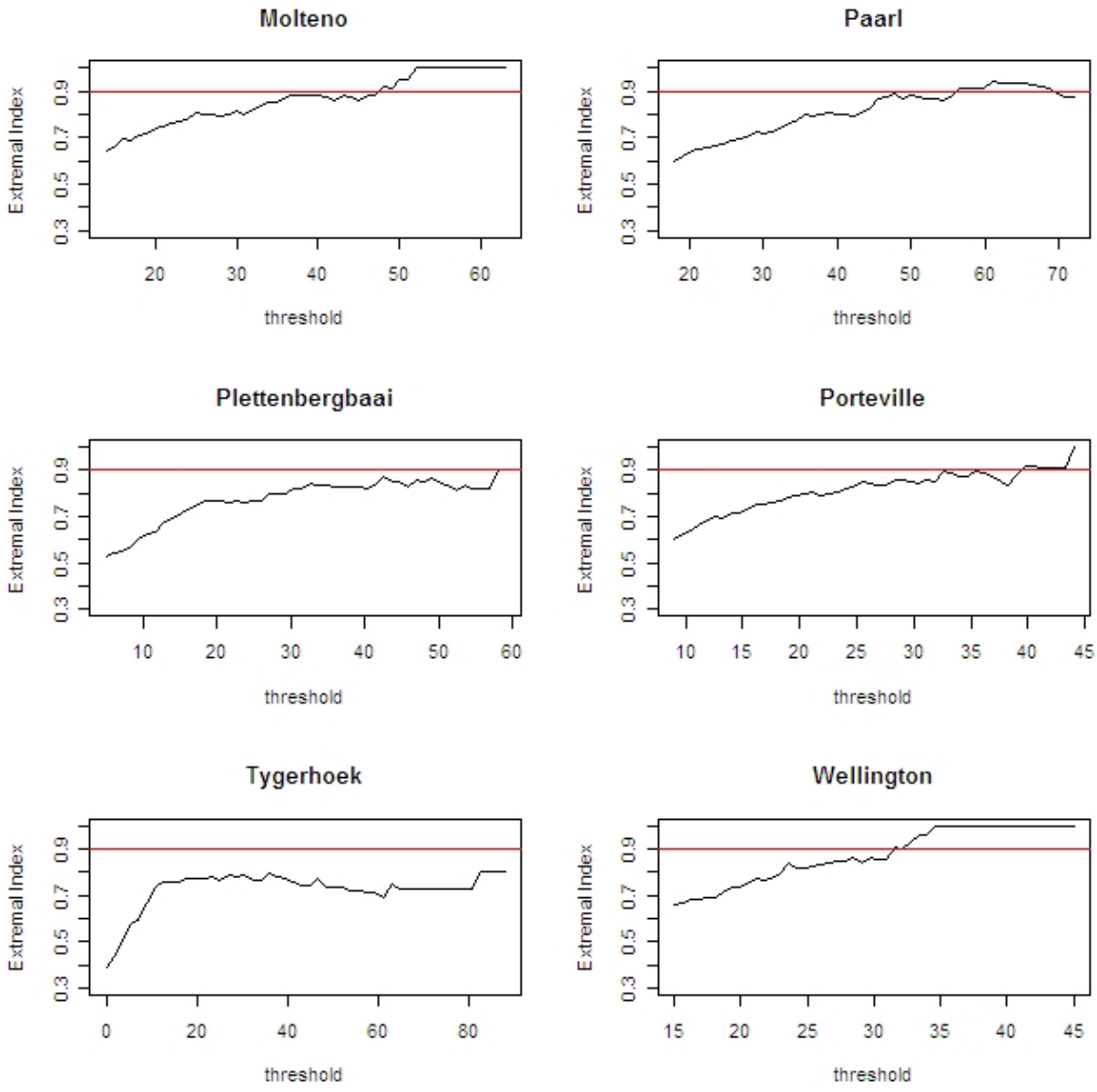


Figure 4.13: Level of clustering of exceedances at each threshold level for the stations Molteno to Wellington

4.3 Site-wise Point Process Extreme Value Models

In the previous section, extensive sensitivity analysis was undertaken to identify potential thresholds. Two thresholds were considered and subject to further analysis for each station. For CPT Int. these are 22 mm and 26 mm and for Tygerhoek these are 18 mm and 38 mm. In this section the point process extreme value model is fitted to threshold exceedances. For some stations, like Tygerhoek, temporal dependence is accounted for, while for others it can be assumed that the exceedances are temporally independent, due to the high extremal index values obtained for those stations.

Generally the sample proportion of exceedances is small. For example, from the CPT Int. station only 111 of the 4600 winter days had rainfall in excess of 22 mm. This amounts to only 2.4% of the observed values from this station (refer to Table 4.3). In other words, over the period 1958 to 2007, there were on average only 2 days every winter[3] where rainfall accumulated over 24 hours was at least 22 mm. When a threshold 26 mm is considered, this drops to a single day[4] per winter. The proportion of exceedances is also low at Tygerhoek at only 1.6% for the threshold 18 mm and 0.6% for 38 mm.

The shape parameter estimate is negative for both thresholds for CPT Int. and positive in the case of Tygerhoek. This is in agreement with MRL plots in Fig 4.6, where a linear decreasing trend was observed for CPT Int. and an increasing trend for Tygerhoek. To investigate the plausibility of an asymptotically zero shape parameter, a point process model with the shape parameter set to zero was fitted and the deviance statistic was calculated. There is insufficient evidence in the data for CPT Int., to dispute the hypothesis of an asymptotically zero shape parameter for both thresholds. On the other hand there is strong evidence in support of a long-tailed maximum distribution for Tygerhoek when the lower threshold 18 mm is considered, but evidence is lacking to reject medium-tailed distribution when a higher threshold 38 mm is considered (see Table A.1).

Table 4.3: Results of the point process model fit for CPT Int. and Tygerhoek.

Site	u	Dev. (Model)	n_u/n	$\hat{\theta}$	$\hat{\sigma}_*$ (s.e)	$\hat{\xi}$ (s.e)
CPT Int..	22	766.16	0.024		9.67 (1.50)	-0.02 (0.12)
	26	557.6	0.015		11.34 (2.14)	-0.19 (0.13)
Tygerhoek	18	731.89	0.016		19.20 (3.69)	0.38 (0.16)
	18	640.84	0.014	0.94	17.94 (3.75)	0.43 (0.18)
	38	348.59	0.006		37.62 (12.72)	0.21 (0.28)

[3]Exceedance rate per year $= \frac{n_u}{n} \times$ number of winter days per year $= \frac{n_u}{n} \times 92$

[4]Note that these values result from rounding off to the nearest integer. For the threshold of 22 mm, the exceedance rate per year is 2.208 which is approximately 2 days when rounded off.

The declustering of exceedances of 18 mm was done by first using the method of Ferro and Segers (2003), described in Section 3.5. In this method runs length is not an input, but is optimally chosen when the extremal index is estimated from a model of inter-exceedance times. For all stations where declustering was performed, this led to the result that excesses separated by a single day of rainfall that is below the threshold can be assumed independent. For Tygerhoek, the number of days with rainfall above 18 mm was reduced from 74 to 63 when the series was declustered. Table 4.3 shows the results for CPT Int. and Tygerhoek, with the columns corresponding to the threshold u, the deviance (Dev. (Model)) as a measure of the goodness of fit, the sample proportion of exceedances n_u/n, the estimate of the extremal index $\hat{\theta}$, the scale parameter estimate corresponding to the GPD $\hat{\sigma}_*$ and the shape parameter estimate $\hat{\xi}$. The row where the threshold appears in bold, corresponds to model results for the declustered series. For Tygerhoek, the parameter estimates change slightly when the excesses are declustered and the loss in precision is small. Generally increasing the threshold led to a loss in precision in both stations, but for Tygerhoek this is very large at nearly 200% for the scale parameter and 75% for the shape parameter.

The diagnostic plots (in Fig 4.14 the first column are the Q-Q plots and the second column are return level plots for CPT Int., whereas for Tygerhoek, Fig 4.15, the first row and second rows consists of the Q-Q plots and the return level plots, respectively) are used to assess the goodness of model fit. There is an improvement in model fit when the higher threshold 26 mm is considered for CPT Int. Additionally, the lower threshold for CPT Int. results in return level estimates that are too optimistic as can be seen from Table 4.4. Therefore, the model fitted to threshold exceedances of 26 mm is chosen as the final model for CPT Int. The similarity in the diagnostic plots for the model fitted to the declustered and original exceedance series concurs with what was deduced earlier, that very small changes in the parameter estimates are observed when the series of exceedances is declustered. Further, an improvement in fit is observed when the threshold 38 mm is considered, however this is not chosen as the final threshold as the number of exceedances are few leading to high imprecision in the parameter estimates. Generally for this station, the return level estimates and the corresponding confidence intervals are large (refer to Table 4.4), which can be attributed to less than 80 exceedances covering a wide range between 18 mm and 262 mm.

Table 4.4: Return level estimates – CPT Int. and Tygerhoek, with 95% profile confidence intervals

Site	Threshold	$\hat{x}_{25}$ (95% Prof. c.i.)	$\hat{x}_{100}$ (95% Prof. c.i.)	$\hat{x}_{500}$ (95% Prof. c.i.)
CPT Int.	22	53.83 (48.60, 60.47)	66.16 (58.84, 75.35)	80.05 (70.58, 92.09)
	26	51.55 (48.13, 58.13)	59.33 (55.00, 70.94)	66.09 (61.25, 78.13)
Tygerhoek	18	125.93 (94.44, 163.89)	236.40 (177.78, 319.44)	464.64 (352.94, 632.35)
	18	119.16 (88.89, 161.11)	235.68 (172.22, 327.78)	494.66 (355.56, 688.89)
	38	131.44 (97.22, 188.89)	224.26 (155.56, 344.44)	372.78 (258.33, 580.56)

Similar analysis was carried out for the rest of the stations. More detailed results are given

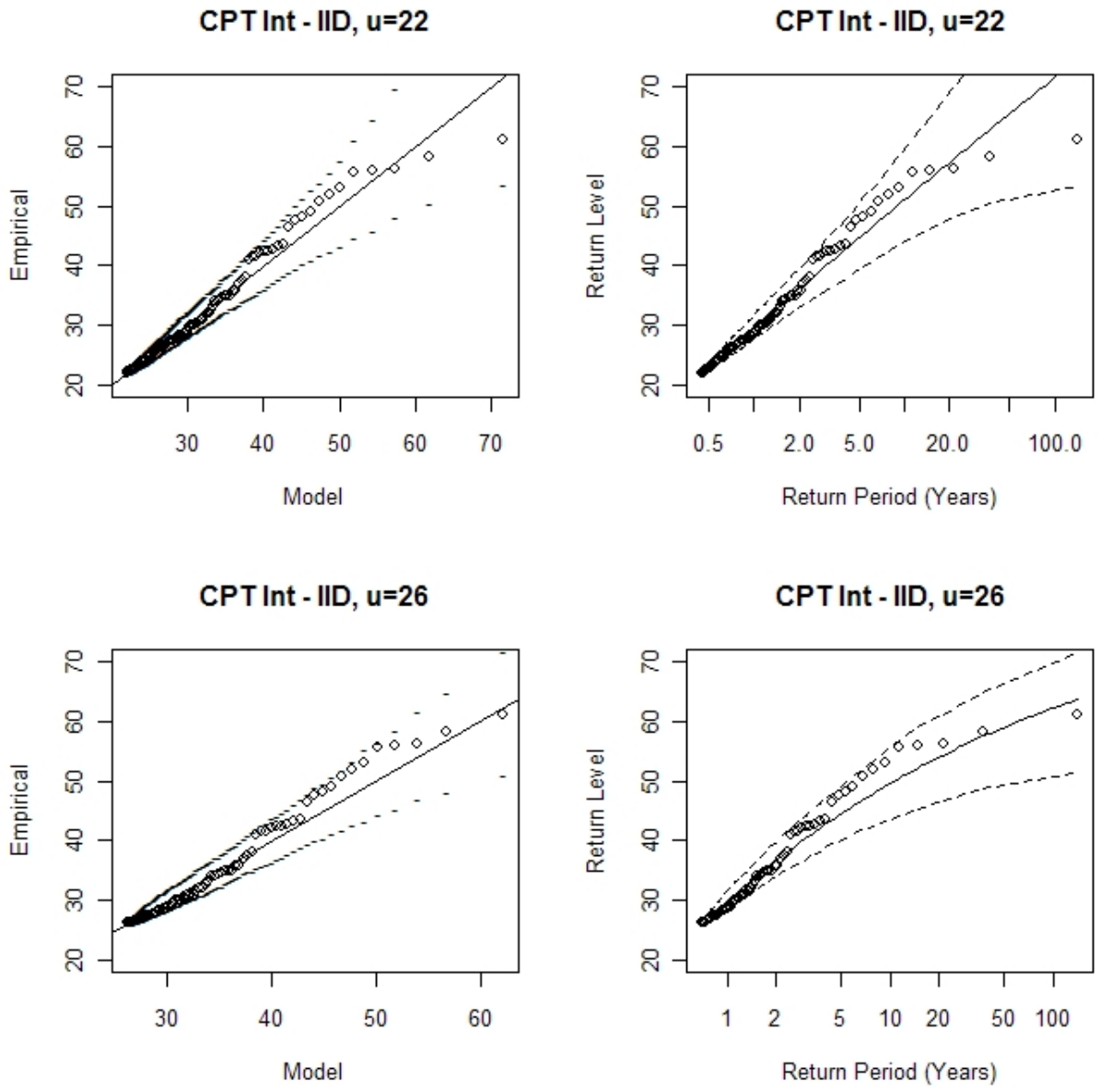


Figure 4.14: Q-Q and return level plots for models fitted to data from CPT Int.

in the appendix, but in Table 4.5, summary of the final point process models chosen for each station are given. Note that for some stations such as Ladismith, the thresholds are very low. Although rainfall accumulated over 24 hours that is below 10 mm may not be considered heavy, the size of the daily rainfall amounts at stations such as Ladismith and even Malmesbury and Langebaan, made it difficult to choose higher threshold from our data-driven threshold selection diagnostics. To emphasize how light the observed rainfall events from these stations are, Fig 4.5 shows that the 75^{th} percentile at these stations is less than

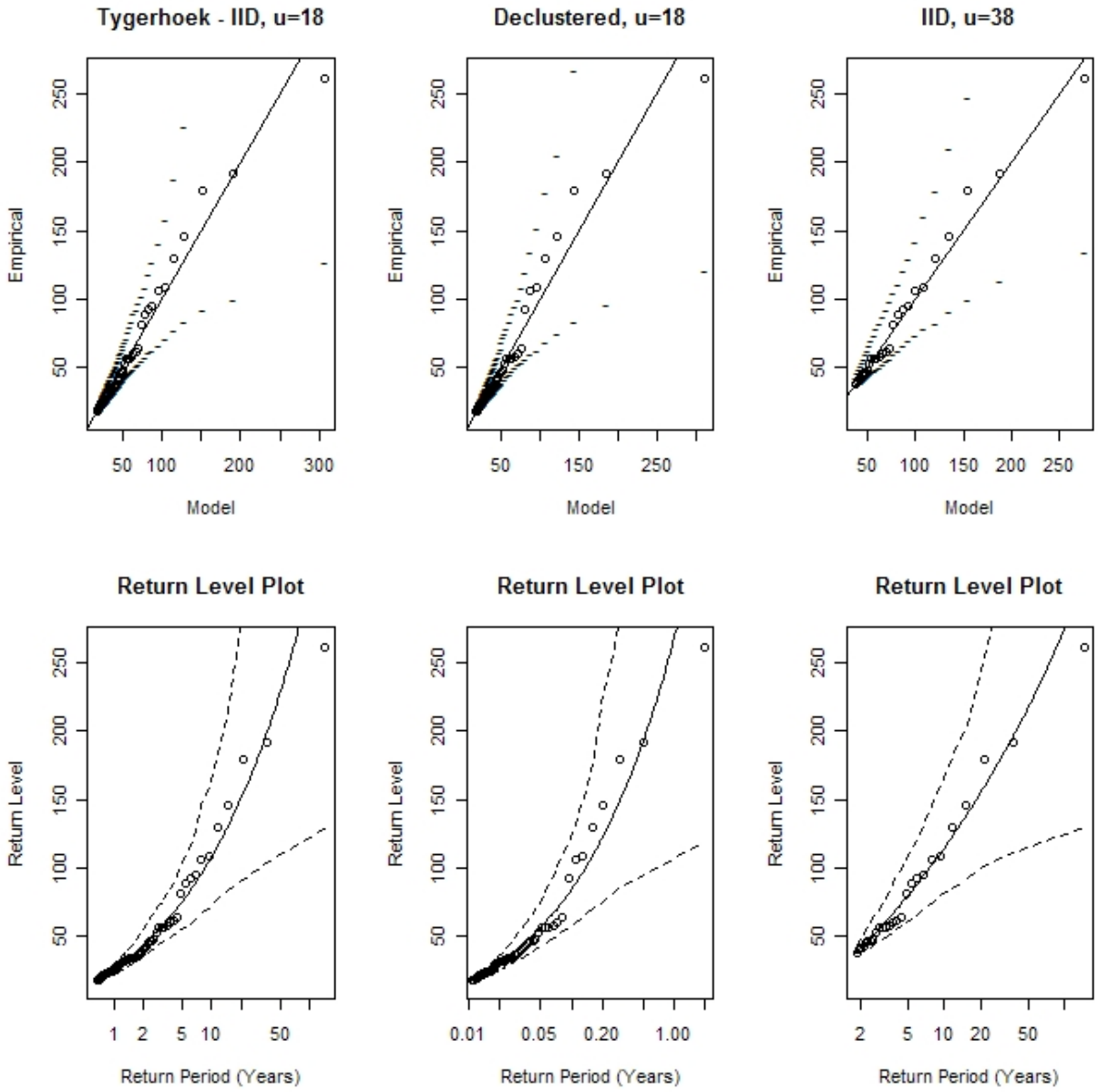


Figure 4.15: Q-Q and return level plots for models fitted to data from Tygerhoek

10 mm. Therefore, the 25 year return level estimates for these stations are also lower for these stations in comparison to the others.

The chosen models appear to have fitted the data well. Fig 4.16 and 4.17 are the Q-Q plots for all the stations. Cases such as for the Jonkersberg and Langebaanweg stations, where the model performed badly in estimating the upper quantiles are few, and even in such cases there discrepancies are within the 95% confidence bounds. High levels of uncertainty,

Table 4.5: Parameter estimates for the extreme value model fitted at each site. Bold values in the column of selected thresholds u indicates series that have been declustered. The maximum likelihood estimates for the location (μ), scale (σ), shape (ξ) and 25 year return level estimates (x_{25}) are accompanied by the standard errors and confidence intervals in brackets

Site	u	$\hat{\mu}$ (s.e)	$\hat{\sigma}$ (s.e)	$\hat{\xi}$ (s.e)	$\hat{x}_{25}$ (95% Prof. c.i.)
Atlantis	18	26.23 (1.24)	7.21 (0.72)	-0.09 (0.11)	39.63 (35.60, 44.80)
CPT Astro	**28**	33.52 (1.50)	11.4 (1.57)	0.19 (0.13)	71.56 (62.50, 85.00)
CPT Int.	26	29.65 (1.44)	10.63 (1.61)	-0.19 (0.13)	51.55 (48.13, 58.13)
Excelsior	**19**	31.43 (3.20)	13.34 (1.93)	-0.22 (0.17)	47.58 (42.50, 56.88)
Jonkersberg	**22**	25.45 (2.17)	16.33 (2.90)	0.17 (0.15)	77.86 (64.84, 100.00)
Ladismith	**5**	9.04 (1.07)	5.55 (0.99)	0.29 (0.17)	22.42 (18.75, 30.00)
Langebaan	**12**	17.54 (1.01)	6.47 (0.80)	0.12 (0.11)	33.99 (29.47, 40.70)
Langgewens	18	22.9 (1.01)	7.64 (0.84)	0.00 (0.12)	42.84 (38.44, 48.75)
Malmesbury	**11**	23.36 (1.49)	6.52 (0.64)	-0.27 (0.12)	30.95 (28.13, 35.25)
Molteno	32	38.81 (1.75)	13.43 (1.77)	0.17 (0.12)	82.80 (71.62, 97.37)
Paarl	**36**	48.65 (2.19)	15.21 (1.52)	-0.10 (0.10)	80.96 (72.86, 91.07)
Plettenberg	40	38.3 (2.04)	13.31 (3.14)	-0.01 (0.10)	72.77 (65.83, 86.33)
Porteville	30	30.3 (1.06)	7.59 (1.57)	0.33 (0.16)	61.52 (52.94, 75.00)
Tygerhoek	**18**	22.36 (2.64)	19.82 (3.89)	0.43 (0.18)	119.16 (88.89, 161.11)
Wellington	**22**	34.5 (2.14)	10.84 (1.42)	0.04 (0.10)	53.45 (46.90, 63.45)

as evident from the widening confidence intervals in Fig 4.16 and 4.17, are observed for stations where the largest observation is nearly twice as large as the second or third largest observation. These observations are very influential, and while it is tempting to remove them from the analysis for this reason, in doing so one risks obtaining return level estimates which are too conservative.

The return level plots also concur with the results of the Q-Q plots, that the models fits are satisfactory. The shape of the return level plots corresponds to the results obtained for the shape parameters as given in Table 4.5. Upward concavity of the return level plot implies an asymptotic maxima distribution belonging to the Frechét family, while downward concavity corresponds to the Negative Weibull family and a straight line to the Gumbel family of distributions. From the return level plots in Fig 4.18 and 4.19, for three of the four stations to the east of the Western Cape province (Jonkersberg, Ladismith and Tygerhoek) the distribution of the maxima can be approximated by the Frechét family of distributions. To the west of the province the shape parameter varies from positive to negative. Two stations, Langgewens and Plettenbergbaai can be considered to have distributions for maxima belonging to the Gumbel family.

The key output in fitting the point process extreme value model was the estimate of the return level. This measure allows us to gauge the rarity of largest events that have been

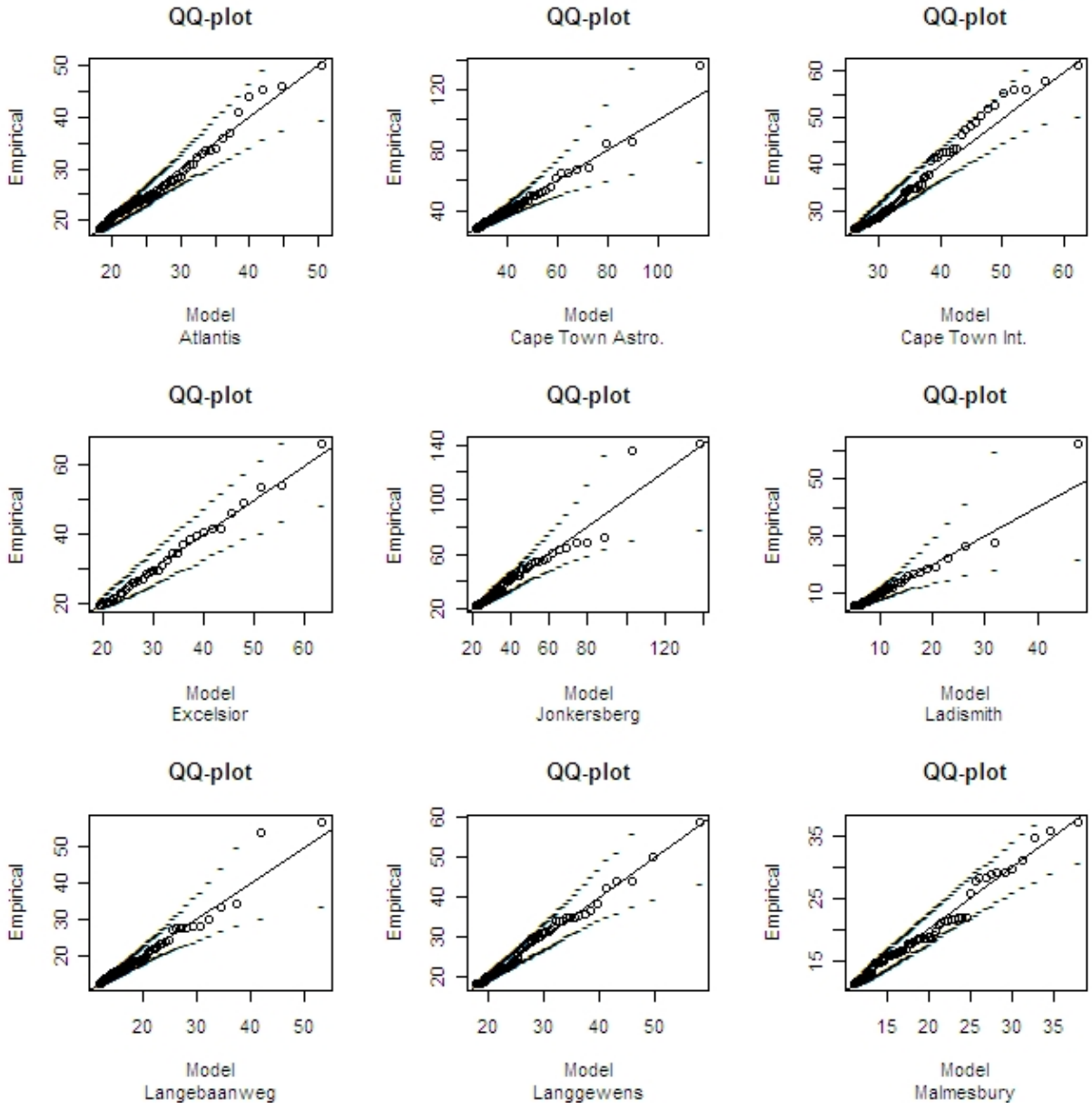


Figure 4.16: Q-Q plots to show goodness-of-fit of the models for the stations Atlantis to Malmesbury

observed for stations in the study area. Using the resulting parameter estimates, it is also possible to quantify the chance of observing rainfall events that are much larger than the ones observed over the study region. From the return level plots (Fig 4.18 and 4.19) and the results in Table 4.5, there is variation in the characteristics of heavy rainfall events over the study area. The concern is whether this observed variation in space is significant. Further, in flood estimation, rainfall return levels at catchment or regional level, rather than at the

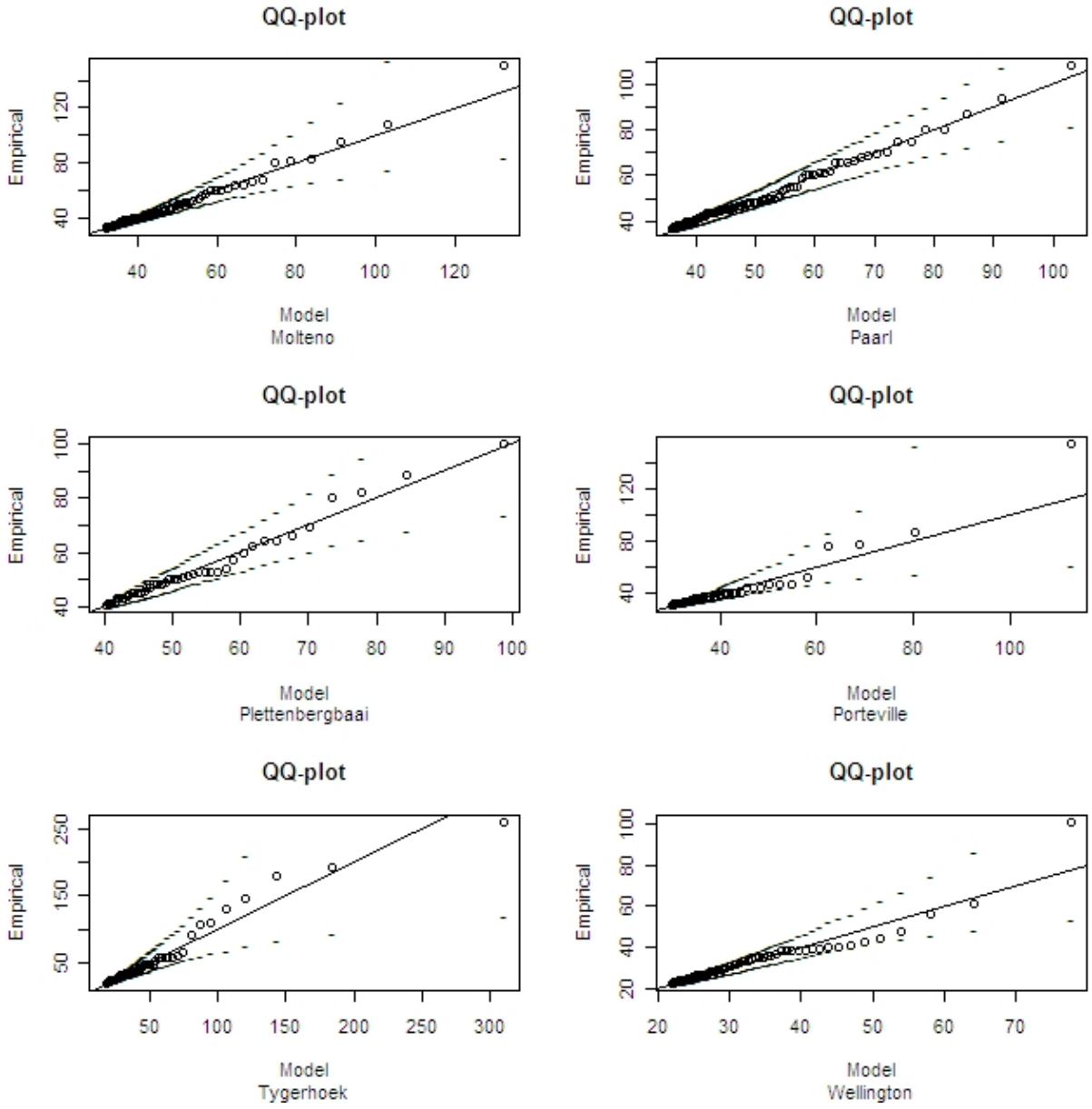


Figure 4.17: Q-Q plots to show goodness-of-fit of the models for the stations Molteno to Wellington

individual gauged site provide more useful information. This necessitates geo-spatial analysis of the rainfall return levels over the study area. Specifically interest is in obtaining the 50 year rainfall return level map of the study area. This is discussed in the next chapter.

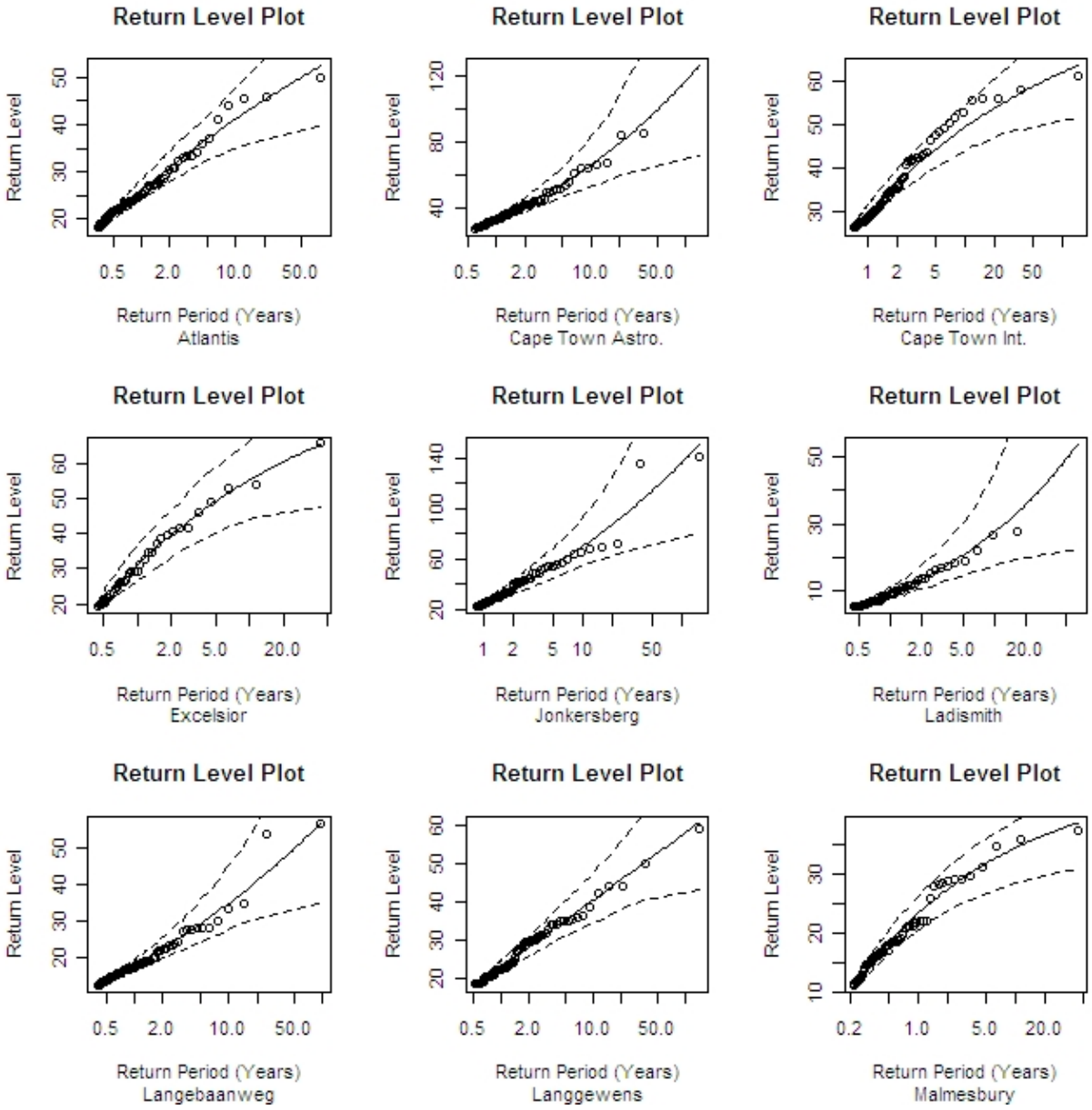


Figure 4.18: Return level plots to show goodness-of-fit of the models for the stations Atlantis to Malmesbury

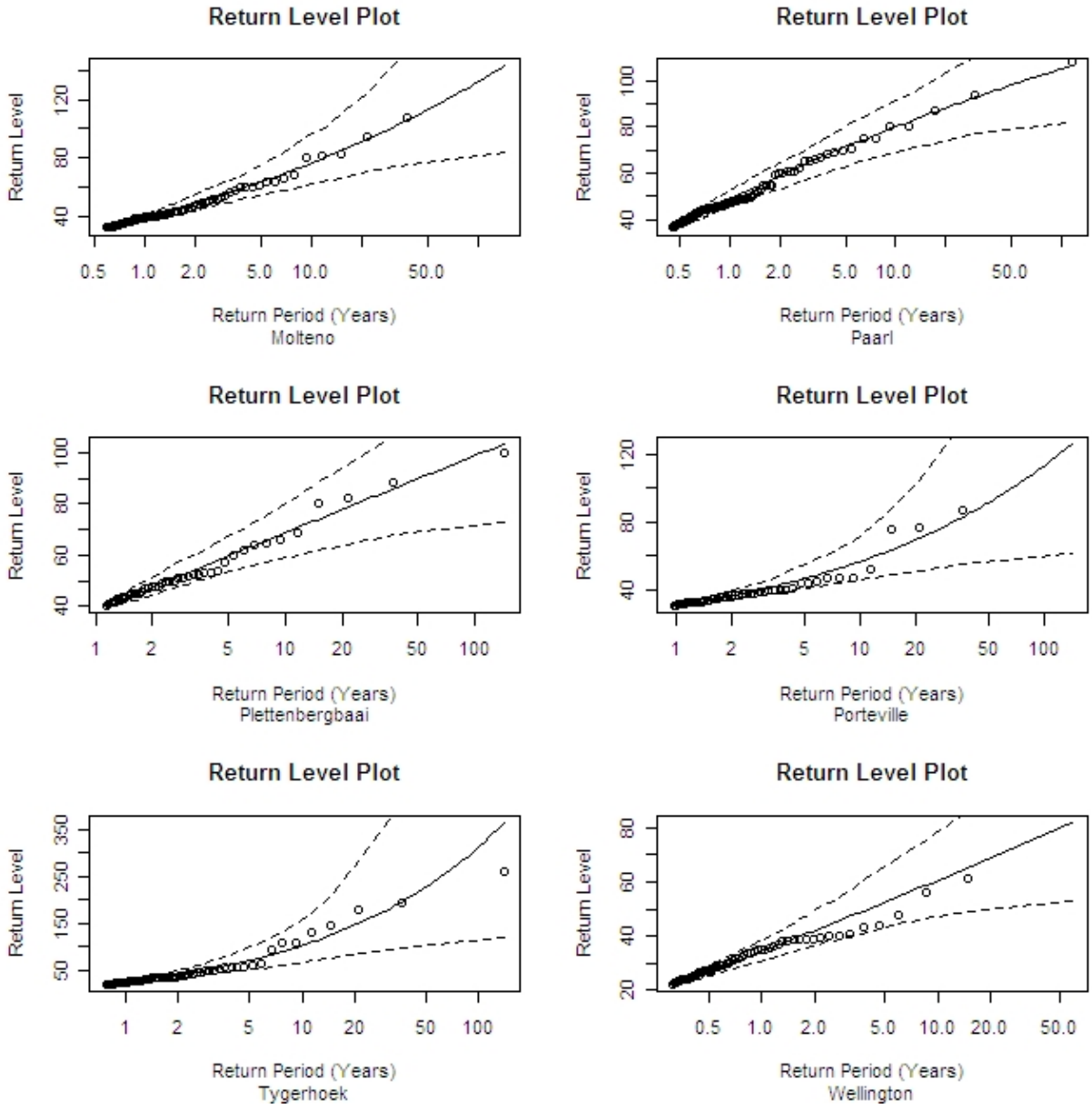


Figure 4.19: Return level plots to show goodness-of-fit of the models for the stations Molteno to Wellington

Chapter 5

Geostatistical Model of Rainfall Return Levels

In the previous section the daily rainfall series at each of the 15 sites was processed through the point process extreme value approach to arrive at the return level estimates. Rainfall return levels (or design values) are important inputs in design flood estimation through rainfall-runoff models. However, for rainfall-runoff models, knowledge of the N-year design value at each gauge site of the catchment is not sufficient, rather interest is in estimating the spatial distribution so that predictions of the runoff can be made for every point within the catchment.

The main objective in this study was to determine whether the estimate 50 year 24-hour rainfall was homogeneous over the study area, which made it necessary to first pre-process the data at each site to arrive at the estimate of the return level. To find out if design values at sites within close proximity show similarity and whether this spatial variation is statistically significant, we begin with estimation of the variogram. The challenge here is the sparseness of the data. With only fifteen design values, the concern is that the sample is too small for obtaining reliable estimates of the parameters of the variogram model. Therefore, the observations were extended in space and in time, increasing the number to used in estimating the variogram (Stein and Sterk, 1999; Sterk et al., 2004). Thereafter, kriging is applied to generate maps for the 50 year 24-hour rainfall return level.

5.1 Exploratory Spatial Analysis of the Return Levels

Firstly, we explore the return level estimates to determine if there are any obvious patterns. In Fig 5.1, the magnitudes of the 25 year return level estimates are compared using the size

of the symbol to identify clusters where they are similar. Large design values are obtained for stations along the coast.

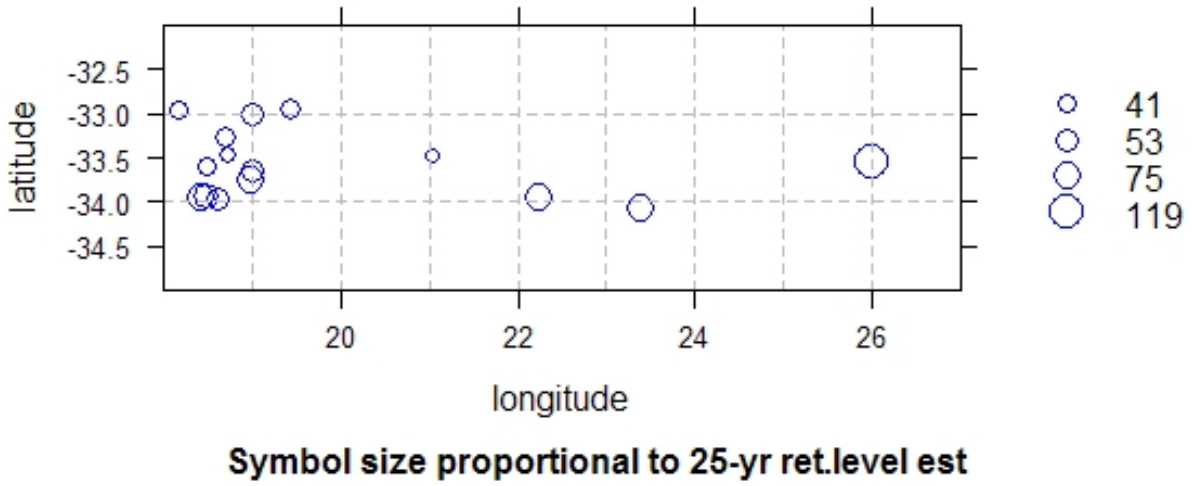


Figure 5.1: The 25 year return level values, where symbol size shows relative magnitudes

It is important to note the irregularity in the sample design, nearly 75% of the points are clustered in one part of the study area – Cape Town and surrounding region. Between the longitudes 19.5 and 21, there are no sample points. This may be due to no installation of weather stations in that area because of a mountain situated in this area.

Initially the relation between the return level estimate and the spatial coordinates is investigated. Changes in longitude seem to have no influence on the 25 year design values (Fig 5.2(a)), however, there seems to be a weak negative relation with latitude as seen in Fig 5.2(b). With only 15 estimated return level values across the study region, it is difficult to be overly confident of the spatial trend in design values. Further investigation is done by regressing the design values to the spatial coordinates. The model reveals that the spatial coordinates do have an overall influence on the return level values (p-value = 0.028), but again the strength of this linear relationship is weak because only 36% of the variation in the return level estimates is explained by its position with respect to the longitude and latitude.

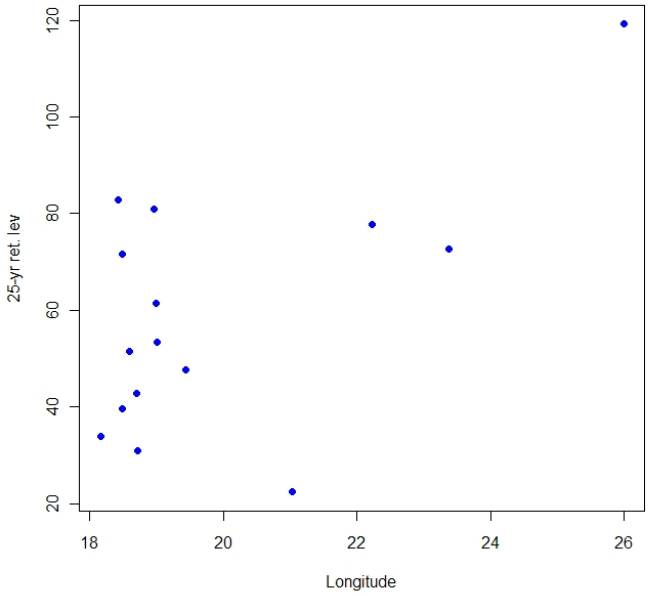
There is approximately a 3% chance that the positive relation between design values and longitude is due to chance, whilst for the latitude this is high at nearly 20%. At a 5% level of significance, the conclusion is that there is evidence in support of a linear spatial trend, although it explains a small proportion of the variation in the design values. It is anticipated that the remainder of the variation will be explained by the model for local spatial variation.

Table 5.1: Investigating the possibility of linear spatial trend in 25 year return level estimates

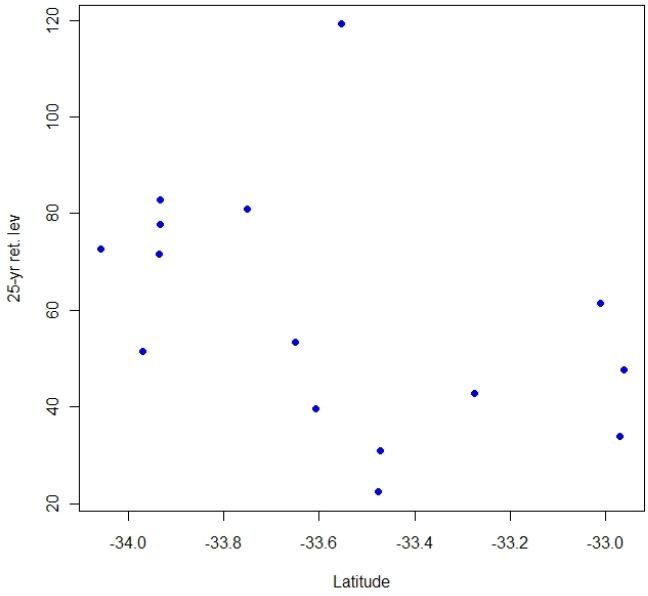
Coefficients	Estimate	Std. Err	Pr($>\|t\|$)
Intercept	-749.37	489.14	0.15
Longitude	5.91	2.45	0.03
Latitude	-20.59	14.84	0.19

The diagnostic plots, Fig 5.3, also concur that the linear spatial trend model is not a perfect fit. The design values for Ladismith, Molteno and Tygerhoek are influential – especially for Ladismith, at 22.42 mm this is small comparison to the others, whilst 119.16 mm at Tygerhoek is more than 70% greater than the values obtained for the other sites. Further, checking whether the residuals violate the assumption of normality, the q-q plot shows that this assumption is violated, especially the centre of the empirical residual distribution is lower than the normal distribution. At this point we also point out that such discrepancies could be the result of there being few observations – the normal approximation is reliable for sample sizes of at least 25–30 observations (Mukhopadhyay, 2000). In previous studies, the strength of the association between rainfall at the daily level of aggregation and elevation has been found to be weak (Goovaerts, 2000; Szolgay et al., 2009). A Similar observation is made in this study, as Fig 5.4 reveals lack of association between the 25-year 24-hour return level and elevation.

Therefore, two cases will be considered, a model for local spatial dependence in the design values without the trend surface and the case where the trend surface model and local spatial dependence of the residuals is considered.

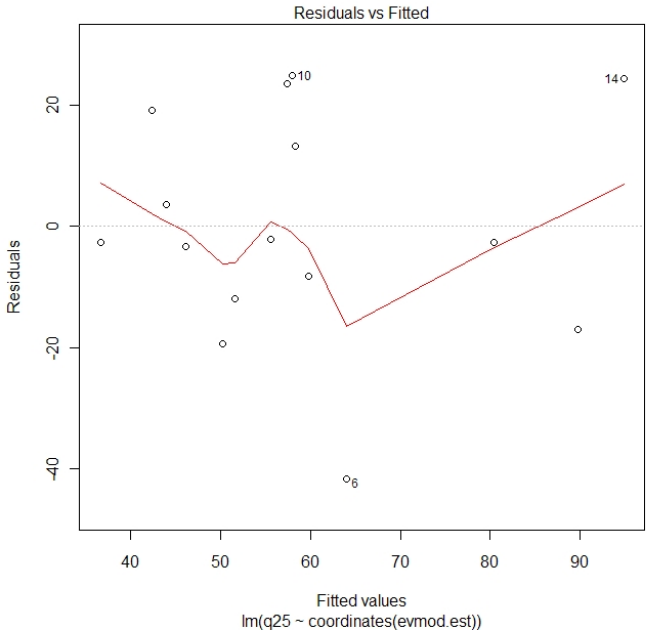


(a) 25 year return level estimate vs longitude

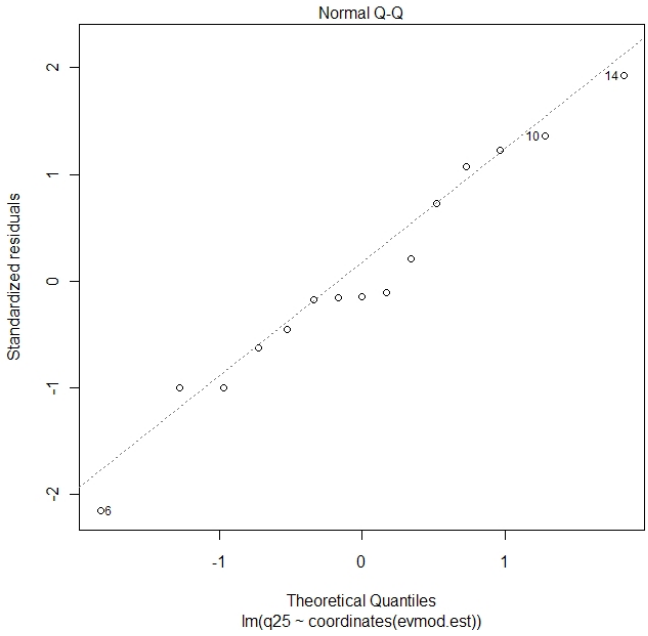


(b) 25 year return level estimate vs latitude

Figure 5.2: Relation of the 25 year return level estimate to the coordinates



(a) Fitted values against residuals



(b) Q-Q plot of residuals

Figure 5.3: Goodness-of-fit of linear spatial trend surface to the 25-year return level estimates in the study area

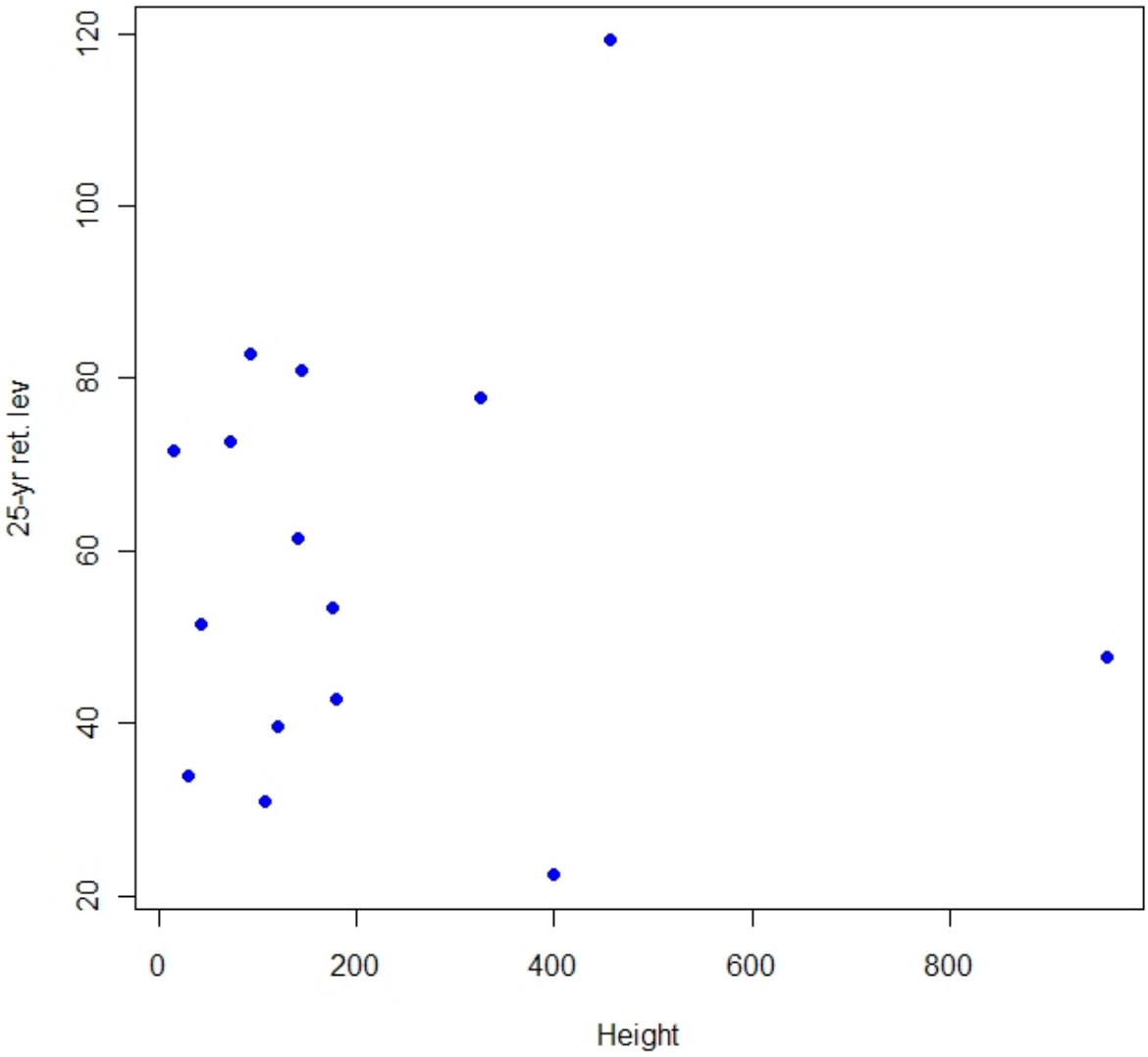


Figure 5.4: The relation of the 25-year return level estimates to altitude

5.2 Model for Spatial Correlation of Rainfall Return Levels

In this section the issue of few observations is initially dealt with, by space-time replication of the design values. For the replication in time, the design values of a specific period are considered as realization of an unobserved return level surface for that period. Specifically, 6 consecutive return periods are considered from 25 to 50 years in intervals of five years (see Table 5.2).

Table 5.2: Replicates of the design values in pseudo-time

Site	Long.	Lat.	q_{25}	q_{30}	q_{35}	q_{40}	q_{45}	q_{50}
Atlantis	18.483	-33.607	39.63	40.72	41.63	42.41	43.09	43.69
CPT Astro.	18.477	-33.935	71.56	75	77.99	80.66	83.06	85.26
CPT Int. Air.	18.597	-33.969	51.55	52.69	53.63	54.42	55.1	55.7
Excelsior	19.43	-32.963	47.58	49.34	50.77	51.98	53.01	53.92
Jonkers.	22.227	-33.934	77.86	82.49	86.52	90.09	93.31	96.24
Ladismith	21.035	-33.476	22.42	24.17	25.73	27.14	28.42	29.61
Langebaan.	18.157	-32.972	33.99	35.56	36.91	38.1	39.17	40.13
Langgewens	18.706	-33.276	42.84	44.23	45.41	46.42	47.32	48.13
Malmesbury	18.718	-33.472	30.95	31.75	32.39	32.92	33.38	33.77
Molteno	18.417	-33.933	82.8	86.66	90.02	92.99	95.68	98.12
Paarl	18.967	-33.75	80.96	83.15	84.96	86.51	87.87	89.06
Plettenberg.	23.372	-34.058	72.77	75.16	77.18	78.92	80.46	81.84
Porteville	18.994	-33.012	61.52	64.9	67.93	70.68	73.22	75.57
Tygerhoek	25.993	-33.553	119.16	130.82	141.41	151.18	160.27	168.8
Wellington	19.006	-33.651	53.45	55.58	57.38	58.96	60.35	61.61

The extension of the variogram to space-time is justified if the design levels for each of the 6 return periods are the same (Sterk and Stein, 1997). Further, the spatial correlation model takes into account the relationship between values located a certain distance apart, therefore after standardization, observations at new sites can be formed by displacement from the original sites by a constant distance. We standardize by the ratio of the overall average design value and each return period's average design value. The overall estimated average design value is 65.44 mm and the individual return period value average design values are given in Table 5.3. The return levels are strictly positive values, hence the use of the ratio of

Table 5.3: Average design values for each return period – 25 to 50 years

$\bar{q}_{25}$	$\bar{q}_{30}$	$\bar{q}_{35}$	$\bar{q}_{40}$	$\bar{q}_{45}$	$\bar{q}_{50}$
59.27	62.15	64.66	66.89	68.91	70.76

averages instead of the normal standardization by ratio of the difference from the mean and

the standard deviation. To check whether the structure of the spatial correlation is the same across the different return periods, the variogram cloud (Fig 5.5) was computed. Notice that the points on the variogram cloud appear stacked, an indication of a similar structure for the different return periods.

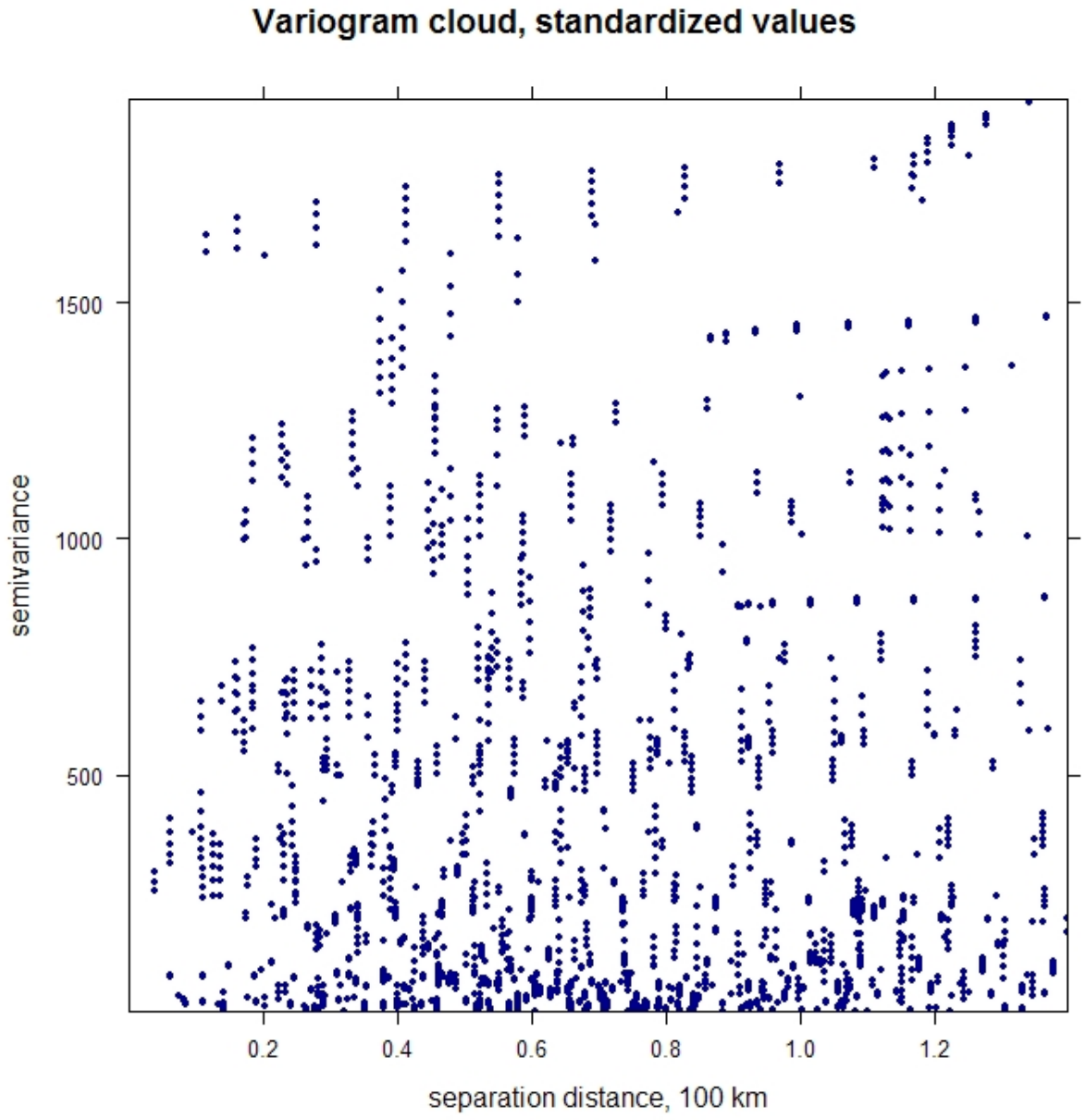


Figure 5.5: Pooled variogram cloud of the standardized 25 up to 50 year return level estimates

It is difficult to make out the structure of the spatial correlation from the cloud, hence the sample variogram plots (Fig 5.6) were computed. Recall that in this study two cases are considered, the first is assuming lack of regional trend and the second is where a linear

trend surface is considered, hence the variogram model (Fig 5.6(b)) is for the residual after removal of the regional trend. As anticipated the semivariances for the residuals would be lower as some of spatial variation has been accounted for by the regional trend, but the spatial correlation structure remains similar for the two cases.

The variogram model was estimated by curve-fitting, namely minimizing by the weighted non-linear least squares. Three models were considered, with the results given in Table 5.4. From the structure of each sample variogram, there appears to be a definite sill, which is gradually reached. Therefore, the spherical class of models were considered. The exponential model was also considered, but it was the first to be eliminated. The elimination of this model was due to the nugget estimate being zero, which is impractical in our case since the design values are estimates and by definition they have a certain degree of uncertainty. From the plots in Fig 5.6, the spherical model does not appear to be close fit in comparison to the penta-spherical model. The sill in the sample variogram is reached gradually and judging by its shoulder, the sill is reached more abruptly in the spherical model. More formally, the curve-fitting procedure is based on minimizing the weighted sum of squared differences of the curve to the data points. In this case, this is minimum for the penta-spherical model.

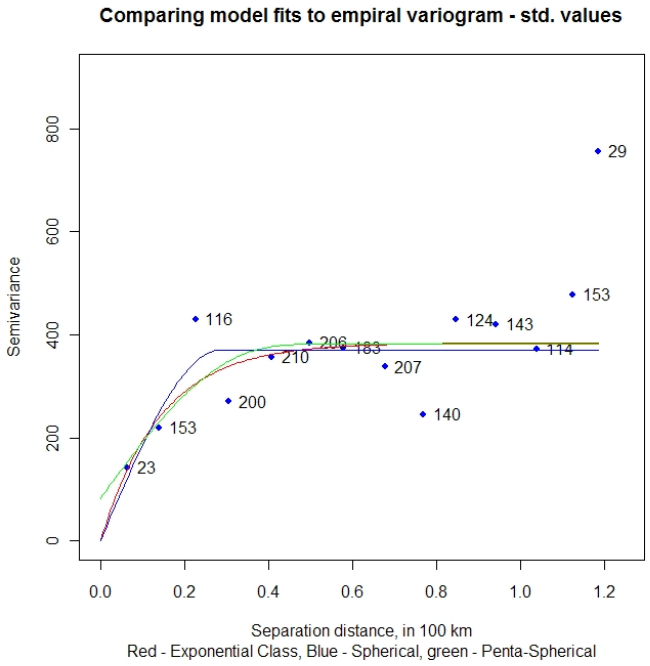
Table 5.4: Variogram model parameter estimates for the pooled design values

Model	Exponential	Spherical	Penta-spherical
Nugget	0.00	0.00	86.38
Partial Sill	384.29	370.97	295.31
Range	0.14	0.29	0.51
Sum of Sq. Err.	155	87.30	69.60

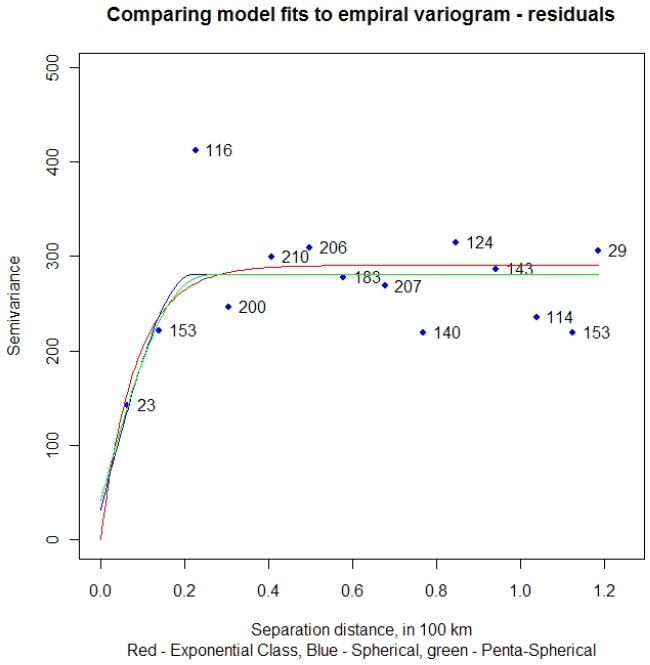
Similar analysis was carried out for the second case, the residuals, with the results given in Table 5.5. Note that the sum of squared errors is again minimum for the penta-spherical model. Recall that these are the model parameter estimates for the pooled variogram, hence to arrive at the parameter estimates for the 50 year return level, we have to multiply the sill and the nugget by the square of the reciprocal of the standardization factor. The nugget and sills for the first case are $101\,\mathrm{mm}^2$ and $446\,\mathrm{mm}^2$ and for the second case $48.21\,\mathrm{mm}^2$ and $328.17\,\mathrm{mm}^2$ respectively. The range for the two cases is as given in Table 5.4 and Table 5.5 respectively.

Table 5.5: Variogram model parameter estimates for the pooled residuals, after removal of regional trend

Model	Exponential	Spherical	Penta-spherical
Nugget	0.00	31.62	41.23
Partial Sill	290.61	249.56	239.45
Range	0.08	0.22	0.28
Sum of Sq. Err.	59.2	56.7	52.2



(a) Variogram for the standardized design values



(b) Variogram for the residuals

Figure 5.6: Empirical variogram and potential variogram model curves superimposed. In red is the exponential model, blue is the spherical and in green the penta-spherical model.

5.3 The Return Level Maps

In this section kriging maps are obtained. In both the ordinary and universal kriging cases, due to the small number of observations and the vastness of the study region, sites where values were available can be clearly seen on Fig 5.7. In ordinary kriging, the predicted values are the same as the observed in a circular area around it, whilst in universal kriging these spread in the direction of the regional trend. The design rainfall level in most of the Western Cape is between 40 mm and 80 mm, but as one moves east this becomes higher, largely due to the influence of the large design value obtained for Tygerhoek. As anticipated for ordinary

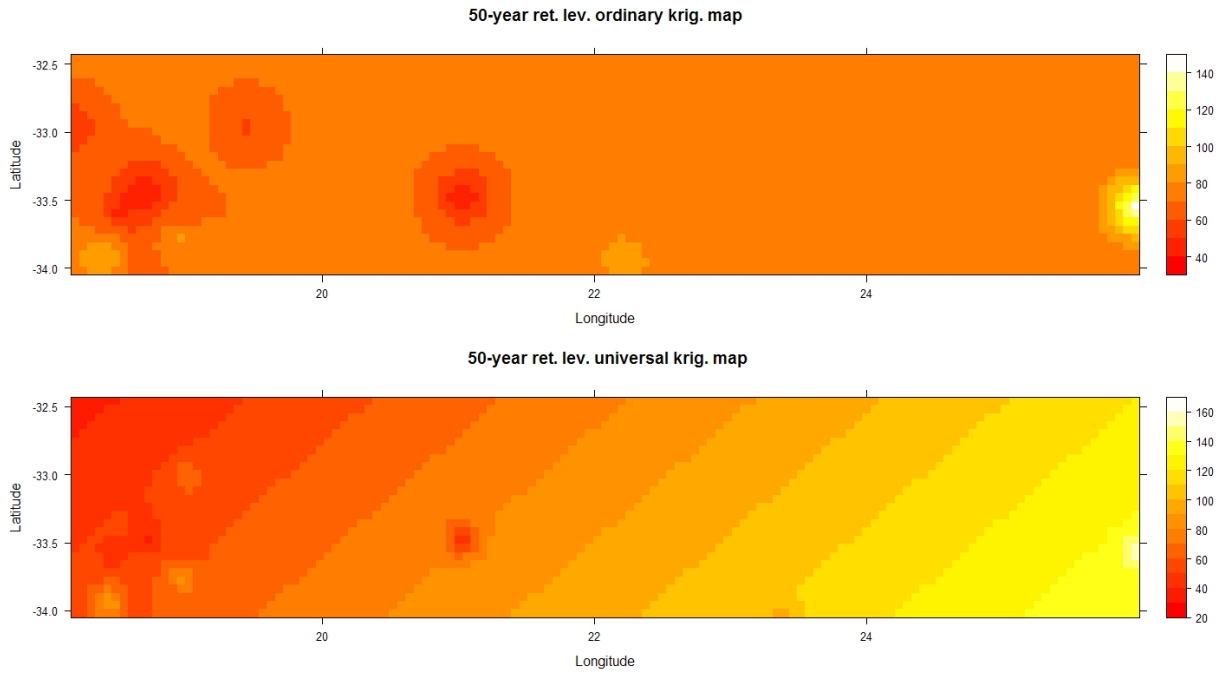


Figure 5.7: Comparison of maps derived by ordinary Kriging against universal kriging for the 50 year 24-hour return level estimate

kriging, the kriging error is greatest at unsampled locations, but the this is lower then in the universal kriging case (refer to Fig 5.8). Region of minimum error for the universal kriging in Fig 5.8 is in the inner spherical area containing all the sites, with the exception of Tygerhoek.

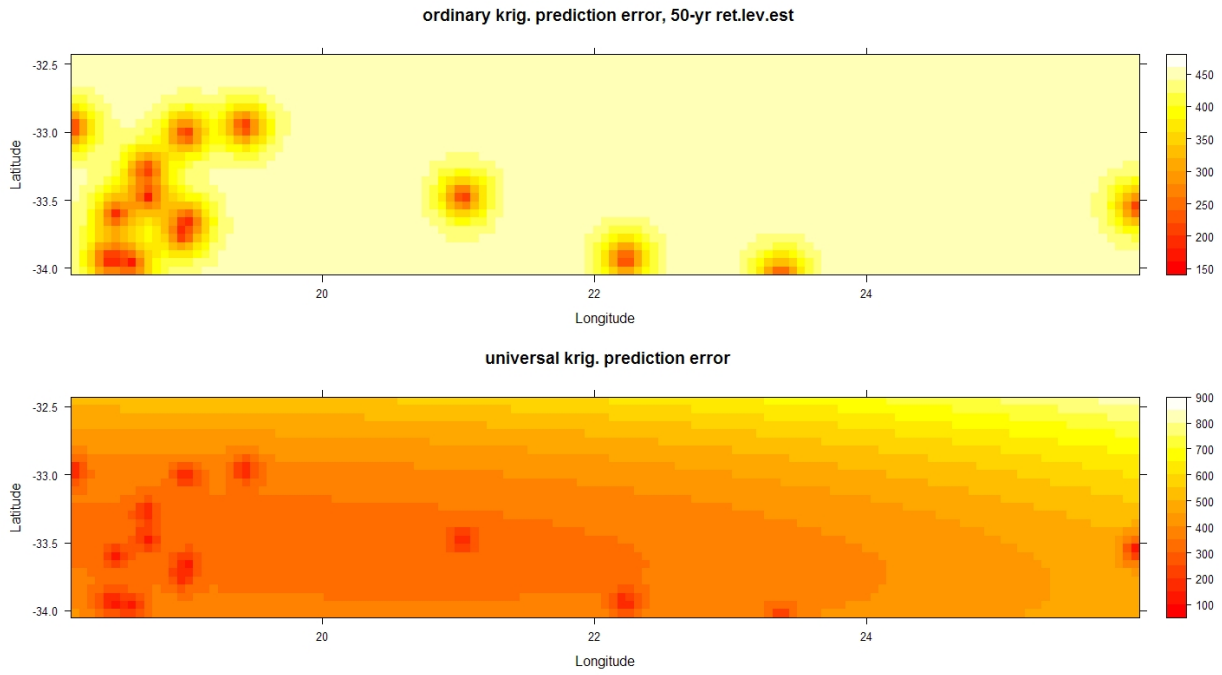


Figure 5.8: Map of the uncertainty about the ordinary and universal kriging estimates for the 50 year 24-hour return level estimate

Chapter 6

Concluding Remarks

The study of heavy rainfall is useful in hydrological applications. The main objective set out for this study included investigating whether any trends temporal and spatial can be detected in heavy rainfall over the study area. Focus was specifically in the winter season since the Western Cape is classified as a winter rainfall region, since this is the season where the area receives most of it's rainfall. This is in contrast to the eastern parts of South Africa which receive the bulk of the rainfall during summer months. To address our objective a two-tiered approach was followed. Firstly, at each site a point process extreme value model was fitted to obtain an estimate of the N-year 24-hour winter rainfall return level (or design values). Secondly, in the geostatistical analysis spatial correlation in design rainfall values obtained from the extreme value analysis was investigated. It was shown that rainfall design values vary spatially over the Western Cape, despite the challenge posed by the small sample size.

The point process approach to extreme value theory was used to quantify rainfall return levels for discrete sites within the Western Cape. The 50-year 24-hour winter rainfall return level is the single-day average amount of winter rainfall that can be expected to be exceeded with 0.02 probability. This gives a measure of the rarity of such events. Quantifying this statistic for this study involved the crucial step where at each station rainfall events are classified as "heavy" if they are above a selected threshold. To select an appropriate threshold, responses of the point process extreme value model parameter estimates to changes in the sample of excesses were analyzed. While it is desirable to use a value obtained from an expert in hydrology, in this study preference was given to data-driven methods to select an appropriate value for each station. Short-range temporal correlation was high for some stations, thus to ensure model accuracy declustering the series of exceedances was performed. At these sites extremal index values were smaller than 0.80 for the chosen thresholds. Model fitting preceded the threshold selection exercise. Quantile-quantile and return level plots were used for evaluating the model's goodness-of-fit. These did not reveal any inadequacies, except for high uncertainty for large quantiles for sites like Porteville, Ladismith and Tygerhoek. The

uncertainty was due to the influence of the largest excess being nearly twice the size of other excesses at these sites. Direct fit of the GPD to threshold exceedances could have been used to quantify the rainfall return levels. In this study the point process approach was chosen because it's parametrization in terms of the generalized extreme value distribution (GEV), ensures that the scale parameter is invariant to threshold choice.

Geostatistical analysis of 24-hour winter rainfall design values was performed to determine if they were spatially dependent. Kriging was chosen over other methods of geostatistical analysis. Theoretically, kriging is the Best Linear Unbiased Estimator, given a model for spatial correlation (i.e. the variogram model). Hence the first step in performing kriging is to choose an appropriate variogram model and then obtain an estimate of it's parameters. There are several variogram models that are available, however the most common in practice are the exponential and spherical classes of models (Cressie, 1993). To estimate variogram model parameters a sufficiently large sample is required, which was a challenge in this study since there were only fifteen stations and therefore only fifteen 50-year winter rainfall design values available for spatial analysis. Our extension of standardized return level estimates spatially through the method of Stein and Sterk (1999); Sterk et al. (2004) was an attempt to obtain a reasonable estimate of variogram model parameters from a larger pool of observations. A weighted least squares approach was followed to fit a variogram model for two cases: one where regional trend in design values was ignored and another which accounted for the trend between design values and geographical co-ordinates. In the first case a definite sill observed from a sample variogram informed the decision to fit a penta-spherical model where variation due to micro-scale and measurement error was found to be of the order of $86.38\,\text{mm}^2$. We found that sites within 50 km of each other can be expected to show similarities in 25-year 24-hour design rainfall levels. When the regional trend was taken into account the range dropped to approximately 30 km. A limitation in this exercise was the assumption that design values are independent across return periods. Possible non-independence might have an effect on the resultant variogram model.

After obtaining variogram model parameters, the 50-year 24-hour winter rainfall return level map of the study area was obtained through kriging. Since the sampled area was sparse in comparison to the spatial extent of the entire study area, uncertainty about the predicted design values was higher than desired. It was interesting that ordinary kriging error was less than that obtained for universal kriging, with the former being nearly $500\,\text{mm}^2$ at it's highest in comparison with the maximum error of $900\,\text{mm}^2$ obtained for universal kriging. The 50-year 24-hour design rainfall over most of the study region was found to be between 40 mm and 80 mm, amounts that are double these levels for stations on the east coast of the Western Cape. Therefore, it is deduced in this study that heavy rainfall does vary spatially over the Western Cape.

6.1 Recommendations

A study looking at the issue of mapping design rainfall was carried out by Szolgay et al. (2009). The objective was to compare the quality of design annual maximum 24-hour rainfall maps produced by three spatial analysis methods. It was also an investigation into whether there would be any differences in the quality of the map produced, when the method of first creating a regular grid of daily precipitation values from point-referenced data, thereafter deriving the design values for each grid site was compared with more traditional approaches of doing site-wise extreme value analysis followed by mapping of design values. The maps produced were the 2-year and 100-year rainfall return levels deemed reliable in engineering hydrology for use in rainfall-runoff studies and flood estimation. This study differs in that the point process model was applied to get the return level estimates, with interest being in the 25-year and 50-year return period due to the length of the data at each site and concerns about extrapolation. The study areas are different, however an issue common to both studies is the small sample size. The authors carried out kriging without trying out ways to mitigate the issue of reliability of variogram estimation when the sample size is small.

In this study the space-time extension of the observation for variogram modelling provided one alternative on the issue of obtaining a return level map given a spatially scarce sample. When densely sampled covariate information which is highly correlated with rainfall can be obtained, a recommendation is to employ multivariate geostatistical techniques such as external drift kriging to obtain the return level map. An area of further research would be to investigate whether an external drift kriging approach leads to an improvement in results obtained in this study. Another limitation of the geostatistical model was the lack of inclusion of precision estimates of the return level estimates. Incorporation of uncertainty from a preprocessing step can result in erroneous maps due to error propagation. In light of this limitation an approach which incorporates standard errors of return level estimates in the geostatistical model such as the hierarchical Bayesian model approach is recommended and will be pursued as an area of further research.

The information provided by this study provides useful insight for planning flood prevention infrastructure and regional development. More work needs to be done to improve the accuracy of the estimates and reduce the high level of uncertainty in the return level surface estimate. However, the result that design values for the east coast are much higher than the rest of the Western Cape is a sufficient indicator that regional development guidelines for the Western Cape need to differentiate between the different areas within the province. For instance areas on the east coast of the province would require construction materials developed to withstand a 24-hour rainfall event that is higher 160 mm. Further, storm-water and drainage systems for the area would need to be planned in anticipation of rainfall events of such magnitude. Disaster managers can also use this information in demarcating areas that are most at risk of flooding as a result of heavy rainfall, which is helpful in deploying resources during an emergency.

Appendix A

Appendix

This chapter consists of graphics of threshold diagnostic indices that were used in evaluating the fit of the point process extreme value model. Estimated model parameters that were short-listed for each site are given as Table A.1.

A.1 Threshold Sensitivity Analysis

In this section threshold sensitivity diagnostics for each site are shown. From a sensitivity analysis two thresholds were chosen for each site. Models were fitted using both thresholds, and for some stations declustering was performed to account for temporal dependence, resulting in three fitted models for those stations.

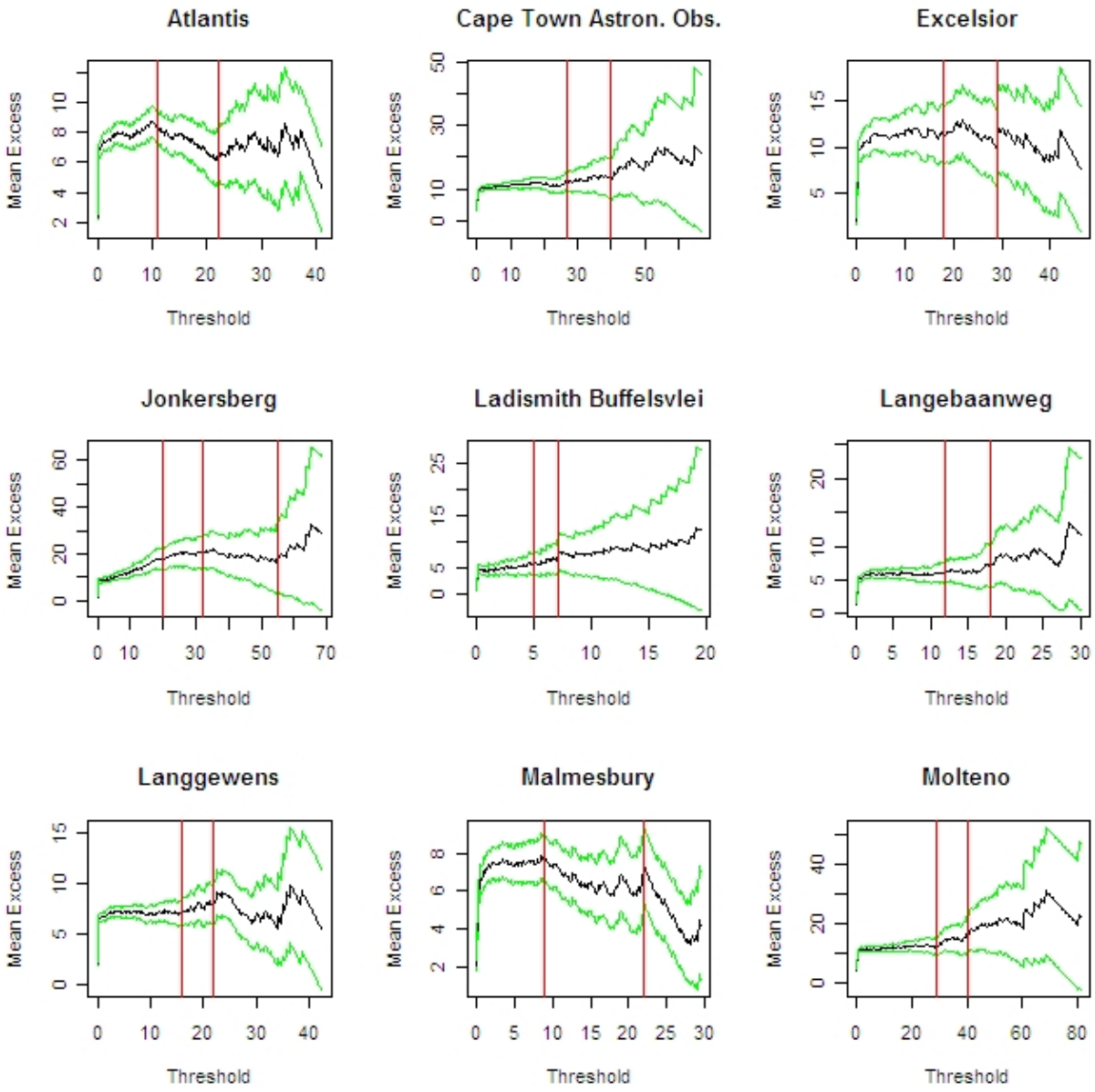


Figure A.1: Mean residual life plots: Atlantis to Molteno

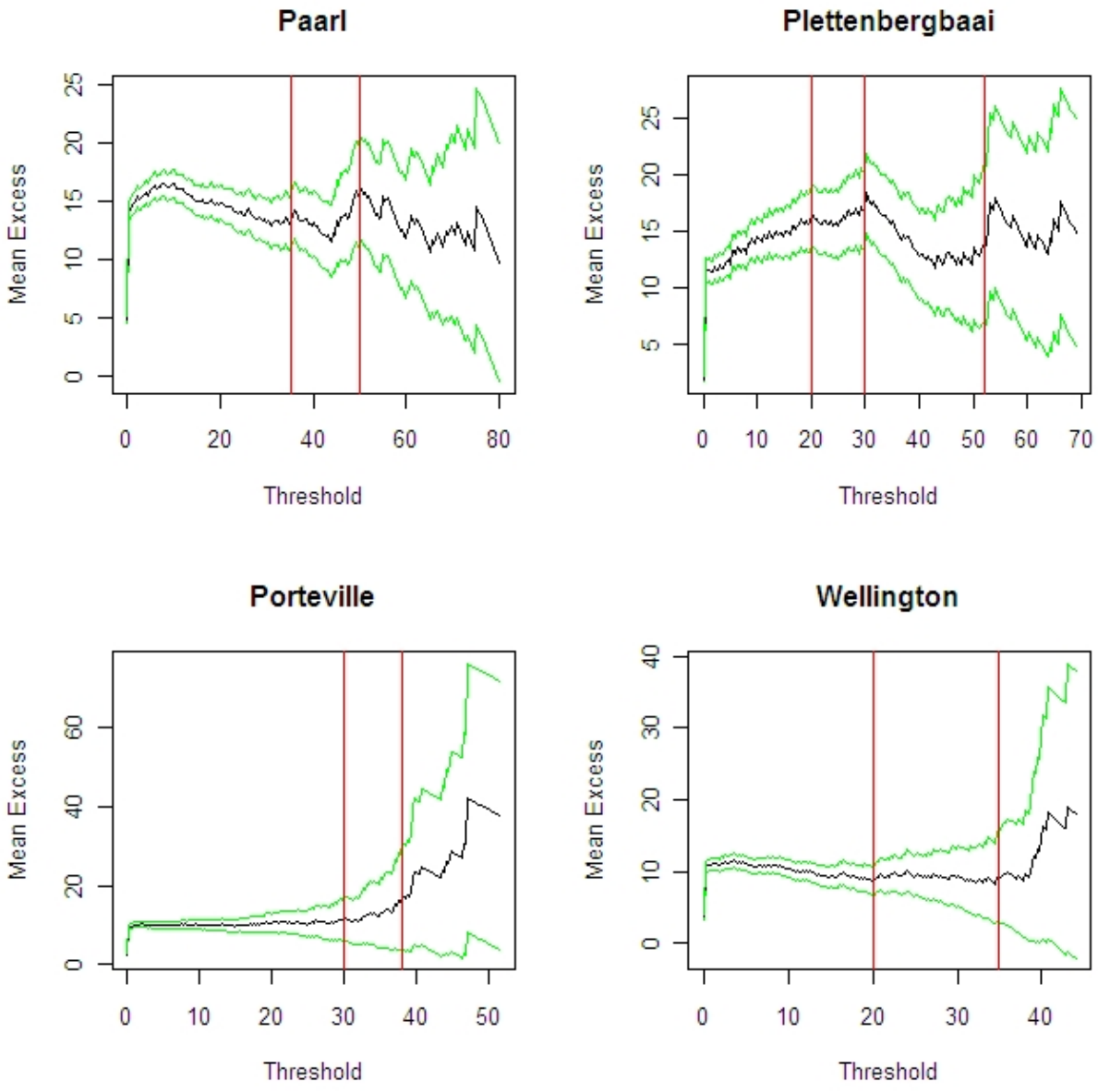
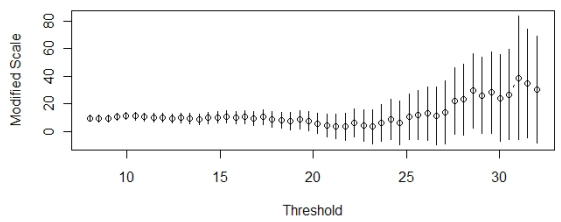
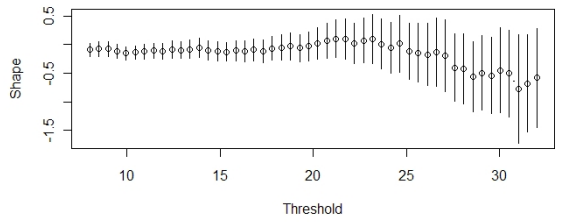
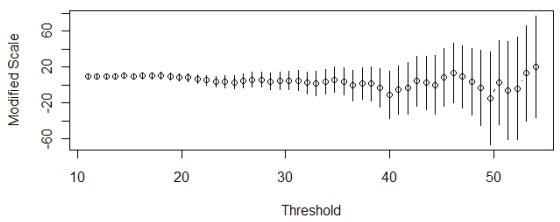


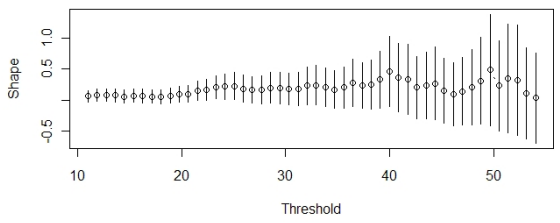
Figure A.2: Mean residual life plots: Paarl – Wellington



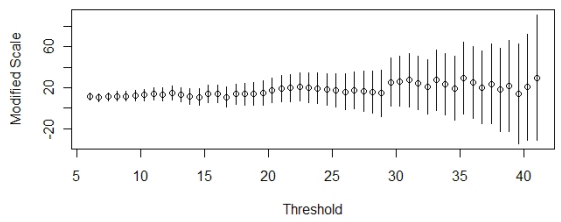


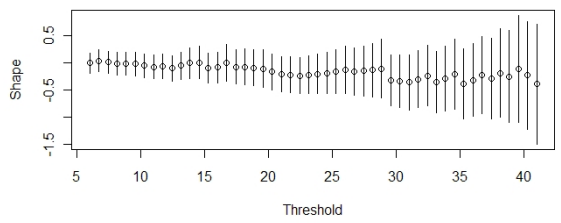
(a) Atlantis



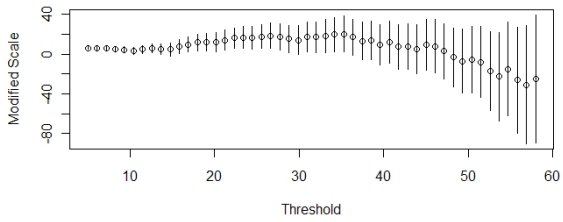


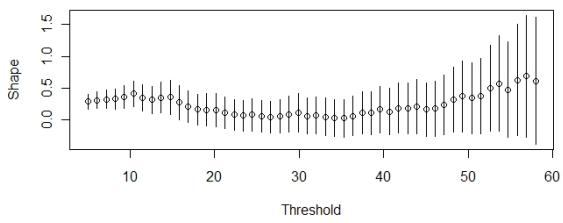
(b) CPT Astr Obs





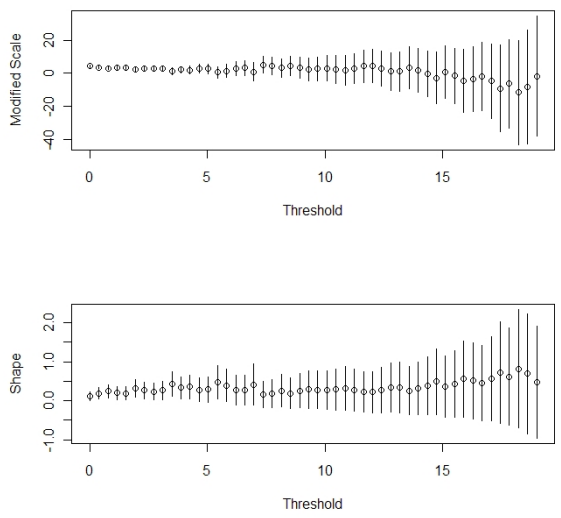
(c) Excelsior



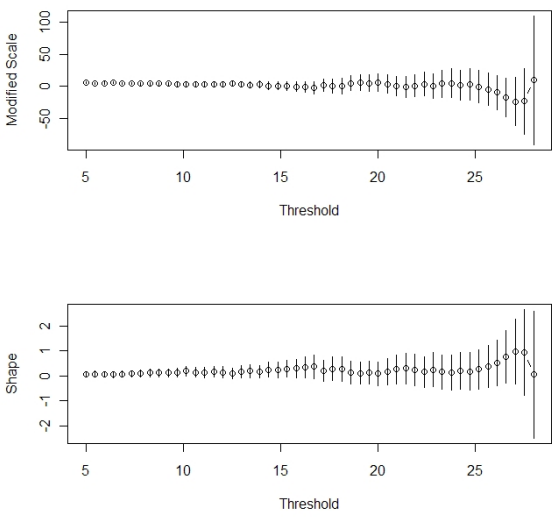


(d) Jonkersberg

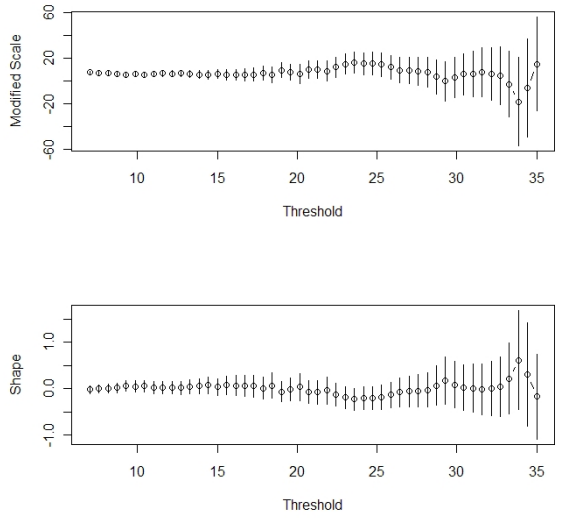
Figure A.3: Threshold stability plots: Atlantis – Jonkersberg



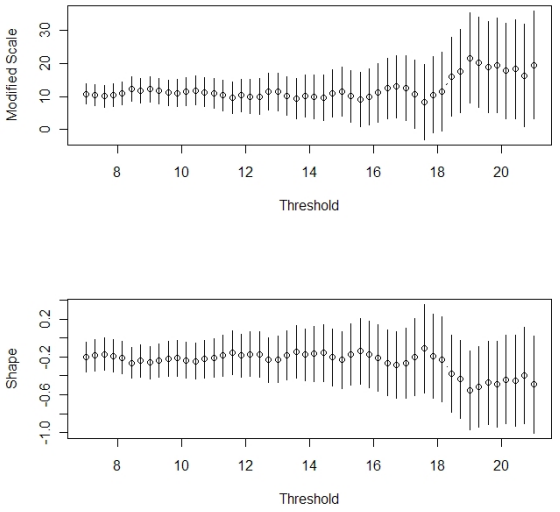
(a) Ladismith



(b) Langebaanweg

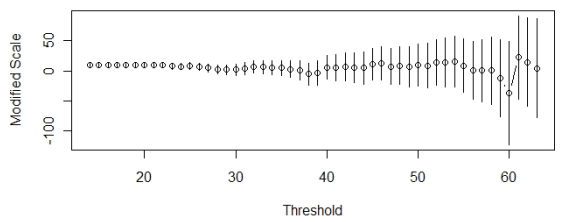


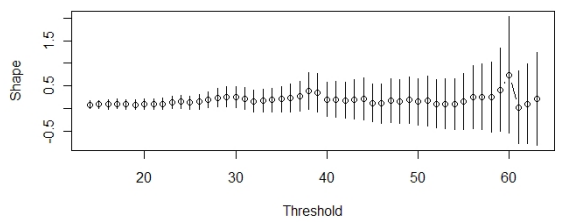
(c) Langgewens



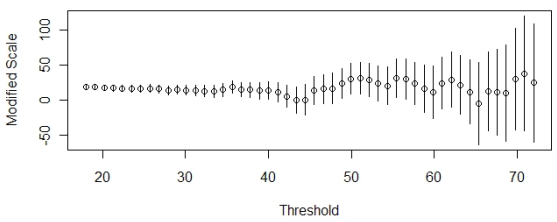
(d) Malmesbury

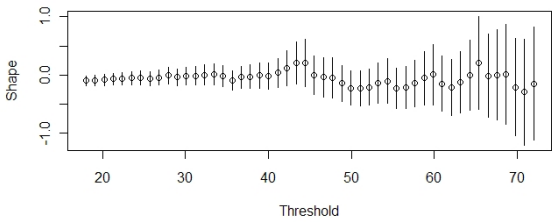
Figure A.4: Threshold stability plots: Ladismith – Malmesbury



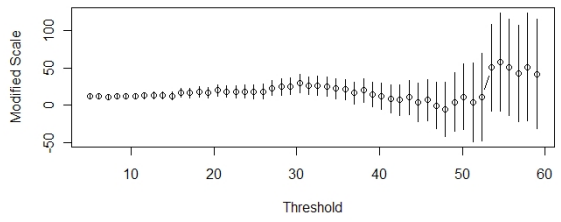


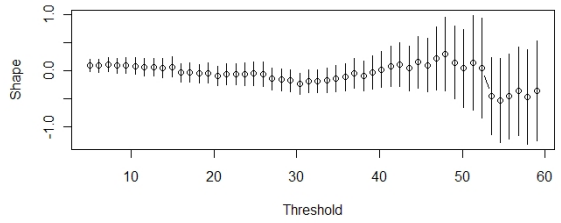
(a) Molteno



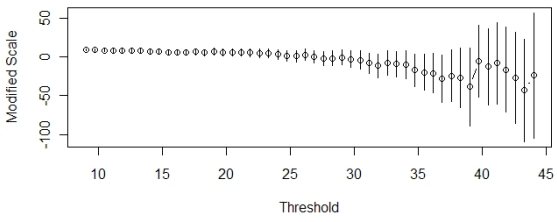


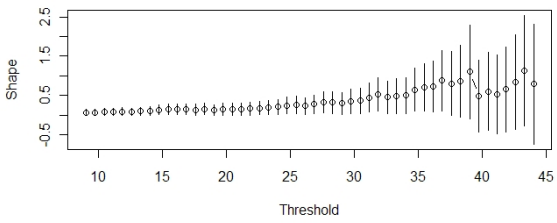
(b) Paarl





(c) Plettenbergbaai





(d) Porteville

Figure A.5: Threshold stability plots: Molteno – Porteville

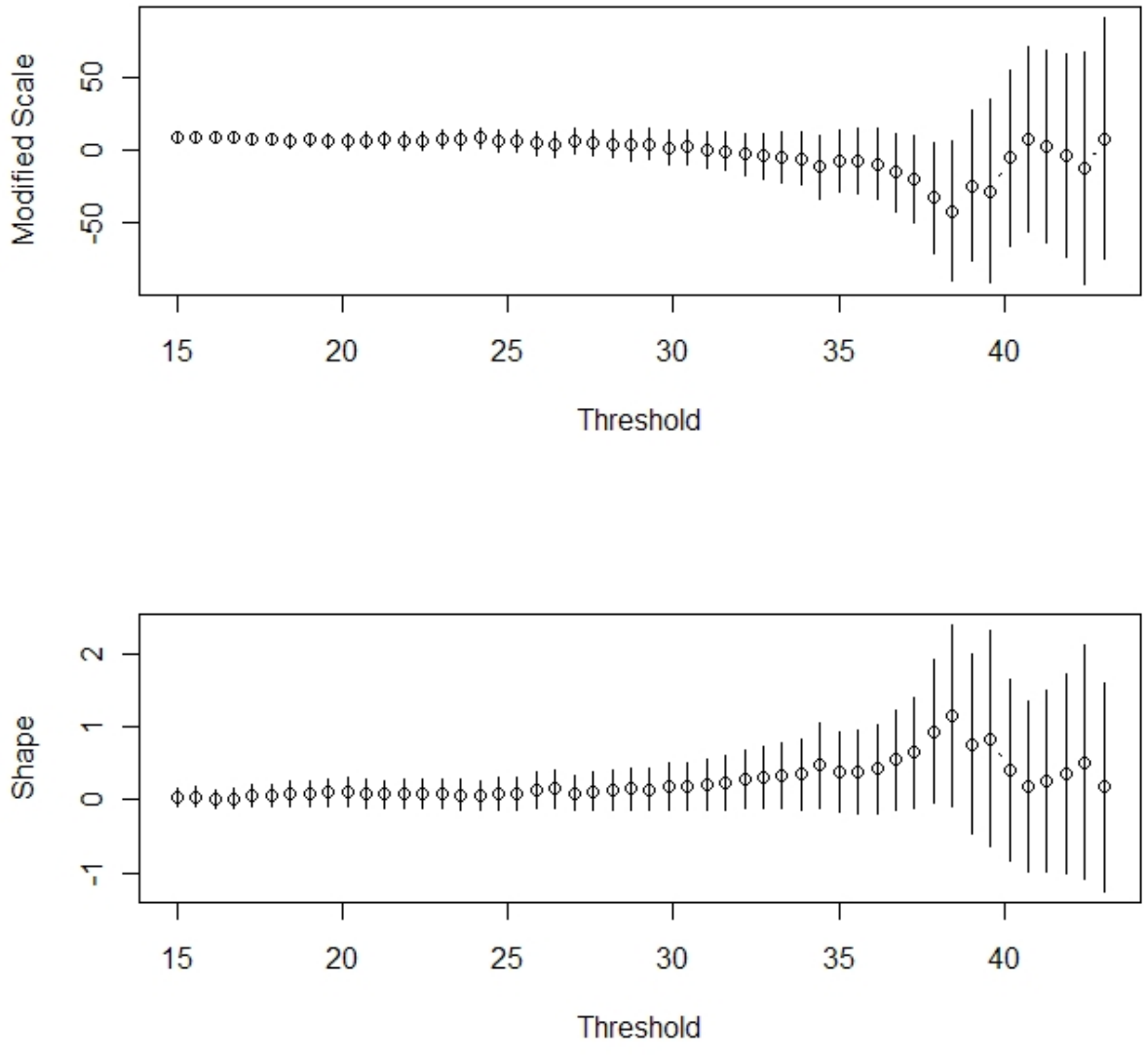
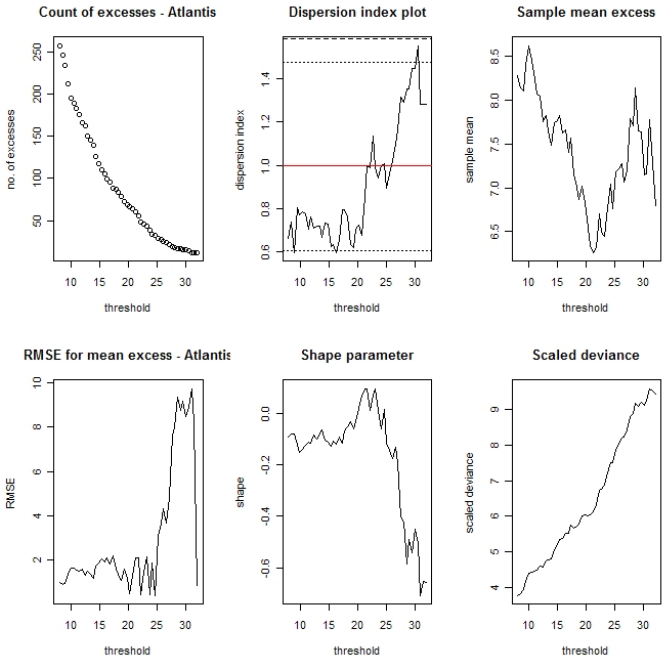
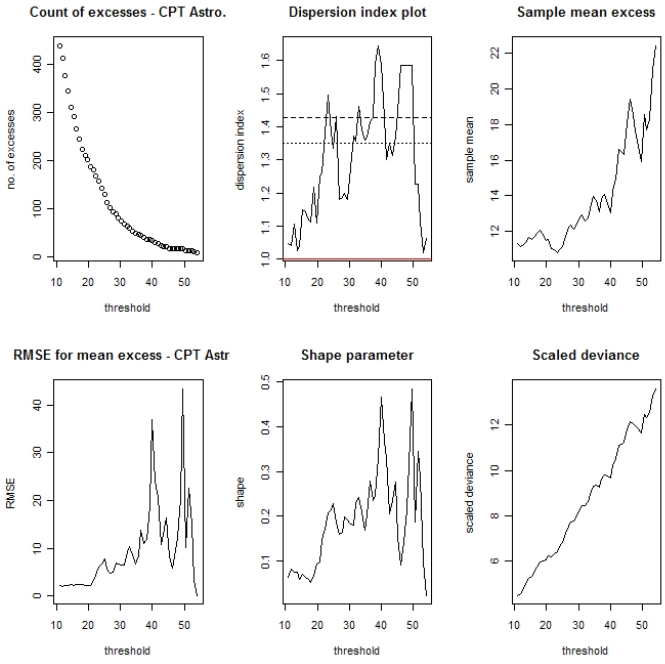


Figure A.6: Threshold stability plot: Wellington

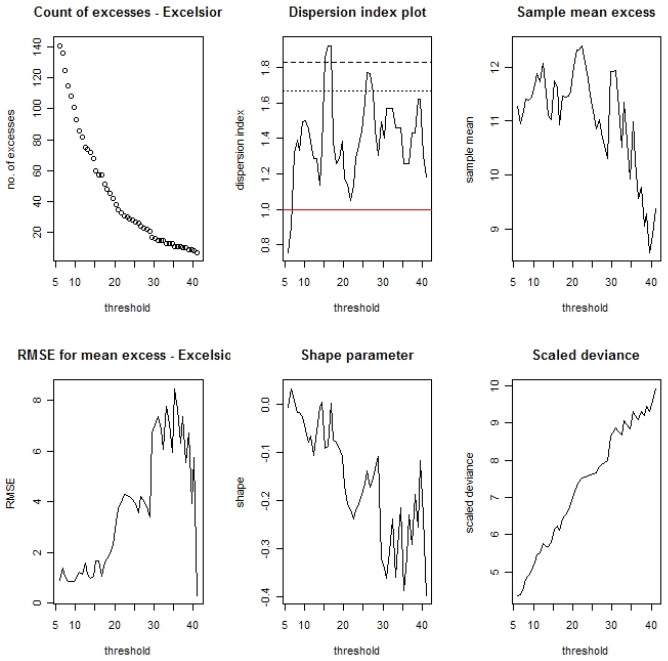


(a) Atlantis

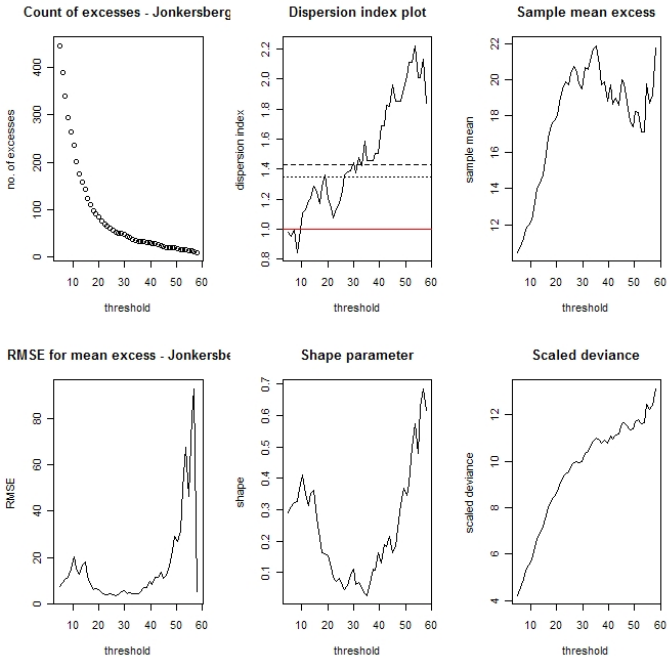


(b) CPT Astr. Obs.

Figure A.7: Additional threshold sensitivity diagnostics: Atlantis – CPT Astro.

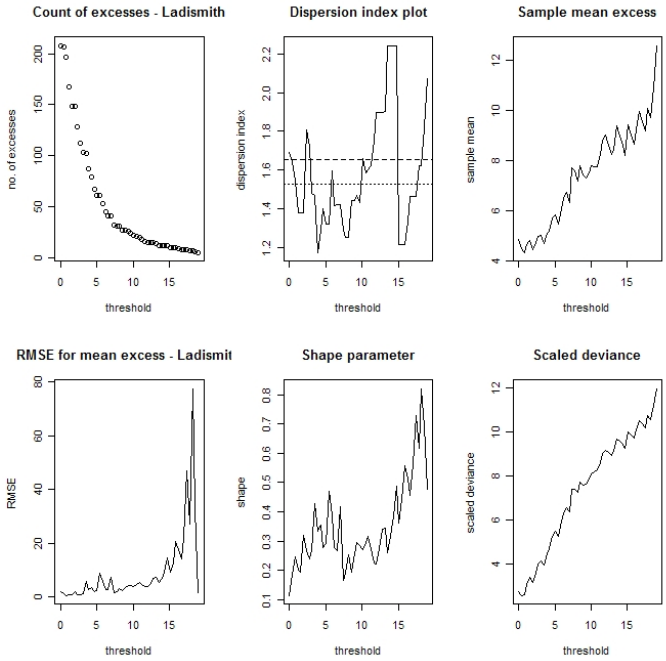


(a) Excelsior

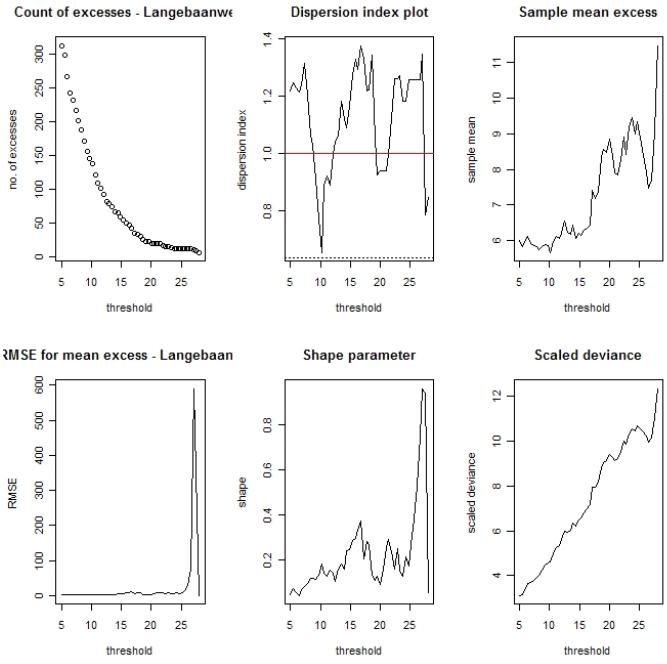


(b) Jonkersberg

Figure A.8: Additional threshold sensitivity diagnostics: Excelsior – Jonkersberg

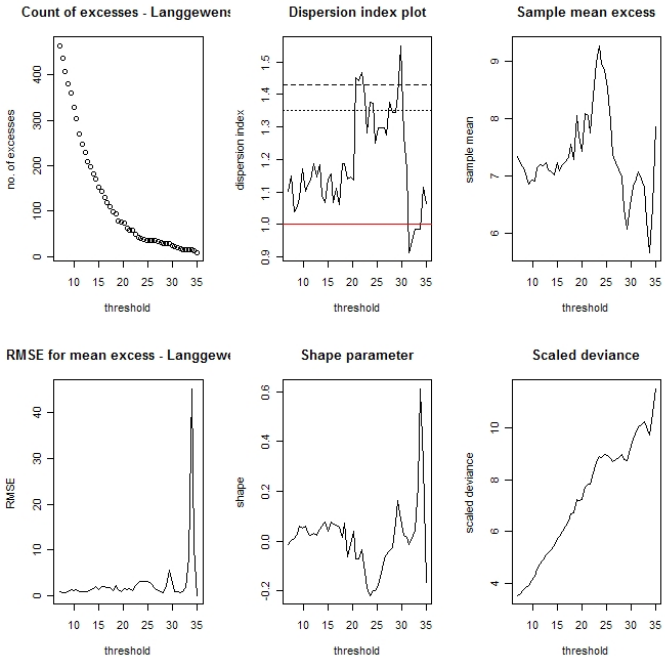


(a) Ladismith

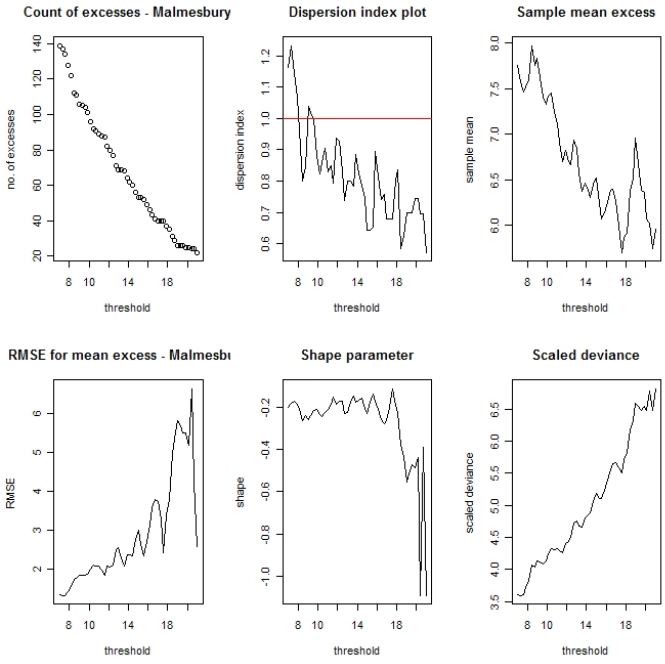


(b) Langebaanweg

Figure A.9: Additional threshold sensitivity diagnostics: Ladismith – Langebaanweg

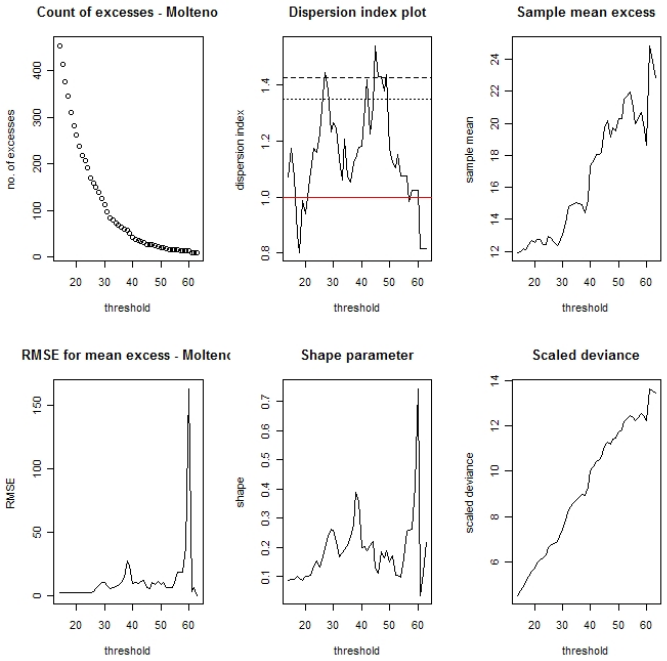


(a) Langgewens

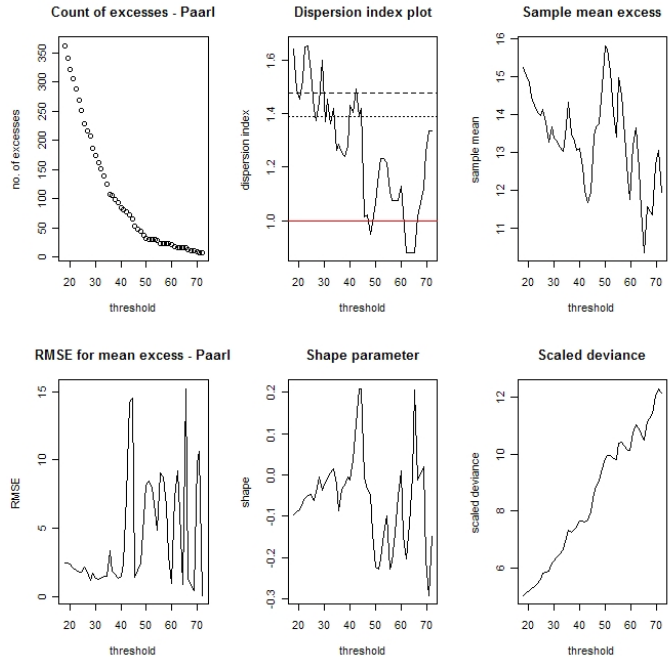


(b) Malmesbury

Figure A.10: Additional threshold sensitivity diagnostics: Langgewens – Malmesbury

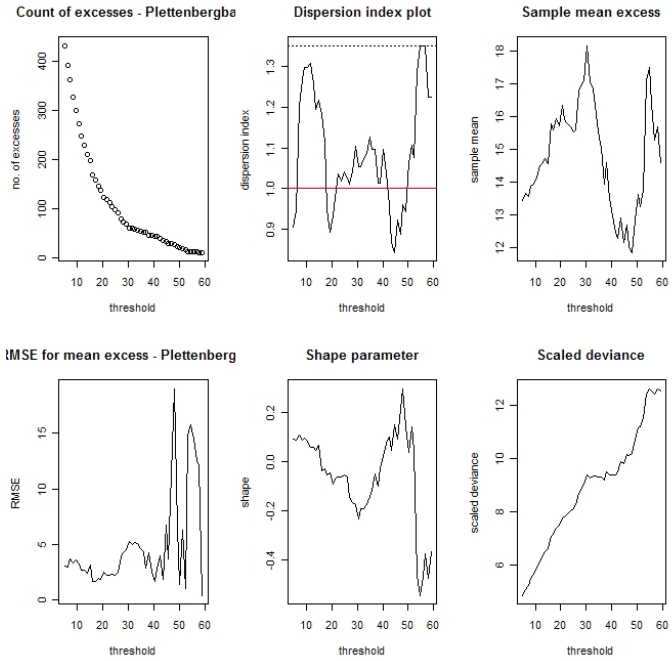


(a) Molteno

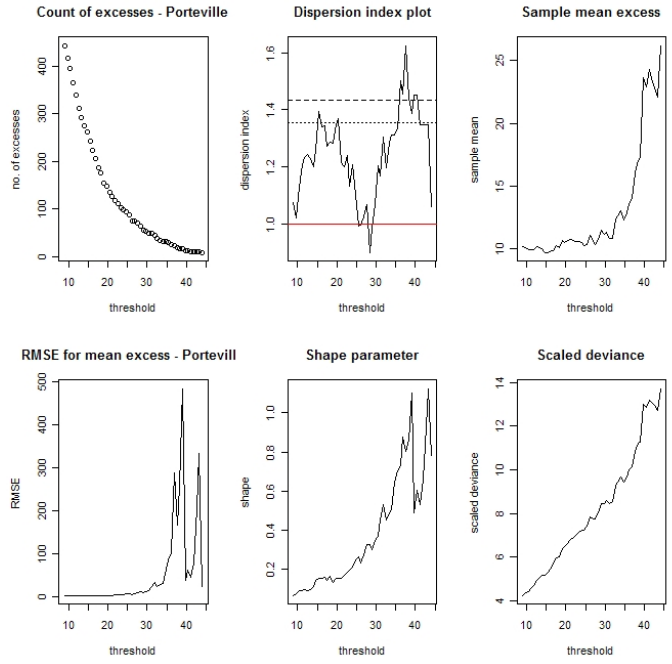


(b) Paarl

Figure A.11: Additional threshold sensitivity diagnostics: Molteno – Paarl



(a) Plettenbergbaai



(b) Porteville

Figure A.12: Additional threshold sensitivity diagnostics: Plettenbergbaai – Porteville

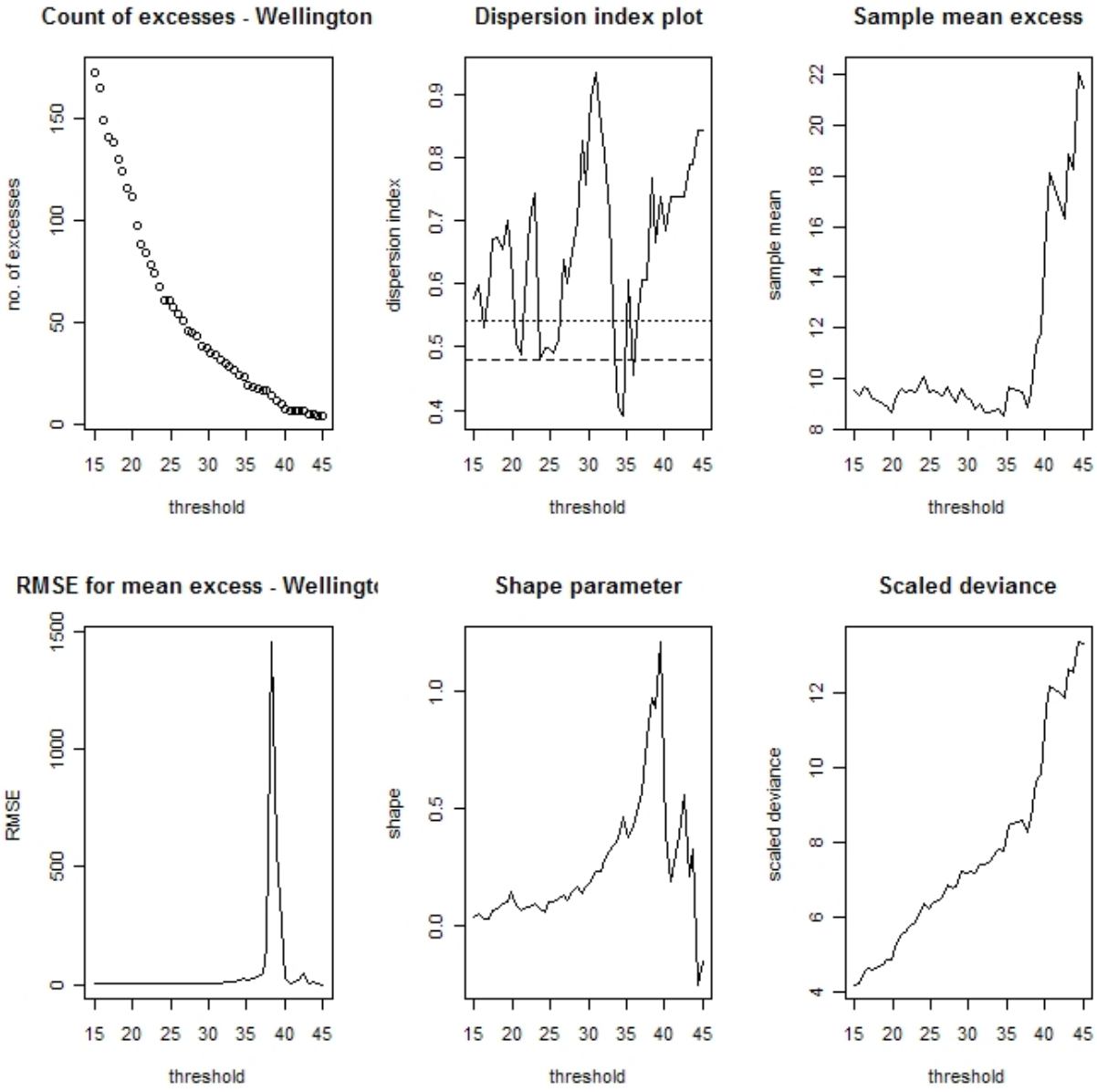
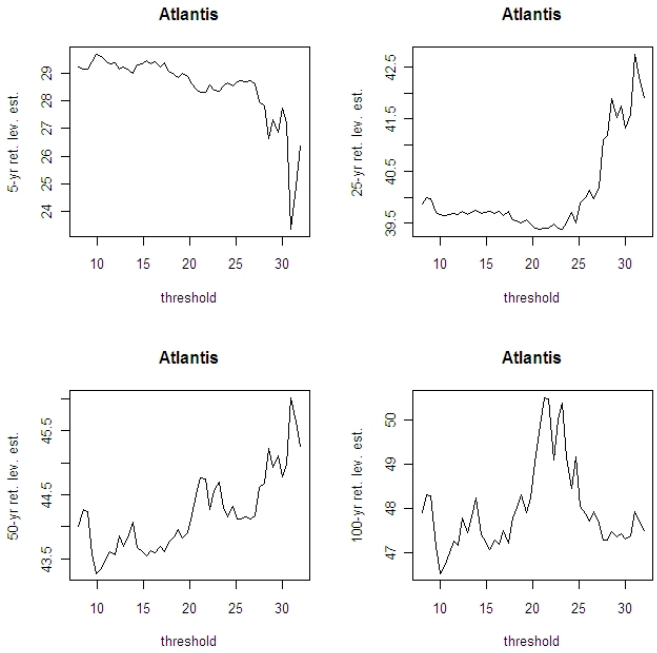
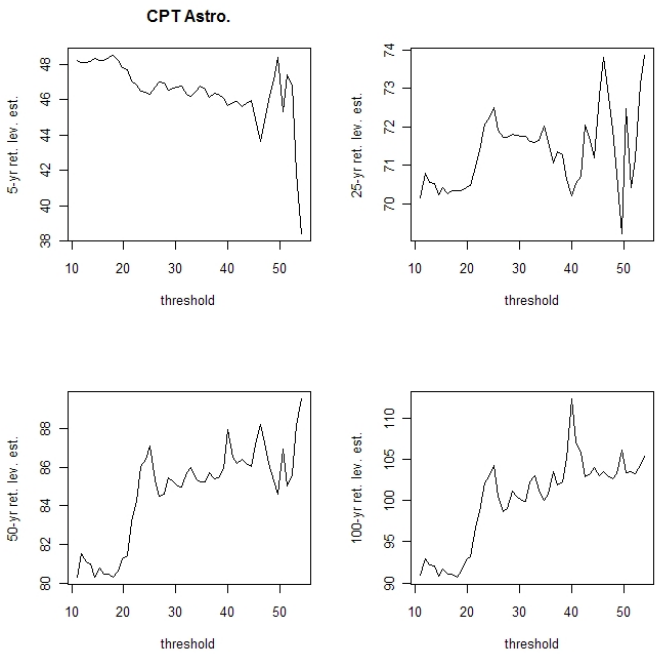


Figure A.13: Additional threshold sensitivity diagnostics: Wellington

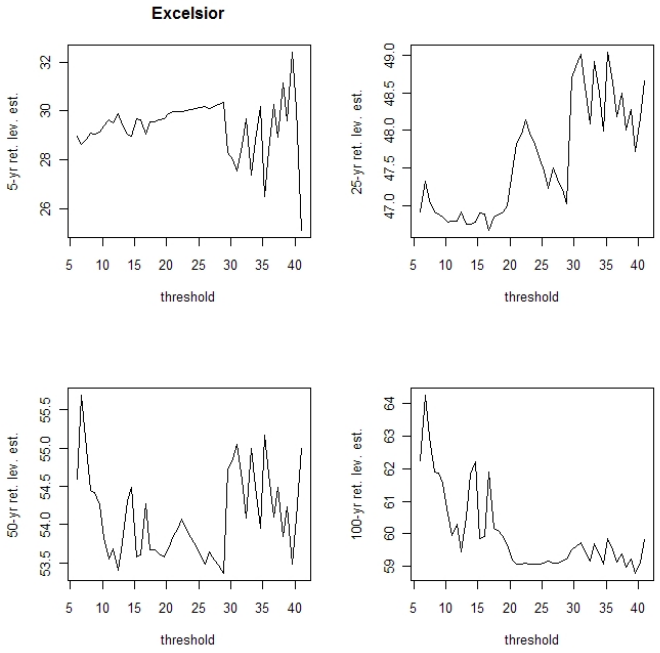


(a) Atlantis

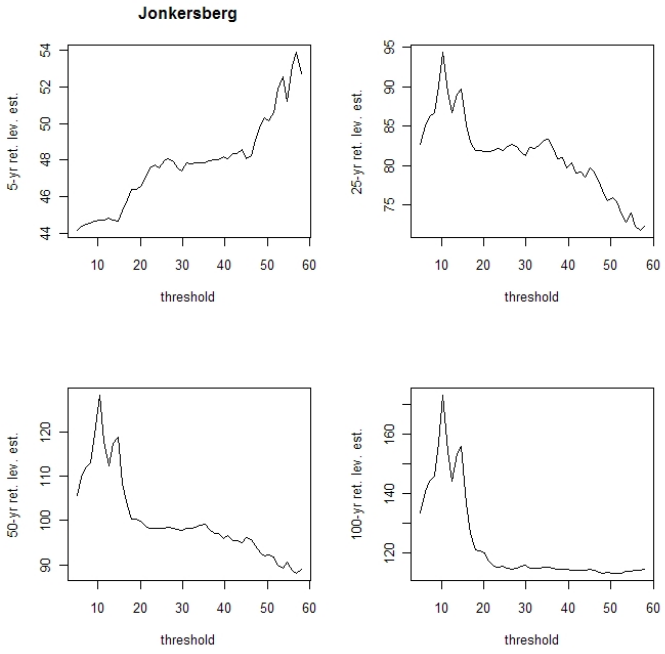


(b) CPT Astr.

Figure A.14: Sensitivity of the return level estimates to the threshold values: Atlantis – CPT Astr. Obs.

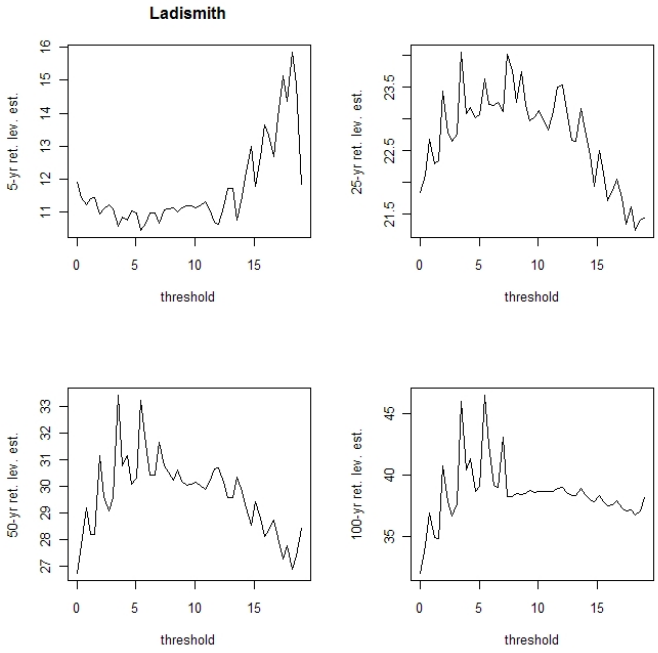


(a) Excelsior

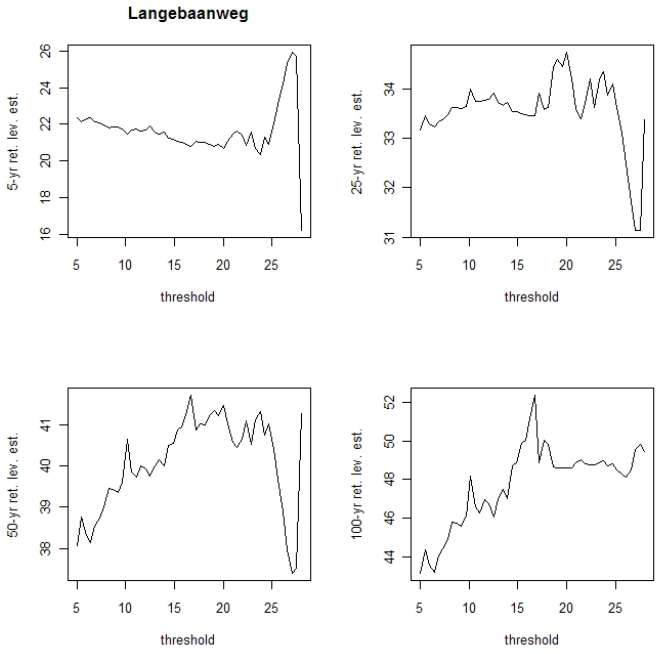


(b) Jonkersberg

Figure A.15: Sensitivity of the return level estimates to the threshold values: Excelsior – Jonkersberg

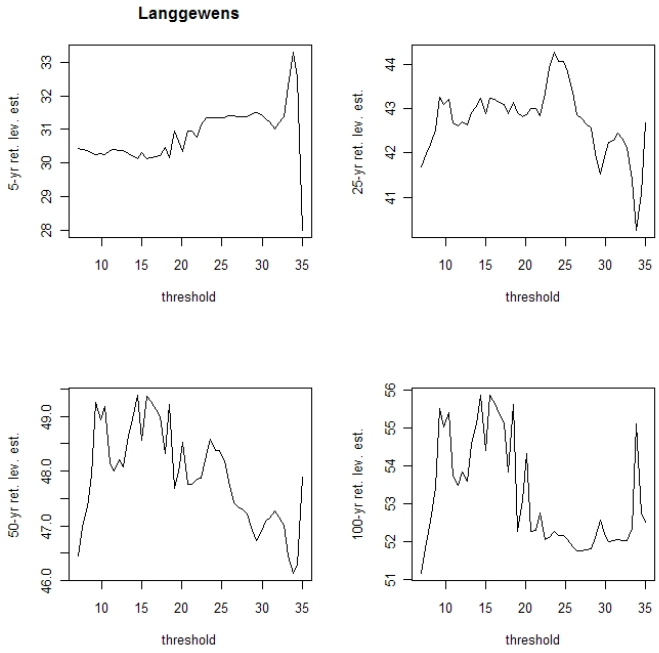


(a) Ladismith

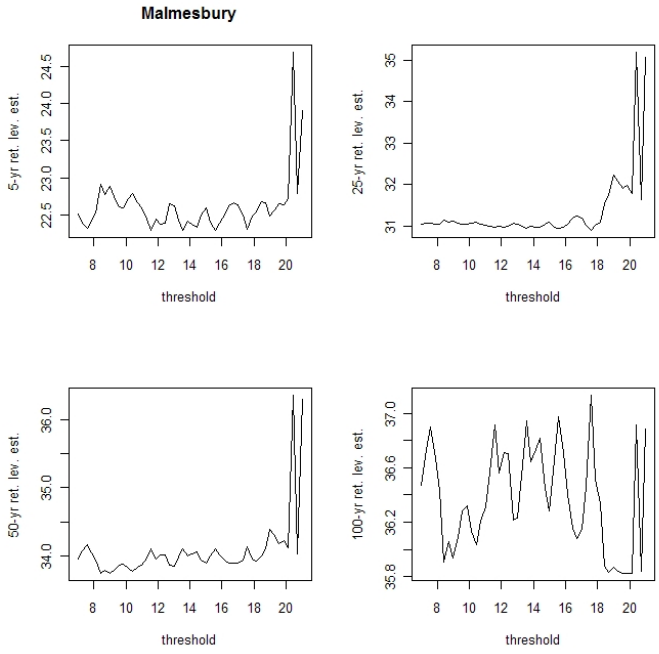


(b) Langebaanweg

Figure A.16: Sensitivity of the return level estimates to the threshold values: Ladismith – Langebaanweg



(a) Langgewens



(b) Malmesbury

Figure A.17: Sensitivity of the return level estimates to the threshold values: Langgewens – Malmesbury

A.2 Ultimate Models

In this section parameter estimates from fitting the point process extreme value model are given. At each station the point process extreme value model was fitted for each of two thresholds chosen from the threshold sensitivity analysis exercise. In cases where high rainfall values were determined to be clustered, declustering of the series was performed. This resulted in the model being fitted three times at these sites. The model parameter estimates for stations where declustering was performed are given as rows where the threshold is in bold text in Table A.1. The threshold is given as u, the number of threshold exceedances as n_u, model deviance as Dev., n_u/n is the threshold exceedance rate, $\hat{\theta}$ is an estimate of the extremal index, $\hat{\mu}$ an estimate of the location for the point process extreme value model and corresponding standard error (s.e.), $\hat{\sigma}$ is an estimate of the scale parameter, $\hat{\xi}$ an estimate of the shape parameter with the corresponding 95% confidence intervals. Lastly, the column $H_0 : \xi \to 0$ gives the results of testing the hypothesis of a zero shape parameter which corresponds to a Gumbel distribution.

Table A.1: Results from fitting point process model

Site	u	n_u	Dev.	n_u/n	$\hat{\theta}$	$\hat{\mu}$(s.e.)	$\hat{\sigma}$(s.e.)	$\hat{\xi}$(95% c.i.)	$(H_0 : \xi \to 0)$
Atlantis	18	83	480.9	0.032		26.23 (1.24)	7.21 (0.72)	-0.09 (-0.30, 0.12)	0.58
	18	74	450.61	0.029	0.86	25.67 (1.30)	7.53 (0.80)	-0.09 (-0.32, 0.13)	0.58
	22	50	331.24	0.019		25.77 (1.17)	6.55 (1.06)	0.03 (-0.32, 0.38)	0.02
CPT Astro.	28	92	711.65	0.02		34.66 (1.48)	11.29 (1.41)	0.17 (-0.05, 0.40)	3.45
	28	82	659.03	0.02	0.93	33.52 (1.50)	11.40 (1.57)	0.19 (-0.06, 0.44)	3.37
	38	38	369.3	0.008		35.36 (1.74)	9.74 (2.87)	0.25 (-0.14, 0.65)	2.56
CPT Int.	22	111	766.16	0.024		29.65 (1.28)	9.52 (0.92)	-0.02 (-0.26, 0.22)	0.03
	26	71	557.6	0.015		29.65 (1.44)	10.63 (1.61)	-0.19 (-0.45, 0.06)	0.09
Excelsior	19	44	296.43	0.032		32.07 (2.74)	11.45 (1.55)	-0.11 (-0.44, 0.23)	0.32
	19	35	259.66	0.025	0.88	31.43 (3.20)	13.34 (1.93)	-0.22 (-0.56, 0.12)	1.12
	29	21	167.11	0.015		32.68 (2.62)	10.74 (3.06)	-0.09 (-0.65, 0.48)	0.08
Jonkersberg	22	71	655.68	0.015		28.37 (2.41)	18.45 (2.69)	0.08 (-0.16, 0.32)	0.54
	22	62	583.73	0.013	0.94	25.45 (2.17)	16.33 (2.90)	0.17 (-0.13, 0.46)	1.75
	32	41	428.07	0.009		28.21 (3.08)	18.96 (4.44)	0.07 (-0.22, 0.36)	0.23
Ladismith	5	61	335.24	0.029		9.75 (1.06)	5.56 (0.93)	0.28 (-0.04, 0.59)	5.90
	5	52	306.85	0.025	0.89	9.04 (1.07)	5.55 (0.99)	0.29 (-0.05, 0.63)	6.45
	7	32	239.78	0.015		9.37 (1.40)	7.34 (1.57)	0.13 (-0.19, 0.44)	0.84
Langebaan	12	92	517.96	0.03		17.83 (0.97)	6.19 (0.74)	0.14 (-0.08, 0.35)	2.14
	12	83	493.74	0.027	0.89	17.54 (1.01)	6.47 (0.80)	0.12 (-0.10, 0.35)	1.57
	17	38	285.96	0.01		17.61 (0.83)	5.04 (1.34)	0.31 (-0.15, 0.77)	2.87
Langgewens	18	95	644.45	0.02		22.9 (1.01)	7.64 (0.84)	0.0003 (-0.23, 0.23)	0.00
	24	38	338.61	0.008		20.92 (1.99)	11.57 (2.67)	-0.21 (-0.45, 0.03)	1.96
Malmesbury	11	89	384.61	0.065		23.86 (1.30)	5.92 (0.63)	-0.21 (-0.42, -0.01)	3.04
	11	69	343.34	0.05	0.79	23.36 (1.49)	6.52 (0.64)	-0.27 (-0.50, -0.04)	3.70
	18	35	205.38	0.025		23.92 (1.56)	6.2 (0.88)	-0.26 (-0.69, 0.16)	1.10
Molteno	32	85	704.81	0.018	0.93	38.81 (1.75)	13.43 (1.77)	0.17 (-0.06, 0.40)	2.94
	40	42	420.82	0.009	0.96	37.61 (2.17)	13.48 (3.62)	0.20 (-0.18, 0.57)	1.58
Paarl	36	106	779.53	0.029		50.42 (2.03)	14.2 (1.27)	-0.08 (-0.25, 0.09)	0.78
	36	89	694.75	0.024	0.9	48.65 (2.19)	15.21 (1.52)	-0.10 (-0.29, 0.09)	0.83
	48	41	380.84	0.011		48.39 (2.52)	13.48 (3.62)	-0.12 (-0.43, 0.20)	0.42
Plettenberg	30	60	564.21	0.013		34.08 (2.89)	21.86 (3.22)	-0.24 (-0.43, -0.06)	4.14
	40	44	414.44	0.01		38.3 (2.04)	13.31 (3.14)	-0.01 (-0.32, 0.31)	0.00
Porteville	25	88	565.19	0.02		30.05 (1.20)	9.21 (1.20)	0.23 (0.02, 0.44)	9.11
	30	51	439.31	0.011		30.3 (1.06)	7.59 (1.57)	0.33 (0.03, 0.64)	10.18
Tygerhoek	18	74	731.89	0.016		26.12 (2.96)	22.30 (3.76)	0.38 (0.07, 0.69)	10.47
	18	63	640.84	0.014	0.94	22.36 (2.64)	19.82 (3.89)	0.43 (0.09, 0.77)	11.76
	38	27	348.59	0.006		16.25 (10.86)	32.99 (15.88)	0.21 (-0.34, 0.77)	0.73
Wellington	22	79	458.08	0.043		35.14 (1.94)	9.95 (1.27)	0.06 (-0.12, 0.25)	0.51
	22	65	415.35	0.035	0.83	34.5 (2.14)	10.84 (1.42)	0.04 (-0.15, 0.24)	0.19
	30	35	256.32	0.019		34.63 (1.74)	8.61 (1.70)	0.16 (-0.14, 0.46)	1.61

Bibliography

Beichelt, F. (2006). *Stochastic Processes in Science, Engineering and Finance*. Chapman and Hall/CRC.

Bierlant, J., Geogebeur, Y., Segers, J., and Teugels, J. (2004). *Statistics of Extremes*. Probability and Statistics. Wiley. ISBN 0-471-97647-4.

Butler, A., Heffernan, J. E., Tawn, J. A., and Flather, R. A. (2007). Trend estimation in extremes of synthetic North Sea surges. *Applied Statistics*, 56(Part 4):395–414.

Cameron, A. C. and Trivedi, P. K. (1998). *Regression analysis of count data*. Economic Society Monographs. Cambridge University Press.

Casson, E. and Coles, S. G. (1999). Spatial regression models for extremes. *Extremes*, 1–4:449–468.

Chavez-Demoulin, V. and Davison, A. C. (2005). Generalized additive modelling of sample extremes. *Applied Statistics*, 54:207–222.

Coles, S. G. (1993). Regional modelling of extreme storms via max-stable processes. *Journal of the Royal Statistical Society*, 55:797–816.

Coles, S. G. (1994). *A Temporal Study of Extreme Rainfall*, volume 2, pages 61–78. Wiley.

Coles, S. G. (2001). *An Introduction to Statistical Modeling of Extreme Values*. Statistics. Springer Verlag, London.

Coles, S. G., Pericchi, L. R., and Sisson, S. (2003). A fully probabilistic approach to rainfall modeling. *Journal of Hydrology*, 273:35–50.

Coles, S. G. and Powell, E. A. (1996). Bayesian methods in extreme value modelling: A review and new developments. *International Statistical Review*, 64(1):119–136.

Coles, S. G. and Tawn, J. A. (1996a). A bayesian analysis of extreme rainfall data. *Applied Statistics*, 45(4):463–478.

Coles, S. G. and Tawn, J. A. (1996b). Modelling extremes of the areal rainfall process. *Journal of the Royal Statistical Society*, 58(2):329–347.

Cooley, D., Nychka, D., and Naveau, P. (2007). Bayesian spatial modelling of extreme precipitation return levels. *Journal of the American Statistical Association*, 102:824–840.

Cox, D. R. and Isham, V. (1980). *Point Processes*. Chapman and Hall, London.

Cressie, N. A. C. (1993). *Statistics for spatial data*. Wiley.

Cunnane, C. (1979). A note on the Poisson assumption in partial duration series model. *Water Resources Research*, 15(2):489–494.

Davison, A. C. (1984). *Statistical Extremes and Applications*, chapter Modelling Excesses Over High Thresholds, with an Application, pages 461–482. Springer.

Davison, A. C. (1986). Approximate predictive likelihood. *Biometrika*, 73:323–332.

Davison, A. C. and Smith, R. L. (1990). Models for exceedances over high thresholds (with discussion). *Journal of the Royal Statistical Society*, 52:393–442.

de Haan, L. (1984). A spectral representation for max-stable processes. *Annals of Probability*, 12:1194–1204.

de Haan, L. and Pickands, J. (1984). A spectral representation for stationary min-stable stochastic processes. *Stochastic Processes and their Applications*, 17:26–27.

de Haan, L. and Pickands, J. (1986). Stationary min-stable stochastic processes. *Probability Theory and Related Fields*, 72:477–492.

de Haan, L. and Resnick, S. (1977). Limit theory for multivariate sample extremes. *Z. Wahr. v. Geb*, 40:317–337.

DPLG (2007). National Disaster Management Centre. inaugural Annual Report 2006/2007. Technical report, Department of Provincial and Local Government – Republic of South Africa.

Draghicescu, D. and Ignaccolo, R. (2009). Modelling threshold exceedance probabilities of spatially correlated time series. *Electronic Journal of Statistics*, 3:149–164.

Engelund, S. and Rackwitz, R. (1992). On predictive distribution functions for the three asymptotic extreme value distributions. *Structural Saftey*, 11:255–258.

Fauchereau, N., Trzaska, S., Rouault, M., and Richard, Y. (2003). Rainfall variability and changes in Southern Africa during the 20^{th} century in the global warming context. *Natural Hazards*, 29:139–154.

Ferro, C. A. and Segers, J. (2003). Inference for clusters of extreme values. *Journal of the Royal statistical Society*, 65:545–556.

Fisher, R. A. and Tippett, L. H. C. (1928). Limiting forms of the frequency distributions of the largest or smallest member of the sample. *Proceedings of the Cambridge Philosophical Society*, 24:180–190.

Gnedenko, B. V. (1943). Sur la distribution limite du terme maximum d'une série aléatoire. *Annals of Mathematics*, 44:423–453.

Goovaerts, P. (2000). Geostatistical approaches for incorporating elevation into the spatial interpolation of rainfall. *Journal of Hydrology*, 228:113–129.

Gumbel, E. J. (1958). *Statistics of Extremes*. Columbia University Press, New York USA.

Hastings, W. K. (1970). Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57:97–109.

Heffernan, J. E. and Tawn, J. A. (2004). A conditional approach for multivariate extreme values. *Journal of the Royal Statistics Society*, 66(Part3):497–546.

Hosking, J. R. M., Wallis, J. R., and Wood, E. F. (1985). Estimation of the generalised extreme value distribution by the method of probability weighted moments. *Technometrics*, 27:251–261.

IPCC (2007). Climate change 2007: Impacts, adaptation and vulnerability. Technical report, Intergovernmental Panel on Climate Change (IPCC) Working Group II 4^{th} Assessment Report, Cambridge UK.

IPCC (2008). Climate change and water. Technical report, Contributors Technical Paper IPCC on Climate Change VI, Geneva.

Isaaks, E. H. and Srivastava, R. M. (1989). *Applied geostatistics*. Oxford University Press.

Katz, R. W. and Naveau, P. (2010). Editorial: Special issue on statistics of extremes in weather and climate. *Extremes*, 13:107–108.

Katz, R. W., Parlange, M. B., and Naveau, P. (2002). Statistics of extremes in hydrology. *Advances in Water Resources*, 25:1287–1304.

Kruger, A. C. (2006). Observed trends in daily precipitation indices in South Africa: 1910-2004. *International Journal of Climatology*, 26:2275–2285.

Kruger, A. C. (2007). Climate of South Africa. precipitation. ws47. Technical report, South African Weather Services, Pretoria, South Africa.

Leadbetter, M. R. (1983). Extremes and local dependence in stationary sequences. *Probability Theory and Related Fields*, 65(2):291–306.

Leadbetter, M. R., Lindgren, G., and Rootzén, H. (1983). *Extremes and Related Properties of Random Sequences and Series*. Springer Verlag, New York.

Leadbetter, M. R. and Rootzén, H. (1988). Extremal theory for stochastic processes. *Annals of Probability*, 16:431–478.

Ledford, A. and Tawn, J. A. (1996). Statistics for near independence in multivariate extreme values. *Biometrika*, 83:169–187.

Ledford, A. and Tawn, J. A. (1998). Concomitant tail behaviour for extremes. *Advances in Applied Probability*, 30:197–215.

Leichenko, R. M. and O'Brien, K. L. (2002). The dynamics of rural vulnerability to global change: The case of southern Africa. *Mitigation and Adaptation Strategies for Global Change*, 7:1–18.

Mason, S. J. and Joubert, A. M. (1997). Simulated changes in extreme rainfall over southern Africa. *International Journal of Climatology*, 17:291–301.

Mason, S. J., Waylen, P. R., Mimmack, G. M., Rajaratnam, B., and Harrison, M. (1999). Changes in extreme rainfall events in South Africa. *Climatic Change*, 41:249–257.

Méndez, F. J., Menéndez, M., Luceño, A., Medina, R., and Graham, N. E. (2008). Seasonality and duration in extreme value distributions of significant wave height. *Ocean Engineering*, 35(1):131–138.

Miller, H. J. (2004). Tobler's First Law and spatial analysis. *Annals of the Association of American Geographers*, 94(2).

Morton, I. D., Bowers, J., and Mould, G. (1997). Estimating return period wave heights and wind speeds using a seasonal point process model. *Coastal Engineering*, 31:305–326.

Mukhopadhyay, N. (2000). *Probability and Statistical Inference*. Marcel Dekker.

NERC (1975). The flood studies report. Technical report, The Natural Environment Research Council, London.

New, M., Hewitson, B., Stephenson, D. B., Tsiga, A., Kriger, A., Manhique, A., Gomez, B., Coelho, S. A. S., Masisi, D. N., Kululanga, E., Mbambalala, E., Adesina, F., Saleh, H., Kanyanga, J., Adosi, J., Bulane, L., Fortunata, L., Mdoka, M. L., and Lajoie, R. (2006). Evidence of trends in daily climate extremes over southern and west Africa. *Journal of Geophysical Research*, 111:1–11.

Pickands, J. (1971). The two-dimensional Poisson process and extremal processes. *Journal of Applied Probability*, 8:745–756.

Prescott, P. and Walden, A. T. (1980). Maximum likelihood estimation of the parameters of the generalised extreme value distribution. *Biometrika*, 67:723–724.

Preston-Whyte, R. A. and Tyson, P. D. (1988). *The atmosphere and weather of southern Africa*. Oxford University Press.

Prudhomme, C. (1999). Mapping a statistic of extreme rainfall in a mountainous region. *Physics and Chemistry of the Earth*, 24(1–2):79–84.

Prudhomme, C. and Reed, D. W. (1999). Mapping extreme rainfall in a mountainous region using geostatistical techniques: A case study in Scotland. *International Journal of Climatology*, 19:1337–1356.

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Reason, C. J. C. (1998). Warm and cold events in the Southeast Atlantic/Southwest Indian Ocean region and potential impacts on circulation and rainfall over southern Afica. *Meteorology and Atmospheric Physics*, 69:49–65.

Reason, C. J. C., Engelbrecht, F., Landman, W. A., Lutjeharms, J. R. E., Piketh, S., Rautenbach, C. J. d. W., and Hewitson, B. C. (2006). A reiew of South African research in atmospheric science and physical oceanography during 2000 – 2005. *South African Journal of Science*, 102:35–45.

Reiss, R. D. and Thomas, M. (2007). *Statistical Analysis of Extreme Values with Applications to Insurance, Finance, Hydrology and Other Fields*. Birkhäuser Basel, fourth edition.

Ribatet, M. (2007). *POT - Generalized Pareto Distribution and Peaks Over Threshold*. R package version 1.0-4.

ROA, I. (2007). Science plan on natural and human-induced hazards and disasters in sub-Saharan Africa. Technical report, International Council for Science Regional Office for Africa (ICSU ROA).

Schulze, G. C. (2007). Atmospheric observations and numerical weather prediction. *South African Journal of Science*, 103:318–323.

Shongwe, M. E., van Oldenborgh, G. J., van den Hurk, B. J. J. M., de Boer B., Coelho, C. A. S., and van Aalst, M. K. (2009). Projected changes in mean and extreme precipitation in Africa under global warming Part I: Southern Africa. *American Meteorological Society*, pages 3819–3837.

Sisson, S. A., Pericchi, L. R., and Coles, S. G. (2006). A case for a reassessment of the risks of extreme hydrological hazards in the caribbean. *Stochastic Environmental Resource Risk Assessment*, 20:296–306.

Smith, A. F. M. and Roberts, G. O. (1993). Bayesian computation via the gibbs sampler and related markov chain monte carlo methods. *Journal of the Royal Statistical Society*, 55:3–23.

Smith, R. L. (1984). *Statistical Extremes and Applications*, chapter Threshold Methods for Sample Extremes, pages 621–638. Springer.

Smith, R. L. (1989a). Extreme value analysis of environmental time series: an application to trend detection in ground-level ozone. *Statistical Science*, 4:367–393.

Smith, R. L. (1989b). A survey of nonregular problems. In *Proceedings of the 47th session of the I.S.I.*, pages 353–372.

Smith, R. L. (1991). Spatial extremes and max-stable processes. Technical report, Department of Statistics University of North Carolina Chapel Hill NC 27599-3260 USA.

Smith, R. L. (2003). Statistics of extremes with applications in environment, insurance and finance. Technical report, Department of Statistics University of North Carolina Chapel Hill NC 27599–3260 USA.

Smith, R. L. and Shively, T. S. (1995). A point process approach to modeling trends in tropospheric ozone. *Atmospheric Environment*, 29:3489–3499.

Smith, R. L., Tawn, J. A., and Yuen, H. K. (1990). Statistics of multivariate extremes. *International Statistical Review*, 58(1):47–58.

Smith, R. L. and Weissman, I. (1994). Estimating the extremal index. *Journal of the Royal Statistical Society*, 56:515–528.

Smithers, J. and Schulze, R. (2000). Development and evaluation of techniques for estimating short duration design rainfall in South Africa. Technical Report 681/1/00, Water Research Commission, Pretoria, RSA.

Smithers, J. and Schulze, R. (2002). Design rainfall and flood estimation in South Africa. Technical Report K5/1060, Water Research Commission, Pretoria, RSA.

StatsSA (2009). Census of commercial agriculture 2007. Preliminary report, p1102, Statistics South Africa.

StatsSA (2010a). General household survey 2009. P0318, Statistics South Africa.

StatsSA (2010b). Gross Domestic Product – First quarter 2010. P0441, Statistics South Africa.

Stein, A. and Sterk, G. (1999). Modeling space and time dependence in environmental studies. *JAG*, 1(2):109–121.

Stephenson, A. and Gilleland, E. (2006). Software for the analysis of extreme events: The current state and future directions. *Extremes*, 8:87–109. DOI 10.1007/s10687-006-7962-0.

Sterk, G. and Stein, A. (1997). Mapping wind-blown mass transport by modeling variability in space and time. *Soil Science Society of America Journal*, 61:232–239.

Sterk, G., Stein, A., and Stroosnijder, L. (2004). Wind effects on spatial variability in pearl millet yields in the Sahel. *Soil and Tillage Research*, 76:25–37.

Szolgay, J., Parajka, J., Kohnová, S., and Hlavčová, K. (2009). Comparison of mapping approaches of design annual maximum daily precipitation. *Atmospheric Research*, 92:289–307.

Williams, C. J. R., Kniveton, D. R., and Layberry, R. (2007). Climatic and oceanic associations with daily rainfall extremes over southern Africa. *International Journal of Climatology*, 27:93–108.

Yee, T. and Wild, C. J. (1996). Vector generalized additive models. *Journal of the Royal Statistical Society*, 58:481–493.

Zhang, Q., Xu, C., Chen, Y. D., and Liu, C. (2009). Extreme value analysis of annual maximum water levels in the Pearl River Delta China. *Frontiers of Earth Sciences in China*, 3(2):154–163.