A COMPREHENSIVE ANALYSIS OF EXTREME RAINFALL

Paulo A. Kagoda

A research report submitted to the Faculty of Engineering and the Built Environment, University of the Witwatersrand, in partial fulfilment of the requirements for the degree of Master of Science in Engineering.

October 2006

DECLARATION

I declare that this research report is my own, unaided work. It is submitted for the Degree of Master of Science in Engineering at the University of the Witwatersrand, Johannesburg. It has not been submitted before for any degree or examination in any other University.

_____

(Signature of candidate)

_____day of _____ (year) _____

To my parents:

Alice Merab and Stephen Paul Kagoda

ABSTRACT

Every now and again, and more so recently, natural disasters occur at magnitudes which standard statistical procedures for forecasting extreme processes failed to predict. While it is easy to blame the failure of the statistical analyses to predict the scale of these natural disasters on changing global trends or some sort of external intervention, this study takes the view that standard statistical methods that do not take comprehensive account of the uncertainties involved in record taking, model selection and predictions, will always result in future estimates that are more likely to be contradicted by observed hydrological events.

Based on daily rainfall data for fifteen rain gauge stations selected from across South Africa, the Bayesian approach to data analysis was followed. The Bayesian approach enables uncertainties to be taken into account comprehensively. The Generalized Pareto Distribution (GPD) was used to model excesses over a threshold that had been chosen from a mean residual life plot of the rainfall data. The shape and scale parameters of the Generalized Pareto Distribution (GPD) were represented by a joint distribution which was sequentially modified by the rainfall data resulting in a posterior distribution from which a Markov chain was generated using the Gibbs Sampler. It is this output of the Gibbs sampler that was used to obtain estimates of return levels (rainfall magnitudes) at various return periods.

For the shorter return periods, the estimates of rainfall magnitudes obtained by the Bayesian approach are reasonably similar to the estimates obtained by the regional index storm methodology. For the longer return periods, the Bayesian estimates were significantly larger with an average percentage increase of 63.2 % for estimates at 100-year return period and 87.5 % increase for estimates at 200-year return period. The difference in estimates can be attributed to the Bayesian methodology's recognition and consequent incorporation of the uncertainty in the analysis. The use of these Bayesian estimates in engineering design would result in reductions in the risk of failure of engineering structures, damage to property and loss of lives that are associated with extreme events.

AKNOWLEDGEMENTS

I wish to sincerely acknowledge with gratitude the invaluable guidance, support and inspiration of my supervisor Dr. J. G. Ndiritu through out the duration of my studies.

I am also thankful to the Water Research Commission of South Africa for making available reports as and when they were required for this research.

The most sincere gratitude goes to my dear parents for always being there through it all as well as to the rest of the family for their distant but consistent support and love.

I also wish to acknowledge the support, both intellectually and morally, of my colleagues and friends within the School notably Nako Sebusang, Mthokozisi Ncube, Zacharia Katambara, Thabo Rasiuba, Jean-Marie Kileshye Onema and Jean-Marc Mwenge Kahinda.

Most importantly, I wish to acknowledge the divine hand of my Lord and Saviour, Jesus Christ, for carrying me this far. Faithful is that Hand!

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1 : INTRODUCTION

1.1 The need for Comprehensive Rainfall Frequency Analysis

Every now and again natural disasters occur such as extensive floods whose magnitudes exceed the design flood magnitudes causing failure of engineering structures such as bridges, dams, etc. This does not necessarily imply that failure of the structures occurs only when the magnitude of the flood exceeds the design flood. In fact, floods of a magnitude far less than the design flood when combined with other unforeseen conditions such as undetected foundation weakness have been known to result in applied loads exceeding the resistance to structural failure resulting in failure (Alexander, 2001). Structural failure, however, regardless of the cause, often results in engineering design standards getting questioned.

The accepted approach to safety in engineering design is such that the maximum load that may be applied to a structural element at sometime in the future is less than the resistance to failure and in this ideal case the probability of failure can be said to be zero. In practice, the resistance of a structural element to failure is not known exactly and consequently, there is always a risk that the resistance to failure will be less than the applied load at some time in the future. In practice, a probability density function is used to describe the variation of the properties of a structural member. This has lead to the development of engineering design standards which ensure that the structural elements are designed in such a manner as to have low probabilities of failure.

Similarly, for design rainfall and design flood determination, engineers have had to make do with statistical analysis. Essentially, the magnitude of rainfall or flood for a specific return period is derived from historical observations. Bearing in mind that the historical records are very unlikely to be repeated in the future, the statistical properties of the past records are examined and from these, estimates of the likelihood of events of a given severity occurring are made. These methodologies that make use of statistical analysis in engineering design have provided reasonable predictions despite the mistrust of statistics in the past (Alexander, 2001). This mistrust, rather than being a weakness on the

part of statistical analysis as a tool, can be seen as a failure on the part of the analysts to fully appreciate the weaknesses as well as the power of the basic statistical methodology.

Nevertheless, frequently, and more so recently, there have been natural disasters at magnitudes which standard statistical procedures for forecasting extreme processes failed to predict. For instance, in December 1999, Maiquetia station in Venezuela recorded a daily rainfall value of 410.4 mm; a value whose return period according to maximum likelihood parameter estimation methodology using a Gumbel distribution attached to pre-1999 data, was 17,600,000 years (Coles et al, 2003). More often, the failure of the statistical analyses to predict the scale of these natural disasters has been explained by the changing global trends or some sort of external intervention. Coles et al (2003), however, dispute the strength of this argument as an explanation for all prediction failures and argue that the explanation for the failure, in part, lies in the fact that the standard statistical methodology does not take into account the uncertainties involved in parameter estimation of the statistical models. For instance, values recorded during an extreme event have to be taken with caution since it is during such events that significant measurement errors arise. For example, a rain gauge could fill-up and overflow during a heavy storm while a stream gauge could be rendered useless by a flood. Since such uncertainties can be significant, analysis and prediction should allow for these effects. Subsequently, using the pre-1999 data, Coles et al (2003) were able to show using Bayesian inference techniques that the return period of the rainfall value that was recorded in Venezuela in December 1999 was 260 years when the Generalized Pareto distribution was used to model exceedances of the rainfall over the 10mm threshold that was selected based on the mean residual life plot of the rainfall data.

Consequently, Bayesian inference, which has been shown to offer a more coherent framework for keeping track of, and incorporating, all uncertainties involved in the prediction process, is increasingly being used in preference over the standard methodology. This is further helped by the fact that advances in the Monte Carlo Markov techniques provide a simpler alternative to the complex analytical calculations that once made the Bayesian analysis difficult

(Coles et al, 2003). For instance, using Bayesian inference, Coles and Tawn (1996) elicited prior information for extreme rainfall from an expert who had specific knowledge of their study area in South-west England which they combined with rainfall data, to make estimates of extreme rainfall behavior. Estimates were then made using the Bayesian approach as well as the maximum likelihood approach of the design level (mm) that would be exceeded in a given year with probability of 0.001. The Bayesian estimate was 248 mm while the maximum likelihood estimate was 145 mm (Coles and Tawn, 1996).

South Africa, like any other part of the world, is not exempt from floods. Floods can and have occurred in the past and caused a lot of damage to property, claimed the lives of some people and left many others homeless (Grobler, 1996). Tables 1-1a and 1-1b show a record of some of the major floods recorded in South Africa from 1970 as well as the details of the damage to property and the lives claimed by each of these floods, where that information is available. Although lists of historical floods, such as this, cannot be used for accurate estimates of the frequency of floods within a region, they can provide useful background information (Alexander, 2001).

Numerous regional- and national-scale studies on estimation of design rainfall in South Africa have been done over the years. The majority of them have focused on estimation of rainfall with durations of less than 24 hours (Smithers and Schulze, 2003). Some of these studies include the works of Midgley and Pitman (1978), Smithers and Schulze (2000a) and Alexander (2001). There are, however, a few others that have estimated design rainfall for durations of one day or longer; the latest being the work of Smithers and Schulze (2000b). With the exception of the work by Smithers and Schulze (2000a, 2000b, 2003), the other studies have all utilized point design rainfall values using at-site data only. Smithers and Schulze (2003) performed a regionalization in an attempt to increase the reliability of the design values at gauged sites and for estimation of design values at ungauged sites. It should be noted that all the previous studies on design rainfall estimation have not offered a comprehensive account of the uncertainties involved in estimation of the parameters of the models used. It is the intention of this study, therefore, to offer a more comprehensive account of the uncertainties involved in parameter estimation.

Table 1.1a: Major Floods in South Africa since 1970

Date	Region	Details
[1]August 1970	East London	447 mm in six hours
[*]4th March 1974	Central Interior	Millions of rand damage along Modder and Riet rivers. At Upington, the Orange river flooded 80 % of the houses on the island and along the river. Fish River Valley experienced the worst flood in 120 years. In Cradock, 200 homes were inundated.
[1]February 1975	Vaal – Orange River system	
[*]9th February 1977	North Eastern Regions	Widespread flooding. Ten people drowned at Tshipise near the Kruger National Park.
[*]28th January 1978	Pretoria	245 mm in 4 hours. Homes, factories and flats flooded. 11 people dead.
[*]25th January 1981	Laingsburg	104 people drowned, 185 homes, old age home and 23 offices destroyed.
[*]31st January 1984	North Eastern Regions	Cyclone Domoina killed more than 200 people in Swaziland, Mozambique, Eastern Transvaal and North-Eastern Natal. Damage to Sugar Cane fields estimated at R 470m. Damage to bridges estimated at R 25m. Highest daily rainfall amount ever recorded in South Africa (597 mm at St Lucia lake).
[*]28th September 1987	Natal	Homes washed away, collapsed or buried in mud. Thousands of kilometers of roads damaged. 14 bridges washed away, all entrance routes to Durban closed. R 3,300m damage, 388 deaths and 68,000 left homeless.

[*] Source: Grobler (1996)
[1] Source: W. J. R Alexander (2001)

Table 1.1b: Major floods in South Africa since 1970

Date	Region	Details
[*]29th February 1988	Central Interior	In free state, 47 bridges destroyed, 1,300 homes evacuated in Northern Cape. 30 magisterial districts declared disaster areas.
[*]2nd February 1994	Ladysmith	R 60m flood damage. More than a thousand left homeless.
[1]December 1997	Edendale	
[1]Feb – Mar 2000	Country wide	
August 2001	Cape flats	10 000 people displaced by flood waters.
15th -19th August 2002	East London	317.2 mm of rainfall recorded. 21.0 mm of which fell between 9 and 12 in the evening of 15 August
November 2003	Durban	Heavy rains caused Palmiet river to bust it banks. 300 occupants of an informal settlement were relocated. In one storm event 40.8 mm of rainfall was recorded.
December 2004 – January 2005	Southern Cape areas	R 25m flood damage. After facing the worst drought conditions in 25 years, pattern changed with heavy rainstorms starting in December 2004.
April 2005	Western Cape	In Cape Aghulas Municipality, 800 houses and 3000 people submerged under 1.5 metres of water one morning.
August 2006	Eastern Cape	R 80 m flood damage. 4 people lost their lives. 7000 people relocated. Bridges destroyed.

[*] Source: Grobler (1996)
[1] Source: W. J. R Alexander (2001)

1.2 Objectives

The objective of this study is to study what impact the comprehensive incorporation of parameter uncertainties would have on design rainfall estimation in South Africa. This will be done by:-

- Use of the Bayesian approach to analyze rainfall records so as to come up with their expected probability distribution. The expected probability distribution will enable the estimation of exceedance probabilities corresponding to extreme rainfall events.

- Comparison of the estimates obtained from using the Bayesian approach to estimates obtained using a statistical procedure used by Smithers and Schulze (2003). The procedure that Smithers and Schulze employed utilizes L-moments for design rainfall estimation.

1.3 The Study Area

The study was carried out on data from 15 rainfall stations taken from across South Africa. The selection of stations was done in such a way that one station was selected from each of the fifteen clusters of relatively homogenous short-duration ($\leq 24\ hrs$) rainfall that the study by Smithers and Schulze (2003) came up with.

1.4 The Expected Outcomes

This study used Bayesian inference, which has been shown to offer a more coherent framework for accounting for uncertainties involved in the parameter estimation process, to make estimates of return periods corresponding to extreme rainfall events. This will involve the derivation and use of appropriate distributions for the respective parameters as opposed to standard methodologies that rely on point estimates for analysis. The account for uncertainties in parameter estimation will result in Bayesian estimates of rainfall magnitudes that are more dependable than the corresponding estimates of rainfall magnitudes obtained using standard methodology for any return period.

CHAPTER 2 : A BACKGROUND TO MONTE CARLO BAYESIAN INFERENCE

2.1 Introduction

Extreme value modeling of environmental processes is standard practice for the design of many large scale construction projects. Standard methodology for modeling extremes consists of adopting an asymptotic model to describe stochastic variation at extreme levels of a process (Coles et al, 2003). Inference and forecasting of the environmental process will then be done on the basis of the adopted asymptotic model. This methodology, however, overlooks model and prediction uncertainty leading to underestimation of the probability of extreme events (Coles et al, 2003). For this reason, Bayesian inference, which has seen only limited application in practice because it is poorly understood and more significantly because of the typical numerical intensity its implementation requires, is beginning to gain preference over the standard methodology (Kuczera, 1999, Coles et al, 2003). This has been made possible by the continuing development in the Markov Chain Monte Carlo approximation techniques that have put to rest concerns about the computational aspects of Bayesian inference (Coles and Tawn, 1996) as well as the wide spread use of computers in hydrology (Kuczera, 1999). Bayesian analysis has found applications in several disciplines ranging from hydrology to ecology to the insurance industry and geophysical sciences (Coles et al, 2003). There are several studies in hydrology where Bayesian analysis has been employed. These include the study by Coles et al (2003) where it was shown using the Bayesian analysis that the daily rainfall amount of 410.4 mm recorded in December 1999 in Venezuela had a return period that was not as long as the standard statistical methods of analysis showed (177 years using the Generalized Extreme Value distribution with the Bayesian approach as opposed to a 737 000 year value when the same data was analyzed using the maximum likelihood estimator method). Coles and Tawn (1996) obtained a 95 % Bayesian interval estimate of the 100 year return period for daily rainfall taken from an area in the south-west of England that was half of the width of the corresponding likelihood-based confidence interval. Moyeed and Clarke (2005) used Bayesian analysis to fit rating curves using discharge and stage data taken from the Amazon at Obidos and the Parana at Corrientes. Kuzcera (1999) developed the FLIKE

software to compute the expected probability distribution as well as the quantile confidence limits for any flood frequency distribution using gauged flow data based on Bayesian inference and Monte Carlo simulation techniques.

A brief introduction to the probability theory of extremes is done in section 2.2. Bayesian inference is discussed in section 2.3 and this is followed by an introduction to the Markov Chain Monte Carlo methodologies in section 2.4. Section 2.5 covers the standard statistical methods of analyzing extreme-value data.

2.2 Probability Theory of Extremes

Consider a sequence of independent, continuous random variables in a stationary and randomly varying context $\{Y_1, Y_2, \ldots\}$. Suppose the point of interest is $M_n = max\{Y_1, \ldots, Y_n\}$. If $F(y) = P(Y_i \leq y)$ is the cumulative distribution function of the individual Y_i, it follows from the laws of probability that the distribution function $G_n(y)$ of M_n is $F^n(y)$ (where $F^n(y) \equiv [F(y)]^n$) (Cox et al, 2002). As $n \to \infty$, $G^n(y)$ takes on a limiting form i.e there exists a limiting shape characteristic of the distribution of extremes. Since for all m, the maximum of nm values can be regarded as the maximum of m individual maxima each of n values, any limiting form $H(y)$ must satisfy (Cox et al, 2002)

$$H^m(y) = H(c_m y + d_m) \qquad 2.1$$

for suitable constants c_m, and d_m.

There are only three possible solutions to (2.1) (Cox et al, 2002). These are

$H_1(y) = \exp(-e^{-y})$ for $-\infty < y < \infty$ Type I

$H_2(y) = \exp(-y^{-\alpha})$ for $y \geq 0$, $H_2(y) = 0$ otherwise, and Type II

$H_3(y) = \exp(-(-y)^{\alpha})$ for $y \leq 0$, $H_3(y) = 1$ otherwise, Type III

where the constant α is positive. These distributions are called ***max-stable***. Type I usually called the Gumbel distribution, holds essentially when the underlying $F(y)$ approaches its limiting value at least exponentially fast. Type II, usually called the Frechet distribution, holds essentially when $F(y)$ approaches its limit at a power law rate. Type III, usually called the Weibull distribution, applies when there is a clear upper bound to the values of Yi (Cox et al, 2002).

The three limiting forms are linked by simple transformations. For example, if Y is a type II. variable, then $ln(Y)$ is type I.; the corresponding transformation for a type III. variable is $-ln(-Y)$. Also, if Y is type I, then $exp(-Y)$ is exponentially distributed. The three extreme-value distributions can all be subsumed in the generalized extreme-value (GEV) distribution which has a distribution function usually parameterized in the form (Cox et al, 2002).

$$G_{EV}(y) = \begin{cases} \exp\left\{-(1+ky)^{-1/k}\right\} & for \quad k \neq 0, \\ \exp\{-\exp(-y)\} & for \quad k = 0, \end{cases}$$

2.2

where $(1 + ky) > 0$. The parameter k determines the shape of the distribution, while arbitrary location and scale parameters can be incorporated by replacing y by $(y- \mu)/\alpha$, where $\alpha > 0$.

Consider the originating sequence $Y_1, \ldots, Y_n$. It can be summarized in a more informative derived aspect, namely the number and magnitude of peaks over a threshold instead of its maximum. For this, a large threshold, u_n, is chosen in principle increasing with n, although in practice likely to be chosen in the light of the data and subject to a sensitivity analysis (Leadbetter, 1991). Each Yi has a small probability (i.e $1 - F(u_n)$) of exceeding the threshold, independently of other Yi's. Thus, in the limit as $n \to \infty$ and assuming that the mean number of exceedances $n\{1 - F(u_n)\}$ tends to a limit ν, these exceedances form a Poisson process. In particular, the probability that there are no exceedances of the threshold tends to $\exp(-\nu)$. Given that there is an exceedance, the size, $Z = Y$

$- u_n$, of the excess is governed by the conditional distribution function $F(z + u_n)/\{1 - F(u_n)\}$. In the limiting form, this gives the generalized Pareto distribution (GPD) (Cox et al, 2002).

$$G_P(z) = \begin{cases} 1-(1+kz)^{-1/k} & when \quad k \neq o, \\ 1-\exp(-z) & if \quad k = 0 \end{cases}$$

2.3

where $z \geq 0$ when $k \geq 0$ and $0 \leq z \leq -1/k$ when $k < 0$. As before the location and scale can be changed by replacing z appropriately. Note that the conditions under which GPD arises as a limiting distribution of excesses are exactly those for which the GEV distribution arises as the limiting distribution of the maxima; the parameter k is the same in both cases (Cox et al, 2002).

Supposing now that $\{Y_i\}$ is a stationary sequence of dependent variables, whose marginal distribution is such that, if they were independent, a limiting distribution for the maxima, M_n, would exist. Then, as long as the dependence in the sequence is rather limited, the extreme-value results for the limiting distribution of M_n still holds, the asymptotic rescaling constants being exactly those that would apply in the independent case. Essentially, two conditions should be satisfied. The first is a mixing condition that ensures that, if the sequence is divided into non-overlapping sections, then the dependence between the section maxima is limited and goes to zero as the distance between the sections increases (Cox et al, 2002).

The second condition ensures that the chance of multiple exceedances within a section of the sequence is asymptotically negligible. These two conditions also guarantee that the results for the point process of exceedances over thresholds, and the distribution of excesses, are as for the independent case. The mixing condition gives the asymptotic independence needed for the limiting point process of exceedances to be a Poisson process, while the second condition guarantees that there are no multiple occurrences in this process (Cox et al, 2002).

2.2.1 Extremes in Non-stationary processes.

Climatic change is one of several factors responsible for the non-stationary characteristic of rainfall data. The effects of climatic change can be detected by the occurrence of a trend in rainfall data of a particular location and the effect of these trends is that they lead to higher exceedance probabilities and as such disregard of the effects of climatic change will lead to analysis with over-optimistic results. Climatic change can be thought of as a monotonic trend that has to be detected against a background of deterministic changes (Cox et al, 2002).

2.3 Bayesian Inference

Suppose that $y' = (y_1, \ldots y_n)$ is a vector of n observations whose probability distribution $p(y|\Theta)$ depends on the values of m parameters $\Theta' = (\Theta_1, \ldots, \Theta_m)$. Suppose also that Θ itself has a probability distribution $p(\Theta)$. Then according to Box and Tiao (1973),

$$p(y|\theta)p(\theta) = p(y,\theta) = p(\theta|y)p(y). \qquad 2.4$$

Given the observed data y, the conditional distribution of Θ is

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)} \qquad 2.5$$

Also, we can write

$$p(y) = E\left[p(y|\theta)\right] = c^{-1} = \begin{cases} \int p(y|\theta)p(\theta)d\theta & \theta \quad continuous \\ \sum p(y|\theta)p(\theta) & \theta \quad discrete \end{cases} \qquad 2.6$$

where the sum or the integral is taken over the admissible range of Θ, (Box and Tiao, 1973) and where $E[f(\Theta)]$ is the mathematical expectation of $f(\Theta)$ with respect to the distribution $p(\Theta)$. Thus alternatively we may write (2.5) *as*

$$p(\theta|y) = cp(y|\theta)p(\theta)$$

2.7

where c is a "normalizing" constant.

This is usually referred to as Bayes' theorem. In this expression, $p(\Theta)$, which tells us what is known about Θ without knowledge of the data, is called the **prior distribution** of Θ. Correspondingly, $p(\Theta|y)$, which tells us what is known about Θ given knowledge of the data, is called the **posterior distribution** of Θ given y (Box and Tiao, 1973). The quantity c is necessary to ensure that the posterior distribution $p(\Theta|y)$ integrates or sums to one.

The methodology provides for the beliefs that are held out of 'experience' i.e the prior, to be modified by the available data resulting in a posterior distribution. The use of the observed data (rainfall records) for the modification can be taken as a means of dealing with the uncertainty arising from the choice of the prior. The second instance in the methodology that enables for the uncertainty to be accounted for is the use of a distribution of the parameters as an input to the predictive distribution as opposed to the standard methodology where a point estimate of each parameter would be used. Given the two instances in which parameter uncertainty is accounted for in the Bayesian methodology, it is reasonable to believe that the estimates of rainfall magnitudes for various return periods obtained by this method will be more reliable than estimates obtained using standard statistical analytical methods.

2.3.1 Bayes' Theorem and the Likelihood Function

Given the data y, $p(y|\Theta)$ in (2.7) may be regarded as a function not of y but of Θ. When so regarded, it is called the likelihood function of Θ for a given y and can be written as $l(\Theta|y)$ (Box and Tiao, 1973). Bayes' formula can thus be written as

$$p(\theta|y) = l(\theta|y)\,p(\theta)$$

2.8

In other words, the probability distribution for Θ posterior to the data y is proportional to the product of the distribution for Θ prior to the data and the likelihood for Θ given y. That is,

$$Posterior \propto Likelihood \times \Pr ior$$

2.9

The likelihood function $l(\Theta|y)$ is the function through which the data y modifies prior knowledge of Θ ; it can therefore be regarded as representing the information about Θ coming from the data (Box and Tiao, 1973).

The likelihood function is defined up to a multiplicative constant, that is, multiplication by a constant leaves the likelihood unchanged. This is in accord with the role it plays in Bayes' formula, since multiplying the likelihood function by an arbitrary constant will have no effect on the posterior distribution of Θ. The constant will cancel upon normalizing the product on the right hand side of (2.8) (Box and Tiao, 1973).

The example in *APPENDIX A* illustrates the concepts prior distribution, likelihood function and the posterior distribution clearly. It also demonstrates the power of Bayesian inference where after a sample of three parts (one of which is found to be defective), the certainty with which one can say there are only two defective parts in the shipment increases to 42 % which is four times what the belief, based on past experience only, was before samples were drawn from the box.

2.3.2 The Standardized Likelihood

When the integral $\int l(\theta|y)d\theta,$ taken over the admissible range of Θ, is finite, then occasionally it will be convenient to refer to the quantity

$$\frac{l(\theta|y)}{\int l(\theta|y)d\theta}$$

2.10

This is the standardized likelihood, that is, the likelihood scaled so that the area, volume, or hypervolume under the curve, surface or hypersurface is one (Box and Tiao, 1973).

2.3.3 Sequential Nature of Bayes' Theorem.

The theorem in (2.8) provides a mathematical formulation of how previous knowledge may be combined with new knowledge.

Suppose an initial sample of observations y_1 is obtained, then Bayes' formula gives

$$p(\theta|y_1) \propto p(\theta)l(\theta|y_1) \qquad 2.11$$

Now suppose a second sample of observations y_2 distributed independently of the first sample is obtained then

$$\begin{aligned} p(\theta|y_2, y_1) &\propto p(\theta)l(\theta|y_1)l(\theta|y_2) \\ &\propto p(\theta|y_1)l(\theta|y_2) \end{aligned} \qquad 2.12$$

Equation (2.12) is precisely of the same form as (2.11) except that $p(\theta|y_1)$, the posterior distribution for Θ given y_1, plays the role of the prior distribution for the second sample (Box and Tiao, 1973). Obviously this process can be repeated any number of times. In particular, for n independent observations, the posterior distribution can, if desired, be recalculated after each new observation so that at the m^{th} stage, the likelihood associated with the m^{th} observation is combined with the posterior distribution of Θ after $m - 1$ observations to give the new posterior distribution

$$p(\theta|y_1,\ldots, y_m) \propto p(\theta|y_1,\ldots, y_{m-1})l(\theta|y_m), \qquad m = 2, \ldots, n \qquad 2.13$$

where

$$p(\theta|y_1) \propto p(\theta)l(\theta|y_1).$$

Thus, Bayes' theorem describes, in a fundamental way, the process of learning from experience and shows how knowledge about the state of nature represented by Θ is continually modified as new data becomes available (Box and Tiao, 1973). Example 2 in *APPENDIX B* illustrates this. The prior probability (certainty) of the 'test mouse being either homozygous (BB) or heterozygous (Bb) are $\frac{1}{3}$ and $\frac{2}{3}$ respectively. Supposing the 7 black offspring are viewed as a sequence of seven independent observations, then, the certainty with which

one can state that the 'test' mouse is homozygous (BB) increases for each black offspring as illustrated in Table B.2.

2.4 Markov Chain Monte Carlo Methods

The proportionality factor in equations 2.9, 2.11 & 2.13 implies the necessity of an integration over the parameter space. However, this is not always analytically possible considering that high-dimensional functions are usually involved. This has led to the development of suitable approximation techniques, with focus on three main strategies: Laplace and related analytic techniques, adaptive quadrature based on classical numerical analysis and the Markov Chain Monte Carlo methods (Smith and Roberts, 1993).

The Markov Chain Monte Carlo methods, owing to their simplicity, are the most widely used of these approximation techniques (Smith and Roberts, 1993). These trace their roots to attempts by mathematical physicists to integrate very complex functions by random sampling (Hastings, 1970). Four Monte Carlo methods: Monte Carlo integration, Importance sampling, Metropolis-Hastings sampling and Gibbs sampling are briefly described following a discussion of Markov chains.

2.4.1 Markov Chains

Before discussing the Monte Carlo Markov Chain methods, a brief introduction to Markov chains is discussed in this section. Consider Y_t, the value of a random variable at time t, and let the space state refer to the range of possible Y values. The random variable is a Markov Chain if the transition probabilities between different values in the state space depend only on the random variable's current state (Smith and Roberts, 1993) i.e.,

$$\Pr\left(Y_{t+1} = s_j \middle| Y_0 = s_k, \ldots, Y_t = s_i\right) = \Pr\left(Y_{t+1} = s_j \middle| Y_t = s_i\right) \qquad 2.14$$

Thus for a Markov random variable, the only information about the past needed to predict the future is the current state of the random variable, knowledge of earlier states do not change the transition probability. A Markov Chain refers to a sequence of random variables $(Y_0, \ldots, Y_n)$ generated by a Markov process. A

particular process is defined most critically by its transition probabilities, $P(i,j) = P(i \to j)$, which is the probability that a process at state space s_i moves to state s_j in a single step (Smith and Roberts, 1993),

$$P(i,j) = P(i \to j) = \Pr(Y_{t+1} = s_j | Y_t = s_i)$$

2.15

The notation $P(i \to j)$ is used to imply a move from i to j.

Let

$$\pi_j(t) = \Pr(Y_t = s_j)$$

2.16

denote the probability that a chain is in state j at time t and let $\pi(t)$ denote the row vector of the state space probabilities at step t. We start the chain by specifying a starting vector $\pi(0)$. Often all the elements of $\pi(0)$ are zero except for a single element of 1, corresponding to the process starting in that particular state. As the chain progresses, the probability values get spread out over the possible state space.

The probability that the chain has state value s_i at step $t+1$ is given by the Chapman-Kolomogrov equation (Smith and Roberts, 1993), which sums up the probability of being in a particular state at the current step and the transition probability from that state into state s_i,

$$\begin{aligned}\pi_i(t+1) &= \Pr(Y_{t+1} = s_i) \\ &= \sum_k \Pr(Y_{t+1} = s_i | Y_t = s_k).\Pr(Y_t = s_k) \\ &= \sum_k P(k \to i)\pi_k(t) = \sum_k P(k,i)\pi_k(t)\end{aligned}$$

2.17

Successive iteration of the Chapman-Kolomogrov equations describes the evolution of the chain.

The Chapman-Kolmologrov equations can be written more compactly in matrix form as follows:

Define the **probability transition matrix P** as the matrix whose i,j^{th} element is $P(i,j)$, the probability of moving from state i to state j, $P(i \to j)$. This implies that the rows sum to one as

$$\sum_j P(i,j) = \sum_j P(i \to j) = 1$$

2.18

The Chapman-Kolomogrov equation becomes

$$\pi(t+1) = \pi(t).\mathbf{P}$$

2.19

Using the matrix form, the iteration of the Chapman-Kolomogrov equation quickly becomes apparent, as

$$\pi(t) = \pi(t-1).\mathbf{P} = (\pi(t-2).\mathbf{P}).\mathbf{P} = \pi(t-2).\mathbf{P^2}$$

2.20

Continuing in this manner shows that

$$\pi(t) = \pi(0).\mathbf{P^t}$$

2.21

Defining the n-step transition probability $P_{i.j}^{(n)}$ as the probability that the process is in state j given that it started in state i n steps ago, i.e.,

$$P_{i.j}^{(n)} = \Pr\left(Y_{t+n} = s_j \middle| Y_t = s_i\right)$$

2.22

It immediately follows that $P_{i.j}^{(n)}$ is just the i,j^{th} element of P^n.

If there exists a positive integer such that $P_{i.j}^{(n)} > 0$ for all i,j, then the Markov chain is said to be **irreducible** (Smith and Roberts, 1993). That is, all states communicate with each other, as one can always go from one state to any other state (although it may take more than one step).

2.4.2 Monte Carlo Integration

Markov Chain Monte Carlo methods attempt to simulate direct draws from complex distributions and are so named because they use the previous sample values to randomly generate the next sample value, generating a Markov Chain.

Suppose the solution to the complex integral below is to be computed.

$$\int_a^b h(y)dy$$

2.23

If $h(y)$ can be decomposed into the product of a function $f(y)$ and a probability density function $p(y)$ defined over the interval (a,b), then note that

$$\int_a^b h(y)dy = \int_a^b f(y)p(y)dy = E_{p(y)}[f(y)]$$

2.24

so that the integral can be expressed as an expectation of $f(y)$ over the density $p(y)$. Thus, if a large number $y_1,\ldots,y_n$ of random variables is drawn from the density $p(y)$, then

$$\int_a^b h(y)dy = E_{p(y)}[f(y)] \approx \frac{1}{n}\sum_{i=1}^{n} f(y_i)$$

2.25

This is referred to as Monte Carlo integration.

Monte Carlo integration can be used to approximate the posterior distributions required for a Bayesian analysis. The integral $I(y) = \int f(y|\theta)p(\theta)d\theta$ can therefore be approximated by

$$\hat{I}(y) = \frac{1}{n}\sum_{i=1}^{n} f(y|\theta_i)$$

2.26

where θ_i are draws from a density $p(\theta)$.

2.4.3 Importance Sampling

Suppose the density $p(y)$ roughly approximates a density of interest $q(y)$, then

$$\int f(y)q(y)dy = \int f(y)\left(\frac{q(y)}{p(y)}\right)p(y)dy = E_{p(y)}\left[f(y)\left(\frac{q(y)}{p(y)}\right)\right]$$

2.27

This forms the basis for the method of importance sampling (Yuan and Druzdzel, 2004, Kuczera and Parent, 1998), with

$$\int f(y)q(y)dy \approx \frac{1}{n}\sum_{i=1}^{n} f(y_i)\left(\frac{q(y_i)}{p(y_i)}\right) \qquad 2.28$$

where the y_i are drawn from the distribution given by $p(y)$.

2.4.4 Metropolis-Hastings Sampling

One problem with applying Monte Carlo integration is in obtaining samples from some complex probability distribution $p(y)$ (Hastings, 1970). Attempts to solve this problem resulted in the Metropolis-Hastings algorithm (Hastings, 1970). Suppose a sample is to be drawn from some distribution $p(y)$ where $p(y) = f(\theta)/K$, where the normalizing constant K may not be known, and very difficult to compute. The Metropolis algorithm generates a sequence of draws from this distribution as follows (Kuczera and Parent (1998), Hastings (1970)):

1. Start with any initial value y_o satisfying $f(y_o) > 0$.
2. Using the current y value, sample a **candidate point** y^* from some jumping distribution $q(y_1, y_2)$, which is the probability of returning a value of y_2 given a previous value of y_1. This distribution is also referred to as the proposal or candidate-generating distribution. The only restriction on the jump density in the Metropolis algorithm is that it is symmetric, i.e., $q(y_1, y_2) = q(y_2, y_1)$.
3. Given the candidate point y^*, calculate the ratio of the density at the candidate point (y^*) and the current y points,

$$\alpha = \frac{p(y^*)}{p(y_{t-1})} = \frac{f(y^*)}{f(y_{t-1})} \qquad 2.29$$

   Notice that because equation (2.29) is the ratio of $p(y)$ under two different values, the normalizing constant K cancels out.
4. if the jump increases the density $(\alpha > 1)$, accept the candidate point (set $y_t = y^*$) and return to step 2. If the jump decreases the density

$(\alpha < 1)$, then with probability α accept the candidate point, else reject it and return to step 2.

The Metropolis sampling can, therefore, be summarized as first computing

$$\alpha = \min\left(\frac{f(y^*)}{f(y_{t-1})}, 1\right) \qquad 2.30$$

and then accepting a candidate point with probability α (the probability of a move). This generates a Markov Chain $(y_0, y_1, ..., y_k, ...)$, as the transition probabilites from y_t to y_{t+1} depends only on y_t and not $(y_0, ..., y_{t-1})$. Following a sufficient **burn-in period,** the chain approaches its stationary distribution and samples from the vector $(y_{k+1}, ..., y_{k+n})$ are samples from $p(y)$.

Hastings (1970) generalized the Metropolis algorithm by using an arbitrary transition probability function $q(y_1, y_2) = \Pr(y_1 \to y_2)$, and setting the acceptance probability for a candidate point as

$$\alpha = \min\left(\frac{f(y^*)q(y^*, y_{t-1})}{f(y_{t-1})q(y_{t-1}, y^*)}, 1\right) \qquad 2.31$$

This is the Metropolis-Hastings algorithm. Assuming that the proposal distribution is symmetric, i.e. $q(y_1, y_2) = q(y_2, y_1)$, recovers the original Metropolis algorithm.

2.4.5 Gibbs Sampling.

The Gibbs Sampler is a special case of Metropolis-Hastings sampling where the random value is always accepted (i.e. $\alpha = 1$) (Geman and Geman (1984), Smith and Roberts (1993)). The Gibbs sampler considers univariate conditional distributions – the distribution when all the random variables but one are assigned fixed values (Smith and Roberts ,1993). Such conditional distributions are far easier to simulate than the complex joint distributions and usually have simple forms (often being normals, Gammas, Betas etc.). Thus n random

variables are simulated from n univariate conditionals rather than generating a single n-dimensional vector in a single pass using a joint distribution.

As an example, consider a bivariate random variable (x,y) where the marginal distributions $P(x)$ and $P(y)$ are supposed to be computed. Let $P(x|y)$ and $P(y|x)$ be the respective conditional distributions of the two variables. Starting with some initial value y_o of y, x_o is obtained by generating a random variable from the conditional distribution $P(x|y=y_o)$. The sampler then uses x_o and $P(y|x=x_o)$ to generate a new value of y_1. The sampler then proceeds as follows

$$x_i \sim P\left(x|y = y_i\right)$$
$$y_i \sim P\left(y|x = x_i\right)$$

2.32

Repeating this process k times, generates a Gibbs sequence of length k, where a subset of points (x_j,y_j), for $1 \leq j \leq m < k$ are taken as the simulated draws from the full joint distribution. The Gibbs sequence converges to a stationary (equilibrium) distribution that is independent of the starting values and this stationary distribution is the target distribution of the complex joint distribution.

Using the Gibbs sampling methodology to approximate the posterior distribution in equation (2.13) where the parameter vector $\theta = \left(\theta_1,\ldots,\theta_n\right)$, will result in a sequence of parameters that will form the joint posterior distribution of parameter vector $\theta = \left(\theta_1,\ldots,\theta_n\right)$.

2.5 Standard Statistical Analysis.

Since it is the prediction of extreme rainfall behavior that is the focus of this study, an asymptotic model to describe the stochastic variation at extreme levels of a process is adopted. If the distribution of rainfall events within a year is such that the tail of the distribution decays exponentially, then a family of extreme value distributions can be applied to the annual maxima (Alexander, 2001). The three member distributions in this family are briefly described below.

2.5.1 Extreme Value Distribution type 1 (Gumbel)

The Gumbel (Extreme Value type I) distribution has a constant positive skewness and is commonly used for hydrological analyses. The maxima from any distribution that converges on an exponential function at the positive tail (normal, Chi-square, lognormal etc.) will have a Gumbel distribution (Alexander, 2001).

The cumulative distribution function is:

$$F(y) = \exp(\exp[-(y-\mu)/\alpha]) \tag{2.33}$$

where μ is the location parameter,

α is the scale parameter.

2.5.2 Extreme Value Distribution type 2 (Frechet)

This is a positively skewed distribution. If the raw data are a Frechet distribution then their logarithms will follow a Gumbel distribution (Alexander, 2001).

The cumulative distribution function is:

$$F(y) = \exp\left[-((y-\mu)/\alpha)^{-\alpha}\right] \tag{2.34}$$

where μ is the location parameter,

α is the scale parameter.

2.5.3 Extreme Value Distribution type 3 (Weibull)

The Weibull distribution is negatively skewed (Alexander, 2001) and the cumulative distribution function is:

$$F(y) = \exp\left\{-\left[-(y-\mu)/\alpha\right]^{\alpha}\right\} \tag{2.35}$$

where μ is the location parameter,

α is the scale parameter.

2.5.4 General Extreme Value (GEV) distribution

The general extreme value (GEV) distribution is the generalized form of the above three extreme value distributions. It is a family of the three sub-types of distributions, which are classified according to the value of skewness coefficient. The skewness coefficient of the Gumbel distribution is 1.1396 while the Frechet and the Weibull distributions have skewness coefficient values greater than and less than 1.1396 respectively (Alexander, 2001).

The General Extreme Value (GEV) distribution is very flexible and is the distribution that is recommended for use in the UK Flood Studies Report (National Environment Research Council, 1975). It has a cumulative distribution function of the form:

$$F(y) = \exp\left\{-\left[1 - k(y-\mu)/\alpha\right]^{1/k}\right\} \qquad 2.36$$

where μ is the location parameter,

α is the scale parameter.

k is the shape parameter and distinguishes the three sub-types of distributions from one another.

The standardized GEV distribution has a cumulative distribution function of the form (National Environment Research Council, 1975):

$$G(z) = e^{-e^{-y/k}} \qquad 2.37$$

where the standardized variate, z, is related to y by

$$z = \left[1 - k(y-u)/\alpha\right] \qquad 2.38$$

The distribution of the standardized variate, z, depends on the shape parameter, k, and is obtained from equation (2.37) as

$$z = \exp\{k \ln[-\ln G(z)]\} \qquad 2.39$$

The moments of the standardized variate are (National Environment Research Council ,1975):

$$Mean\ \ (z)=\Gamma(1+k)$$

$$\text{var}\ \ (z)=\Gamma(1+2k)-\Gamma^2(1+k)$$

$$\text{and}\ \ skewness\ \ (g)=\frac{\Gamma(1+3k)-3\Gamma(1+2k).\Gamma(1+k)+2\Gamma^3(1+k)}{\left[\Gamma(1+2k)-\Gamma^2(1+k)\right]^{3/2}}$$

where the gamma function, $\Gamma(x)$, is defined as $\Gamma(x)=\int_0^\infty t^{x-1}e^{-t}dt$

2.5.5 GEV Parameter Estimation using Conventional Moment Estimators.

The mean, standard deviation and skewness of the annual maxima are calculated as follows:

$$Mean\ \ (\bar{y})=\frac{1}{n}\sum_{i=1}^{n}y_i \qquad 2.40$$

$$S\tan dard\ Deviation\ \ (s)=\left[\frac{1}{(n-1)}\sum_{i=1}^{n}(y_i-\bar{y})^2\right]^{1/2} \qquad 2.41$$

$$Skewness\ \ (g)=\frac{n}{(n-1)(n-2)s^3}\sum_{i=1}^{n}(y_i-\bar{y})^3 \qquad 2.42$$

The location parameter, μ, is estimated using the mean, $(\bar{y})$ of the data. The shape parameter, k, given the skewness (g) can be estimated from the following polynomial (National Environment Research Council ,1975).

$$k = a + bg + cg^2 + dg^3 + eg^4$$

2.43

where

$a = 0.27715$

$b = -0.32773$

$c = 0.09027$

$d = -0.01260$

$e = 0.00070$

The scale parameter, α, is related to the shape parameter , k, as follows (National Environment Research Council ,1975)

$$\alpha = -Bk$$

2.44

where B is given by

$$B = \left(s^2 / \mathrm{var}(z)\right)^{1/2}$$

2.45

Conventional moment estimators suffer from the effects of sampling uncertainty and bias i.e the higher moments are more sensitive to the outlying data values due to the fact that these higher moments involve powers of the data values that are greater than one (Hosking et al., 1985).

2.5.6 GEV Parameter Estimation using Maximum Likelihood method.

If a random variable Y has a GEV distribution function of

$$F(y) = \exp\left\{-\left[1 - k(y-\mu)/\alpha\right]^{1/k}\right\}$$

2.46

The corresponding p.d.f is given by:

$$f(y) = \frac{1}{\alpha}\left(1 + k\frac{(y-\mu)}{\alpha}\right)^{-1/k-1} \exp\left\{-\left(1 + k\frac{y-\mu}{\alpha}\right)^{-1/k}\right\}$$

2.47

$$for \quad 1 + k\frac{y-\mu}{\alpha} > 0 \qquad and \qquad 0 \qquad elsewhere$$

Consequently, the likelihood function based on data $Y=(Y_1,\ldots Y_n)$ is given by

$$L(\alpha,k,\mu:Y)=\prod_{i=1}^{n} f_{\alpha,k,\mu}(y_i)=\prod_{i=1}^{n}\frac{1}{\alpha}\left(1+k\frac{y_i-\mu}{\alpha}\right)^{-1/k-1}\exp\left\{-\left(1+k\frac{y_i-\mu}{\alpha}\right)^{-1/k}\right\}$$

$$=\alpha^{-n}\left[\prod_{i=1}^{n}\left(1+k\frac{y_i-\mu}{\alpha}\right)\right]^{-1/k-1}\exp\left\{-\sum_{i=1}^{n}\left(1+k\frac{y_i-\mu}{\alpha}\right)^{-1/k}\right\}$$

2.48

for $1+k\frac{y_i-\mu}{\alpha}>0$ for all y_i, and 0 elsewhere.

The corresponding log-likelihood is

$$l(\alpha,k,\mu:Y)=-n\ln\alpha-(1+k)\sum_{i=1}^{n}x_i-\sum_{i=1}^{n}e^{-x_i}$$

2.49

where $x_i=k^{-1}\ln\left(1+k\frac{y_i-\mu}{\alpha}\right)$

then by definition, the maximum likelihood estimator (MLE) $\hat{\theta}=(\hat{\alpha},\hat{k},\hat{\mu})$ for the unknown parameters $\theta=(\alpha,k,\mu)$ equals

$$\hat{\theta}=\arg\max_{\theta\in\Theta} l(\alpha,k,\mu:Y)$$

2.50

As far as the estimation of large quantiles of y_p is concerned, the equi-variance property of maximum likelihood implies that the MLE of a quantile is obtained by substitution of the MLE's (α,k,μ) into the quantile function (generalized inverse distribution function) of the generalized extreme-value distribution.

$$y_p=\mu-\frac{\alpha}{k}\left[1-(-\ln p)^{-k}\right]$$

2.51

Differentiating equation (2.49) with respect to the parameters (α, k, μ) yields the likelihood system of equations. Clearly, no explicit solution exists for these equations, so the likelihood equations must be solved iteratively. Numerical procedures such as variants of Newton-Raphson algorithm are required.

MLE for generalized extreme value distribution has an advantage that it has good asymptotic properties which depend on the parameters' values, in particular, where k falls in the range $(-\infty, +\infty)$.

The major disadvantage of MLE is the fact that its small-sample properties are not as good as its asymptotic properties.

2.5.7 GEV Parameter Estimation using L-moment Estimators.

Probability-weighted moments are a generalization of the usual moments of a probability distribution, which give increasing weight to the tail information (Hosking et al, 1985). The probability weighted moments (PWM's) of a continuous random variable Y with distribution function F are the quantities

$$M_{p,r,s} = E\left[Y^p \{F(Y)\}^r \{1 - F(Y)\}^s\right] \qquad 2.52$$

for real p, r, and s. If the inverse distribution function, $y(F)$ can be written in closed form, then equation (2.52) may be written more conveniently as

$$M_{p,r,s} = \int_0^1 \{y(F)\}^p F^r (1-F)^s \, dF \qquad 2.53$$

The quantities $M_{p,0,0}$ $(p = 1, 2, \ldots)$ are the usual product moments of Y. The moments $M_{1,r,s}$ are, however, preferable for estimating the parameters of the distribution of Y, since the occurrence of only the first power of Y in the expression for $M_{1,r,s}$ means that the relationship between the parameters and the moments often takes a simpler form in this case than when using the

conventional moments (Hosking et al., 1985). When r and s are integers, $F^r(1-F)^s$ may be expressed as a linear combination of either powers of F or powers of $(1-F)$, so it is natural to summarize a distribution either by the moments $M_{1,r,0}$ $(r=0,\ 1,\ 2,...)$ or by $M_{1,0,s}$ $(s=0,\ 1,\ 2,...)$. Hosking et al. (1985) used $M_{1,r,0}$ and defined the moment

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

2.54

The moments (equation 2.54) are more conveniently defined in terms of the exceedance probability i.e. $q=(1-F(Y))$, for a random variable, as follows

$$\xi_r = \int_0^1 q^r\, y(q)\, dq \qquad (r=0,\ 1,\ 2,...)$$

2.55

Given a random sample of size n from the distribution F, estimation of β_r is most conveniently based on the ordered sample $y_1 \le y_2 \le ... \le y_n$. The statistic

$$b_r = n^{-1}\sum_{j=1}^{n}\frac{(j-1)(j-2)..(j-r)}{(n-1)(n-2)..(n-r)}y_j$$

2.56

is the unbiased estimator of β_r (Landwehr et al., 1979).

Another method of estimating β_r is by

$$\hat{\beta}_r\left[p_{j,n}\right] = n^{-1}\sum_{j=1}^{n} p_{j,n}^r\, y_{j,}$$

2.57

where $p_{j,n}$ is the plotting position. Reasonable choices of $p_{j,n}$, such as $p_{j,n}=(j-a)/n,\ 0<a<1,$ or $p_{j,n}=(j-a)/(n+1-2a),\ -\frac{1}{2}<a<\frac{1}{2}$, yield estimators

$\hat{\beta}_r\left[p_{j,n}\right]$, which are asymptotically equivalent to b_r and therefore, consistent estimators of β_r (Hosking et al., 1985).

Hosking (1990) showed that certain linear combinations of probability weighted moments, referred to as L-moments, can provide valid measures of dispersion, skewness and kurtosis analogous to conventional moments. The mean, dispersion and the third moments can be written as

$$\lambda_1 = \xi_0 \tag{2.58}$$

$$\lambda_2 = \xi_0 - 2\xi_1 \tag{2.59}$$

$$\lambda_3 = \xi_0 - 6\xi_1 - 6\xi_2 \tag{2.60}$$

The L-skewness is defined as λ_3/λ_2, and λ_2/λ_1 is analogous to the coefficient of variation.

The first three L-moments of a GEV distribution in terms of its location (μ), scale (α) and the shape (k) parameters are given as

$$\lambda_1 = \mu + \frac{\alpha}{k}\left[1 - \Gamma(1+k)\right] \tag{2.61}$$

$$\lambda_2 = \left(1 - 2^{-k}\right)\frac{\alpha}{k}\Gamma(1+k) \tag{2.62}$$

$$\frac{\lambda_3}{\lambda_2} = 2\frac{\left(1 - 3^{-k}\right)}{\left(1 - 2^{-k}\right)} - 3 \tag{2.63}$$

By inverting equations (2.61, 2.62 and 2.63) the distribution parameters can be calculated directly. Hosking (1990) has proposed the following simple approximation for calculating the shape parameter

$$k = 7.859c + 2.9554c^2$$

2.64

and

$$c = \frac{2}{3 + \lambda_3 / \lambda_2} - \frac{\log 2}{\log 3}$$

2.65

The other parameters can then be calculated as

$$\alpha = \frac{\lambda_2 k}{\Gamma(1+k)(1-2^{-k})}$$

2.66

$$\mu = \lambda_1 - \frac{\alpha}{k}(1 - \Gamma(1+k))$$

2.67

The method of L-moments has been shown to be superior to other conventional methods of moments as demonstrated by Hosking et al (1985) through extensive simulations. The method is less sensitive to sampling variability or measurement errors in the extreme data series, therefore giving more robust and accurate estimates. In addition, they are less subject to bias, and approximate normality more closely in finite samples and are more accurate than maximum likelihood estimators (MLE) where small samples are concerned.

2.5.8 Estimation of the Magnitude of an event for a given Return Period Using the GEV distribution.

The following steps are followed:

1. Determine the parameters of the GEV, (α, μ, k), as explained in sections *2.5.5* or *2.5.6* or *2.5.7* depending on the method of parameter estimation chosen.

2. Determine the standardized variate z_T for the required return period, T, from equation 2.68.

$$z = \exp\{k \quad \ln(1 - \ln[1 - 1/T])\}$$

2.68

3. Determine the magnitude y_T from equation 2.69.

$$y_T = \mu + \alpha/k - (\alpha/k)z_T$$

2.69

CHAPTER 3 : METHODOLOGY

3.1 Brief Description of the Methodology

The Bayesian analysis of the data involved the use of the Generalized Pareto distribution to model exceedances of rainfall over a threshold and this was expressed in the form of a likelihood function as shown in section 3.3. Using unpatched daily rainfall records, the threshold, u_n, was selected from the mean residual life plot of the rainfall data as described in section 3.3. The formulation of the joint prior distribution for the Generalized Pareto Distribution (GPD) parameters, α and k, was then done following the procedure outlined in section 3.4. The posterior was then formulated according to equation (2.9) resulting in the expression for the posterior distribution shown as equation (3.9). The presence of a proportionality symbol in equation (3.9) implies a need for an integration over the parameter space for the posterior distribution of the GPD parameters to be evaluated. However, this is not analytically possible considering that high-dimensional functions involved. For this reason, the Gibbs sampler, a Monte Carlo Markov Chain technique, was used to generate a sequence of sample parameter sets whose distribution approximated the posterior distribution. Using this output of the Gibbs Sampler as an input into the predictive distribution (equation 3.13), rainfall magnitudes corresponding to return periods from 1 to 200 years were generated. Figure 3.1 is a graphical representation of the methodology.

Smithers and Schulze (2003) carried out a study on design rainfall estimation in South Africa using a regional index storm approach which utilizes L-moments for parameter estimation. They developed a computer program that gave as an output estimates of rainfall magnitudes corresponding to durations of 5 minutes to 7 days and return periods of 2 to 200 years, based on the regional storm index approach.

For a given return period, the estimates of rainfall magnitudes obtained using the Bayesian analysis in this study were compared to the corresponding rainfall magnitudes obtained from running the program that was developed by Smithers and Schulze (2003) based on the regional storm index approach.

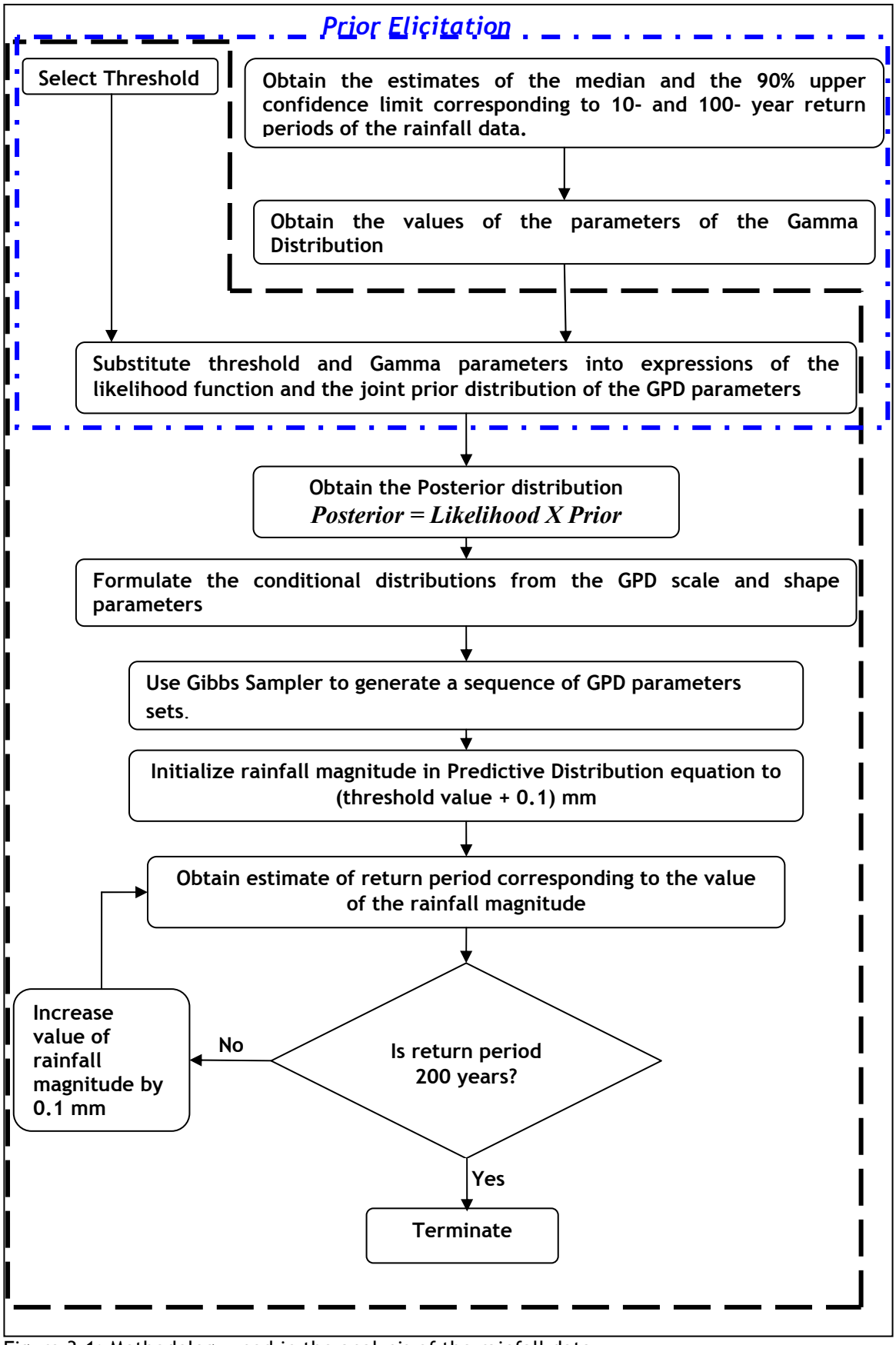

Figure 3.1: Methodology used in the analysis of the rainfall data

3.2 Data utilized.

The data that was used for analysis was unpatched daily rainfall records of the selected stations (Figure 3.2). The data is from the raster database of annual, monthly and daily rainfall for South Africa that was developed for the Water Research Commission by Lynch (2004).

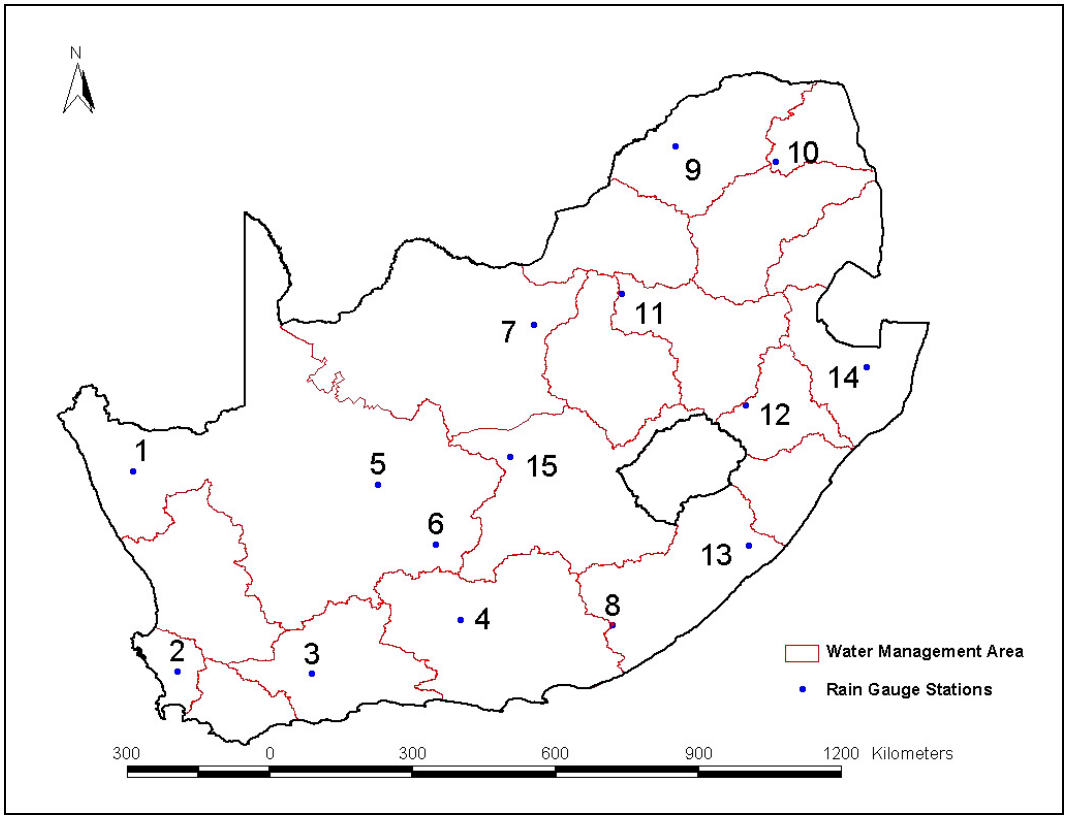

Figure 3.2: Location of rainfall stations used in the study

Table 3.1 shows the details of the stations that are indicated by numbers in Figure 3.2. The selection of the stations was done in such a way that one station was selected from each of the fifteen clusters of relatively homogenous short-duration (≤ 24 *hrs*) rainfall that the study by Smithers and Schulze (2003) came up with. The record length of the data used differs, and ranges from 43 years (Drielingspan station) to 150 years (Malmesbury (Mun) station). Figures 3.3a and 3.3b show time series of the rainfall data of the respective stations that were used in this study.

Table 3.1: Details of the selected rainfall Stations

	Station Name	SAWS No.	Longitude	Latitude	Years on Record
1	SPRINGBOK (MUN)	0214670 W	17.884	29.667	120
2	MALMESBURY (MUN)	0041417 W	18.734	33.451	150
3	LADISMITH (MUN)	0046479 W	21.267	33.501	124
4	ABERDEEN (MUN)	0095119 W	24.067	32.484	124
5	DRIELINGSPAN	0224721 A	22.501	29.917	43
6	LEKKERVLEI	0142153 W	23.601	31.051	100
7	POORTJIE	0433804 W	25.450	26.900	96
8	HOGSBACK (BOS)	0078755 W	26.934	32.584	124
9	VILLA NORA (POL)	0675182 W	28.117	23.534	96
10	DE HOEK (BOS)	0679019 W	30.017	23.817	96
11	ZAMENKOMST	0474198 W	27.117	26.317	99
12	MOORSIDE	0333805 A	29.451	28.417	63
13	FLAGSTAFF (MUN)	0153875 W	29.501	31.084	102
14	ZILVERHOUT	0374402 W	31.734	27.700	125
15	KOFFIEFONTEIN (POL)	0258894 W	25.001	29.401	122

3.3 Selection and application of Distribution

Assuming that the daily rainfall values are independent and identically distributed, the Generalized Pareto distribution was chosen to model the excesses of rainfall values above a chosen threshold. The use of the 'exceedances over a threshold' approach to analyze extreme events ensures that as much of the extreme data present in the rainfall records as possible is included (Coles et al., 2003). This is in contrast with the use of annual maxima where only the maximum recorded value of a year is considered. This leaves out some records which may not be the maximum in years they were recorded but are greater than some of the annual maximums of other years on the record.

The form of the Generalized Pareto distribution is

$$G(y \mid k,\alpha,u) = \begin{cases} 1-\left(1+k\dfrac{y-u}{\alpha}\right)^{-1/k} & if \quad k \neq 0 \\ 1-\exp\{-(y-u)/\alpha\} & if \quad k = 0 \end{cases} \qquad 3.1$$

where y, u, α, k is the rainfall observation (mm), the threshold (mm), scale parameter and the shape parameter, respectively.

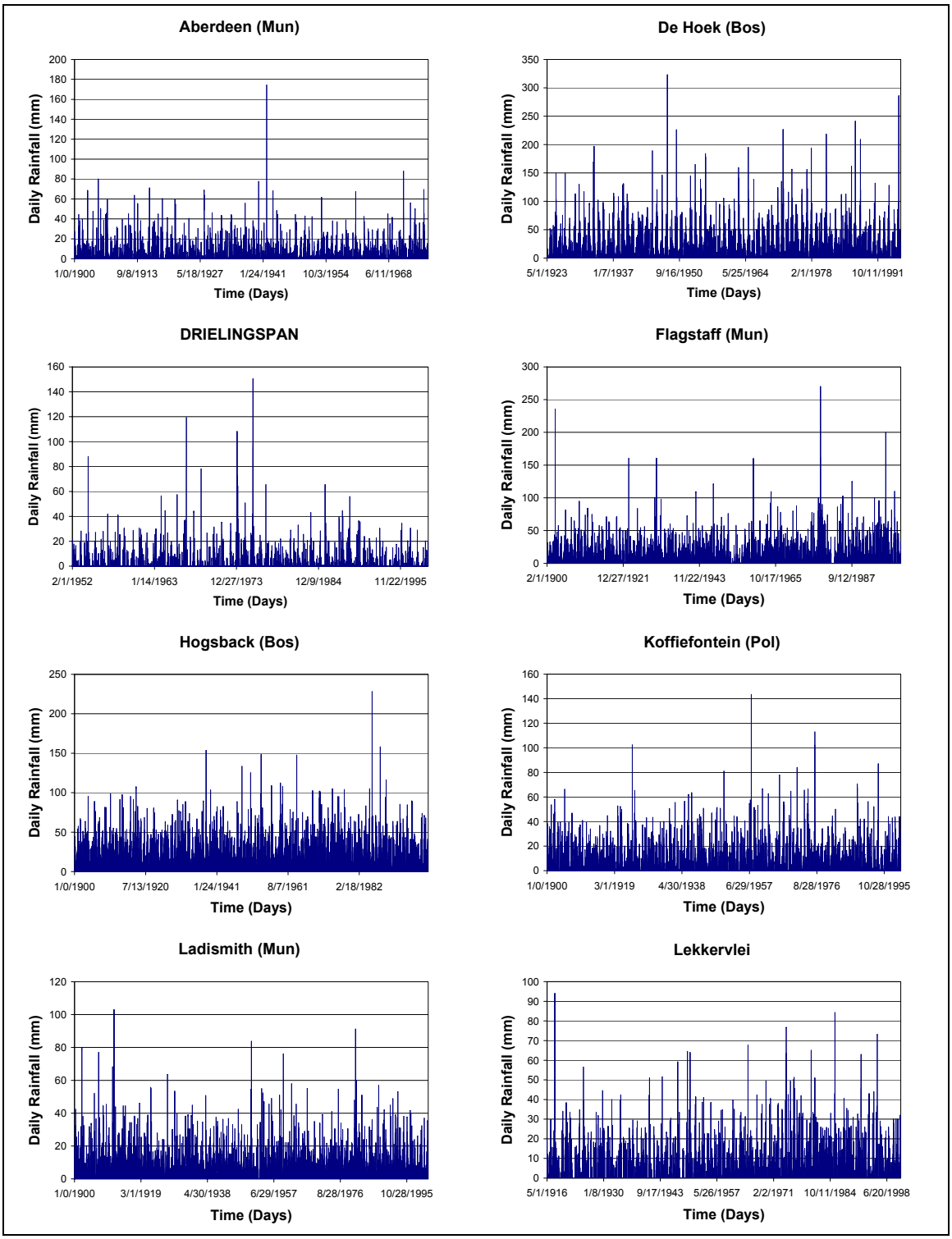

Figure 3-3a: A time series plot of daily rainfall for the respective gauge stations.

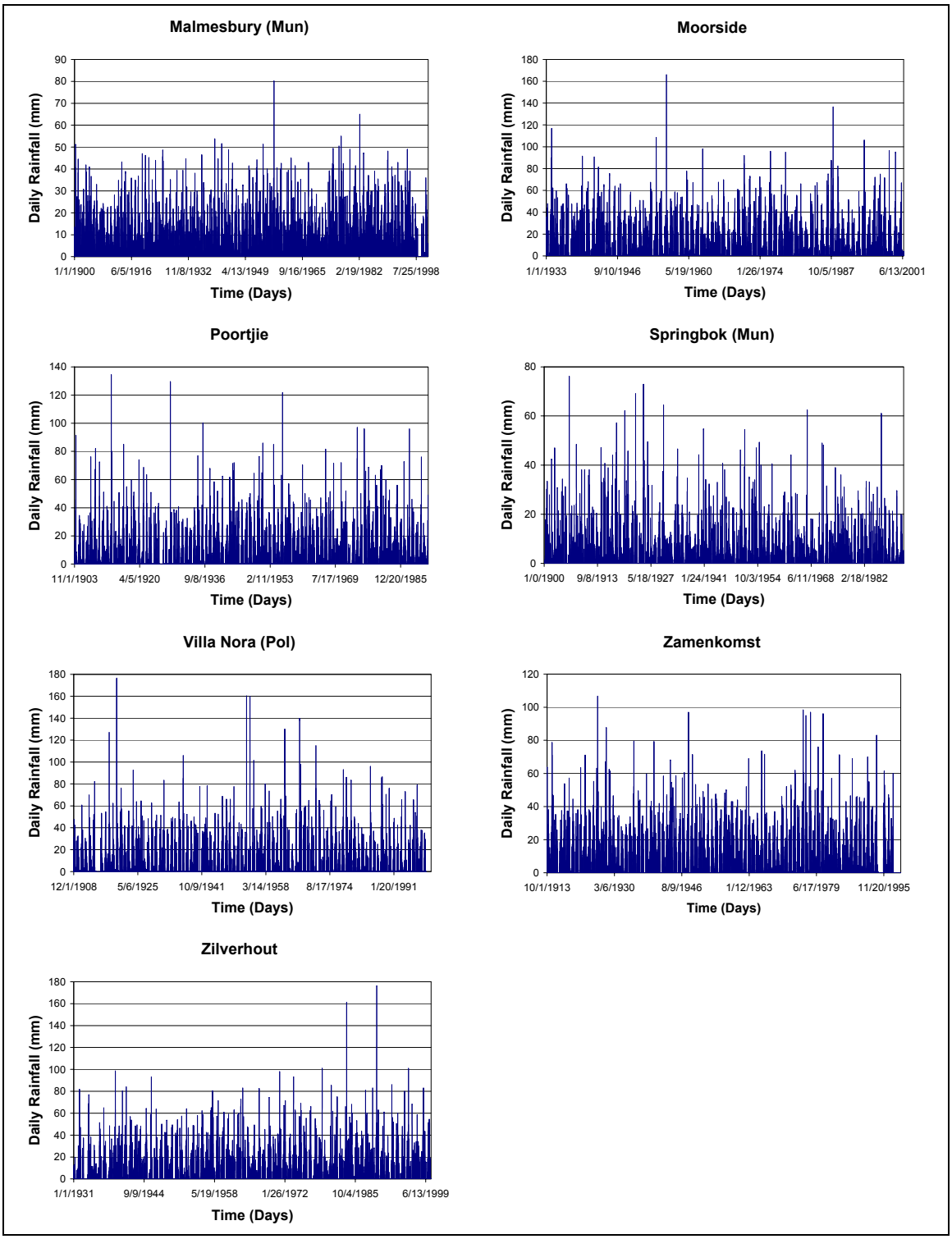

Figure 3-3b: A time series plot of daily rainfall for the respective gauge stations.

The choice of threshold requires some care since a balance between bias and variance has to be found. Too low a value is likely to compromise the asymptotic justification of the model, leading to bias; too large a value will lead to few exceedances, and therefore a large variance of estimators (Coles et al., 2003). The Mean Residual Life plot, first used by Davison and Smith (1990), was used to select the threshold u. The mean Residual life Plot is a plot of $(u_i, e(v_i))$ for a range of possible thresholds, u_i, where $e(V_i)$ is the empirical mean of the set $\{y_i - u : y_i > u\}$. Because of the identity

$$E(y-u|y>u) = \frac{\alpha - ku}{1-k} \qquad 3.2$$

valid for a generalized Pareto model of threshold excesses in the usual case of $k < 1$, it follows that the plot of points $(u_i, e(v_i))$ should be approximately linear above a level u for which the model is valid.

The Generalized Pareto model for the excesses of u was then used to develop the likelihood function. The excesses over the threshold, $v_1, \ldots, v_n$ were assumed to be independent resulting in the likelihood function of the form.

$$L(\theta; v) = \prod_{i=1}^{n} \left(\frac{1}{\alpha} \left[1 + \frac{k v_i}{\alpha} \right]^{\frac{-(1+k)}{k}} \right) \quad for \quad k \neq 0 \qquad 3.3$$

$$L(\theta; v) = \prod_{i=1}^{n} \left(\frac{1}{\alpha} \left(\exp\left\{ v_i / \alpha \right\} \right) \right) \quad for \quad k = 0$$

where $v_i = y_i - u$

3.4 Prior Elicitation

Ordinarily, expert hydrologists would provide their beliefs about the extremal behavior of rainfall in the study area(s). Their beliefs are meant to guide in elicitation of prior information. It is reasonable to hope that an expert can provide relevant prior information about extremal behavior, since they have specific knowledge of the characteristics of the data under study. However,

there was no one who could provide that expert information for the chosen study area. For this reason, the prior information was obtained using the General Extreme Value (GEV) distribution which is a generalized form of the three extreme value distributions (Coles and Tawn, 1996). Using the standard statistical analysis following the procedure described in section 2.5, predictions for rainfall magnitudes(mm) corresponding to return periods of 10 and 100 years were obtained as well as the 90 % upper quantile magnitude (mm). Parameter estimation, for the GEV parameters (α, k, μ) is done using the conventional moment estimators as described in section 2.5.5.

By the inversion of equation (equation 3.1), the $(1-p)$ quantile of the distribution is obtained as:

$$q = u + \frac{\alpha}{k}\left(p^{-k} - 1\right) \qquad 3.4$$

where q can be viewed as the magnitude of rainfall associated with a return period of $1/p$ time units. The elicitation of the prior information was therefore done in terms of q_1 and q_2, instead of the shape-scale parameterization of the GPD, for specific values of $(p_1 > p_2)$ where $q_1 < q_2$.

Coles and Tawn (1996) suggest to work with the differences $d_i = q_i - q_{i-1},\ i = 1,\ 2,\ 3$ with $q_o = e_1$. Where e_1 is the physical lower bound of the variable. Since it is rainfall that is being analyzed, the lower bound naturally is zero, as such $e_1 = 0$.

These differences $(d_i's)$ were assumed to be independent and to follow a gamma distribution, $Ga(a,b)$ as suggested by Coles and Tawn (1996) and Behrens et al. (2004). The prior information was elicited obtaining the GEV estimates of the median, q_{i50} and 90% quantile, q_{i90} estimates for specific values of p using the procedure outlined in sections 2.5.5 and 2.5.8. The 10 and 100 year return periods were considered, which correspond, respectively, to $p_1 = 0.1$ and $p_2 = 0.01$. With the elicited information, the parameters of the

gamma distributions (equations 3.5 and 3.6) were obtained using the ***Solver*** function in the Microsoft Excel program.

$$d_1 = q_1 \sim Ga(a_1, b_1)$$

3.5

$$d_2 = q_2 - q_1 \sim Ga(a_2, b_2)$$

3.6

.

The assumption that the differences, $d_i's$, were independent lead to the formulation of the joint prior distribution for the GPD parameters, α and k, as

$$\pi(\alpha, k) \propto d_1 \times d_2$$

3.7

Substitution for q in equations 3.5 and 3.6 using equation 3.4 yielded equations for the differences, $d_i's$ in terms of the GPD parameters, α and k. These expressions were substituted into equation 3.7 to give the joint prior distribution for α and k as

$$\pi(\alpha, k) \propto \left(u + \frac{\alpha}{k}\left(p_1^{-k} - 1\right)\right)^{a_1 - 1} \times \left(\frac{\alpha}{k}\left(p_2^{-k} - p_1^{-k}\right)\right)^{a_2 - 1} \times \exp\left[-b_1\left(u + \frac{\alpha}{k}\left(p_1^{-k} - 1\right)\right)\right] \times \exp\left[\frac{-b_2\alpha}{k}\left(p_2^{-k} - p_1^{-k}\right)\right]$$

3.8

Note: The detailed derivation of equation 3.8 is shown in *APPENDIX C*.

3.5 Formulation of the Posterior distribution

Having obtained both the likelihood (equation 3.3) and the prior distribution for the GPD parameters, α and k, (equation 3.8), the posterior distribution for the parameters was obtained using equation 2.9 as

$$\pi(\alpha, k|Y) \propto \left(u + \frac{\alpha}{k}\left(p_1^{-k} - 1\right)\right)^{a_1 - 1} \times \left(\frac{\alpha}{k}\left(p_2^{-k} - p_1^{-k}\right)\right)^{a_2 - 1}$$

$$\times \exp\left[-b_1\left(u+\frac{\alpha}{k}\left(p_1^{-k}-1\right)\right)\right]\times \exp\left[\frac{-b_2\alpha}{k}\left(p_2^{-k}-p_1^{-k}\right)\right]$$

$$\times \prod_{i=1}^{n}\left(\frac{1}{\alpha}\left[1+\frac{k\nu_i}{\alpha}\right]^{\frac{-(1+k)}{k}}\right)$$

3.9

for $k \neq 0$, while for $k = 0$:

$$\pi\left(\alpha,k|Y\right)\propto \left(u+\frac{\alpha}{k}\left(p_1^{-k}-1\right)\right)^{a_1-1}\times\left(\frac{\alpha}{k}\left(p_2^{-k}-p_1^{-k}\right)\right)^{a_2-1}$$

$$\times \exp\left[-b_1\left(u+\frac{\alpha}{k}\left(p_1^{-k}-1\right)\right)\right]\times \exp\left[\frac{-b_2\alpha}{k}\left(p_2^{-k}-p_1^{-k}\right)\right]$$

$$\times \prod_{i=1}^{n}\left(\frac{1}{\alpha}\left(\exp\left\{\nu_i \middle/ \alpha\right\}\right)\right)$$

3.10

where $\nu_i = y_i - u$.

Explicit analytical calculation of $\pi\left(\alpha,k|Y\right)$ (equations 3.9 & 3.10) is very difficult. However, using Markov Chain Monte Carlo techniques, direct simulation from a Markov chain whose equilibrium distribution is $\pi\left(\alpha,k|Y\right)$ is not only possible but simple (Coles and Tawn, 1996, Kuczera, 1999, Coles et al., 2003). As such the Gibbs Sampler (described in section 2.4.5) was used to generate the Markov chain for the GPD parameters.

The Gibbs sampler considers only univariate conditional distributions (Smith and Roberts, 1993). Consequently, the conditional distributions $\pi\left(\alpha|k\right)$, (when k is considered a constant) and $\pi\left(k|\alpha\right)$, (when α is considered a constant) were obtained from equation 3.9, as

$$\pi(\alpha|k) \propto \alpha^{a_2-1} \times \left(u + \frac{\alpha}{k}\left(p_1^{-k} - 1\right)\right)^{a_1-1} \times \exp\left(-\frac{b_1\alpha}{k}\left(p_1^{-k} - 1\right)\right) \times \exp\left(-\frac{b_2}{k}\left(p_2^{-k} - p_1^{-k}\right)\right) \times \prod_{i=1}^{n}\left(\frac{1}{\alpha}\left[1 + \frac{k\nu_i}{\alpha}\right]^{\frac{-(1+k)}{k}}\right)$$

3.11

and

$$\pi(k|\alpha) \propto \left(\frac{\left(p_2^{-k} - p_1^{-k}\right)}{k}\right)^{a_2-1} \times \left(u + \frac{\alpha}{k}\left(p_1^{-k} - 1\right)\right)^{a_1-1} \times \exp\left(-\frac{b_1\alpha}{k}\left(p_1^{k} - 1\right)\right) \times \exp\left(-\frac{b_2}{k}\left(p_2^{-k} - p_1^{-k}\right)\right) \times \prod_{i=1}^{n}\left(\frac{1}{\alpha}\left[1 + \frac{k\nu_i}{\alpha}\right]^{\frac{-(1+k)}{k}}\right)$$

3.12

It is these conditional distributions from which the Gibbs sampler generated the Markov Chain for the respective parameters of the GPD.

3.6 Predictive distribution and determination of extreme rainfall magnitudes

The Predictive distribution, $\Pr(Z|y)$, (Coles and Tawn, 1996) was estimated from the Gibbs sampler output by

$$\Pr(Z|y) = m^{-1}\sum_{j=1}^{m}\left[1 - \left(1 + k_j\,\frac{z-u}{\alpha_j}\right)^{-1/k_j}\right] \quad for \quad z > u$$

3.13

where (α_j, k_j) is the output of the j^{th} iteration of a sample of size m taken from the Gibbs sampler of the posterior distribution of (α, k).

The use of the Gibbs Sampler output ensures that a distribution of the parameters is used in the predictive distribution (equation 3.13) to estimate the rainfall magnitudes at various return periods as opposed to the standard methodology where a point estimate of each parameter would be used and as

such parameter uncertainty is accounted for reasonably much better than any point estimate could have been expected to.

The parameter vector (α_j, k_j) is one of the inputs for the program that was developed based on the estimate of the predictive distribution (equation 3.13) whose output is a series of estimates of rainfall magnitudes corresponding to a series of return periods from 1 year up to 200 years. The other input that was required for the estimation process were a series of values, z, all greater than the selected threshold for the particular rainfall station since equation (3.13) holds only for $z > u$. The tasks performed by this program are enclosed by the dotted lines in Figure 3.1. Starting with $z = u + 0.1$, the program computes the non-exceedance probability, $\Pr(Z|y)$ (equation 3.13) as well as the return period, $N(Z|y)$ (equation 3.14) corresponding to the individual values of z.

$$N(Z|y) = \frac{1}{1 - \Pr(Z|y)} \qquad 3.14$$

The program then increases that value of z by 0.1 and computes the corresponding non-exceedance probability and return period. It continues making increments of 0.1 to the value of z until it meets the termination criterion which is obtaining a return period of 200 years.

As such the rainfall magnitudes corresponding to various return periods were obtained and these were compared with the rainfall magnitudes for the same return periods obtained using an approach based on regional index storm as described by Smithers and Schulze (2003). The approach that Smithers and Schulze (2003) used is a regional rainfall frequency procedure which makes use of L-moments (described in section 2.5.7) for GEV parameter estimation. This program gives estimates of rainfall magnitudes, as well as the 90 % upper and lower bounds, corresponding to durations of 5 minutes up to 7 days and return periods ranging from 2 years to 200 years for rainfall stations in South Africa. The estimates of the 1-day duration rainfall magnitude estimates were

obtained from the program (Smithers and Schulze, 2003) for comparison with the estimates obtained using the Bayesian methodology.

CHAPTER 4 : RESULTS AND DISCUSSION

4.1 Preamble

Analysis of the rainfall data began with formulation of the joint prior distribution of the Generalized Pareto Distribution (GPD) parameters shown in equation 3.8 in which the threshold, u, and gamma parameters (a_1, b_1, a_2, b_2) had to be determined first. Section 4.2 gives a description of how the gamma parameters were obtained and these are presented in Table 4.2. The thresholds that were selected from the mean residual life plots of the rainfall data are presented in Table 4.3 which also shows the number of rainfall values that exceed the selected thresholds for each set of rainfall data. The Gibbs sampler was then used to generate 15 000 samples of the Generalized Pareto Distribution (GPD) scale and shape parameters from the posterior distribution. However, before these could be used to estimate rainfall magnitudes at various return periods using equation 3.13, the point at which the generated sequence of parameters reach equilibrium had to be determined so that all those parameter sets generated before that point are disregarded. This was so because the Gibbs sampler is initialized at some arbitrarily chosen starting value whose influence on the subsequently generated parameter sets diminishes only gradually up to such a point when any other sample in the sequence generated depends only on the previous value. As such, three Gibbs sampler sequences, each made up of 15 000 parameter sets, were generated using the data from the Hogsback(Bos) gauge station. Predictions of rainfall magnitudes for various return periods were made using the first 2 500, the second 2 500 and the third 2 500 parameter sets of each of the three sequences as shown in Figure 4.2, Figure 4.3, and Figure 4.4. This led to the decision to disregard the first 5 000 parameter sets of the Gibbs sampler in the subsequent estimation of the rainfall magnitudes for various return periods as it was evident from the Figures 4.2, 4.3 and 4.4 that any parameter set generated after the first 5 000 sets was under negligible influence from the initial arbitrarily chosen value. These estimates were compared with the estimates obtained by a program designed by Smithers and Schulze (2003) who took a regional index storm approach for design rainfall estimation. Their approach utilizes L-moments for the GEV parameter estimation. The details of this

comparison are shown in Table 4.4 and Table 4.5. On average the Bayesian estimates were greater than the corresponding regional index storm estimates by 63.2 % for the 100 year return period while for the 200 year return period, the average difference was 87.5%. The implication of this is that use of estimates based on Bayesian analysis would result in safer designs since parameter uncertainty is accounted for under the Bayesian approach.

4.2 Gamma Parameter Estimates.

Estimation of the Gamma parameters required that the median and the 90% upper confidence limit rainfall magnitudes corresponding to the 10- year and the 100- year return periods be computed. Table 4.1 shows the magnitudes of rainfall corresponding to these return periods for all the stations in this study. These were computed assuming that they follow the Generalized Extreme Value (GEV) distribution using the method of product (conventional) moments for parameter estimation.

Table 4.1: Estimates of the median and 90% upper confidence limit for the 10-year and 100-year rainfall estimate based on the GEV distribution.

Rainfall Station	10 year- Return		100 year- Return	
	Median (mm)	90% Upper Limit (mm)	Median (mm)	90% Upper Limit (mm)
SPRINGBOK (MUN)	64.360	73.858	103.462	112.961
MALMESBURY (MUN)	56.774	62.335	80.221	85.782
LADISMITH (MUN)	67.226	74.852	99.409	107.035
ABERDEEN (MUN)	76.384	90.866	130.764	145.246
DRIELINGSPAN	86.597	102.950	150.188	166.542
LEKKERVLEI	63.536	70.941	95.007	102.411
POORTJIE	99.477	110.125	145.540	156.189
HOGSBACK (BOS)	126.75	142.937	191.377	207.562
VILLA NORA (POL)	115.552	130.426	177.391	192.264
DE HOEK (BOS)	216.66	241.392	322.190	346.914
ZAMENKOMST	88.301	97.568	127.754	137.022
MOORSIDE	109.616	120.827	157.164	168.376
FLAGSTAFF (MUN)	133.685	159.041	230.048	255.404
ZILVERHOUT	108.455	122.123	163.443	177.111
KOFFIEFONTEIN (POL)	77.693	88.967	122.473	133.746

The median and 90 % upper confidence limit rainfall values in Table 4.1 were assumed to follow a Gamma distribution, $Ga(a,b)$. That is, the median and 90% upper confidence values for the 10- year return period were assumed to follow

a Gamma distribution with parameters, a_1 and b_1, while those values corresponding to the 100-year return period were assumed to follow a gamma distribution with parameters, a_2 and b_2. Then, bearing in mind that a median value has a probability of 0.5 and the 90% upper confidence limit has a probability of 0.9, the parameters of the Gamma function were obtained using the solver function in Microsoft Excel 2003 and are presented in Table 4.2.

Table 4.2: Gamma parameter estimates

Rainfall Station	Gamma Parameters			
	a_1	b_1	a_2	b_2
SPRINGBOK (MUN)	83.02	0.778	206.99	0.501
MALMESBURY (MUN)	182.54	0.312	357.73	0.224
LADISMITH (MUN)	137.48	0.490	293.54	0.339
ABERDEEN (MUN)	51.66	1.488	143.98	0.910
DRIELINGSPAN	52.05	1.674	148.77	1.012
LEKKERVLEI	130.52	0.488	284.63	0.334
POORTJIE	153.74	0.648	321.94	0.453
HOGSBACK (BOS)	109.48	1.161	242.74	0.789
VILLA NORA (POL)	107.83	1.075	246.87	0.720
DE HOEK (BOS)	135.93	1.598	293.38	1.099
ZAMENKOMST	159.73	0.554	327.40	0.391
MOORSIDE	167.90	0.654	338.29	0.465
FLAGSTAFF (MUN)	51.62	2.606	145.31	1.587
ZILVERHOUT	112.29	0.969	248.13	0.660
KOFFIEFONTEIN (POL)	85.75	0.909	205.93	0.596

4.3 Estimation of Thresholds and Generation of the Markov Chain of the Generalized Pareto Distribution (GPD) Parameters.

The values of the gamma parameters in Table 4.2 were then substituted into the equations of the conditional distributions $\pi(\alpha|k)$ and $\pi(k|\alpha)$ (3.11 and 3.12) leaving the GPD parameters, α and k, and the threshold, u, as the only unknowns. The threshold, u, was estimated from the mean residual life plot of the rainfall data for each of the rainfall stations included in this study. For instance, Figure 4.1 is the mean residual life plot for the Aberdeen (Mun) rainfall station. The value that was chosen as the threshold is the value for which linearity of the plot was first observed. For example, 55 mm was chosen as the threshold of the rainfall data of the Aberdeen (Mun) station because that is where linearity in the plot (Figure 4.1) is first observed. Table 4.3 shows the thresholds that were selected from the mean residual life plots of the rainfall

data from the 15 gauge stations as well as the number of rainfall values that exceeded the selected thresholds for each station. The use of the 'exceedances over a threshold' approach to analyze extreme events ensures that as much of the extreme data present in the rainfall records as possible is included as explained in section 3.2. This as noted in section 3.2 is in contrast with the use of annual maxima where only the maximum recorded value of a year is considered. This as explained in section 3.2 leaves out some records which may not be the maximum in years they were recorded but are greater than some of the annual maximums of other years on the record. From Table 3.1 and Table 4.3, 8 gauge stations have the numbers of exceedances greater than would have been the case if annual maxima were selected and used for analysis. 7 of the stations, however, have the number of exceedances less than the years on record. This might be as a consequence of considering only unpatched rainfall records for analysis and, therefore, having some years where some of the records are missing.

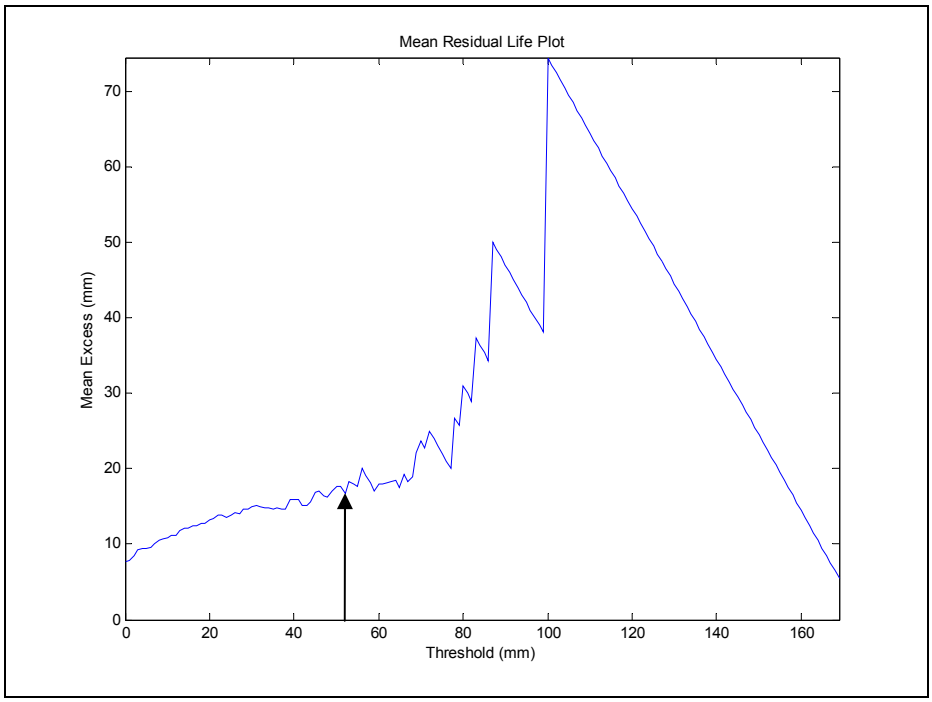

Figure 4.1: Mean Residue Life Plot of the rainfall data of the Aberdeen (Mun) rainfall station.

Table 4.3: Selected Thresholds (mm) and number of rainfall magnitudes that exceed the selected threshold.

Rainfall Station	Selected Threshold (mm)	Number of Exceedances
SPRINGBOK (MUN)	24.0	156
MALMESBURY (MUN)	30.0	154
LADISMITH (MUN)	28.0	185
ABERDEEN (MUN)	55.0	24
DRIELINGSPAN	32.5	35
LEKKERVLEI	27.0	129
POORTJIE	47.0	81
HOGSBACK (BOS)	74.0	89
VILLA NORA (POL)	50.0	101
DE HOEK (BOS)	125.0	36
ZAMENKOMST	42.5	110
MOORSIDE	65.0	45
FLAGSTAFF (MUN)	62.0	79
ZILVERHOUT	48.5	106
KOFFIEFONTEIN (POL)	35.5	154

Use was made of the algorithm developed based on the Gibbs sampling methodology whose description is in section 2.4.5 to generate a Markov Chain. The sampler was initialized at an arbitrary starting point and run for 15 000 iterations. Figure 4.2 shows plots of the rainfall magnitudes versus return periods produced using the three sets of the first 2500 parameter samples generated by the Gibbs Sampler and substituting these into equation 3.13. There is a difference in the plots and this can be attributed to the fact that the equilibrium distribution might not have been reached and expect that another set of Gibbs Sampler outputs would produce a curve quite different from the three plots shown. Figure 4.3 are plots produced by using three sets of the next 2500 parameter samples i.e $\theta = (\theta_{2501},...,\theta_{5000})$, generated by the Gibbs sampler where $\theta_i = (\alpha_i, k_i)$ and i is the iteration number. Again there is a difference in the plots which can be attributed to equilibrium not having been reached. Figure 4.4 shows plots produced based on three sets of the next 2500 parameter samples generated by the Gibbs Sampler i.e $\theta = (\theta_{5001},...,\theta_{7500})$. The plots are similar and as such equilibrium is assumed to have been reached after 5000 iterations. Subsequently, analysis was based on parameter vector $\theta = (\theta_{5001},...,\theta_{15000})$.

4.4 Comparison of the Prior and Posterior Distribution of the Generalized Pareto Distribution (GPD) parameters

The modification of the prior distribution by the rainfall data is illustrated by the plots of the prior and posterior univariate distributions of the GPD parameters in figures D-1a, D-1b and D-1c in Appendix D. The prior distributions of the parameters were obtained using equation 3.8 by varying only the parameter of interest with everything else fixed. The univariate posterior distribution was obtained in a similar manner using equation 3.9. The posterior distributions of the scale parameter for all the stations have standard deviations that are smaller than their corresponding prior distribution. While the standard deviations for the prior and posterior distributions of shape parameter are nearly the same with the difference being the values of the shape parameter at which the peaks occur. A possible explanation for these differences in the prior and posterior distribution plots could be attributed to the uncertainty involved in formulating the prior distribution and the model choice (the Generalized Pareto Distribution) as well as the effect of observed data on modifying the prior beliefs about the parameter distribution. The uncertainty in model choice arises from the assumption that an extreme rainfall distribution can be described explicitly by a mathematical expression (the Generalized Pareto Distribution). While this was done because the Generalized Pareto distribution is one of those distributions best suited for analysis of extreme processes (refer to section 2.2), it would not be reasonable to assume that the exceedances of rainfall over the selected threshold fit the Generalized Pareto distribution perfectly. Also as described in section 3.2 the prior was formulated based on gamma distributions (refer to equation 3.7) of the differences d_i defined in equations 3.5 and 3.6. The gamma parameters were estimated from the median and the 90% upper confidence limit rainfall magnitudes values that had been obtained using point estimates of the Generalized Extreme Value (GEV) parameters.

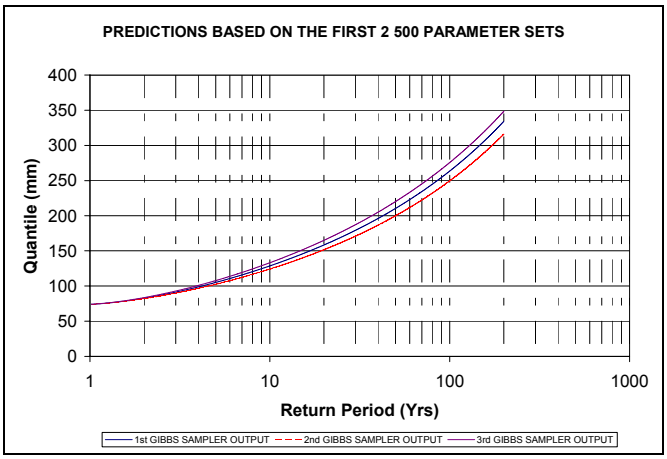

Figure 4.2: Estimates of rainfall magnitudes for various return periods based on the first 2 500 parameter sets of three Gibbs sampler outputs.

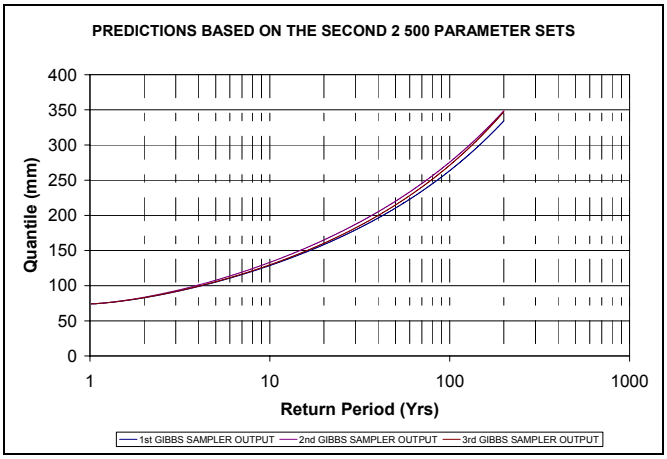

Figure 4.3: Estimates of rainfall magnitudes for various return periods based on the second 2 500 parameter sets of three Gibbs sampler outputs.

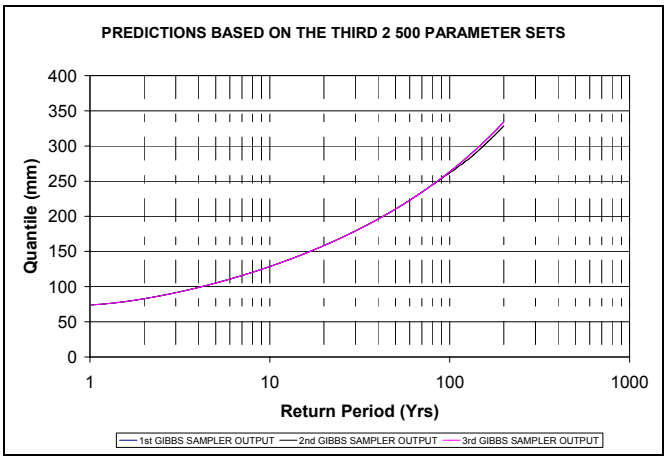

Figure 4.4: Estimates of rainfall magnitudes for various return periods based on the third 2 500 parameter sets of three Gibbs sampler outputs.

The Generalized Extreme Value (GEV) was chosen to model the annual maxima of the rainfall data. The use of point estimates of the GEV parameters does not completely take into account parameter uncertainty while the use of only annual maxima leaves out some extreme data values that may have been useful in the analysis. The power of the Bayesian approach is demonstrated by the fact that it recognizes these uncertainties and by incorporating the observed data which helps capture the actual nature of the hydrological process and therefore reduces the uncertainty. Therefore, the use of the available rainfall data in the form of the likelihood function has the effect of modifying the prior distribution resulting in a posterior distribution that is differently shaped and located.

4.5 Prediction of Rainfall Magnitudes for various Return Periods.

Having obtained the parameter vector $\theta = (\theta_{5001},...,\theta_{15000})$ from the Gibbs Sampler output, the predictive distribution was estimated by equation 3.13. The parameter vector θ is one of the inputs for the program developed based on the estimate of the predictive distribution (equation 3.13) whose output is a series of estimates of rainfall magnitudes corresponding to a series of return periods of 1 year up to 200 years. Figure 4.5 shows plots of estimates of rainfall magnitudes for various return periods obtained using the above methodology (Bayesian estimate) as well as the estimates of the same that were obtained using the regional storm index based approach that utilizes L-moments for design rainfall estimation (Smithers and Schulze, 2003).

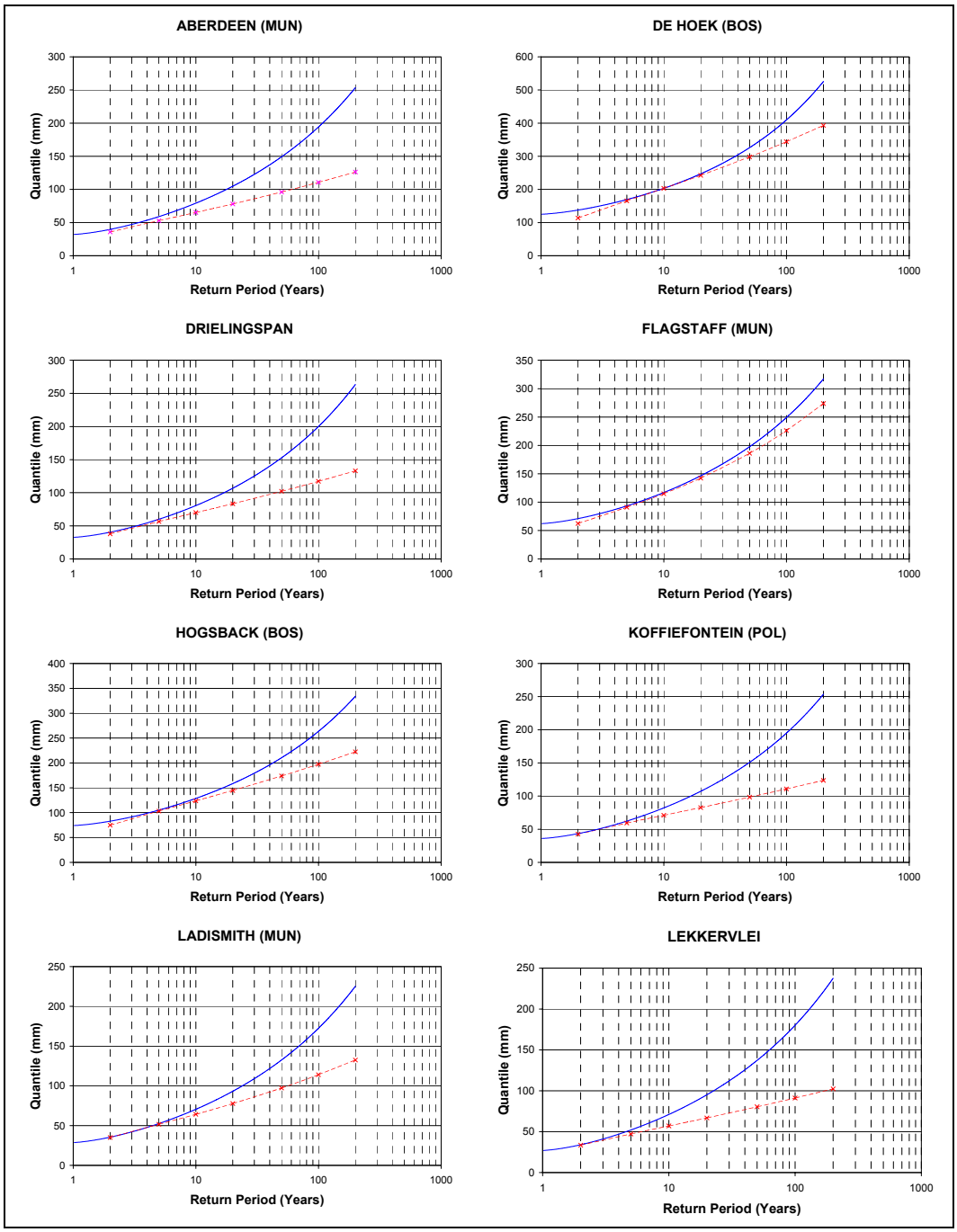

Figure 4.5a: Estimates of rainfall magnitudes (mm) at various return periods (Years) using the Bayesian approach and the Regional Index Storm approach.

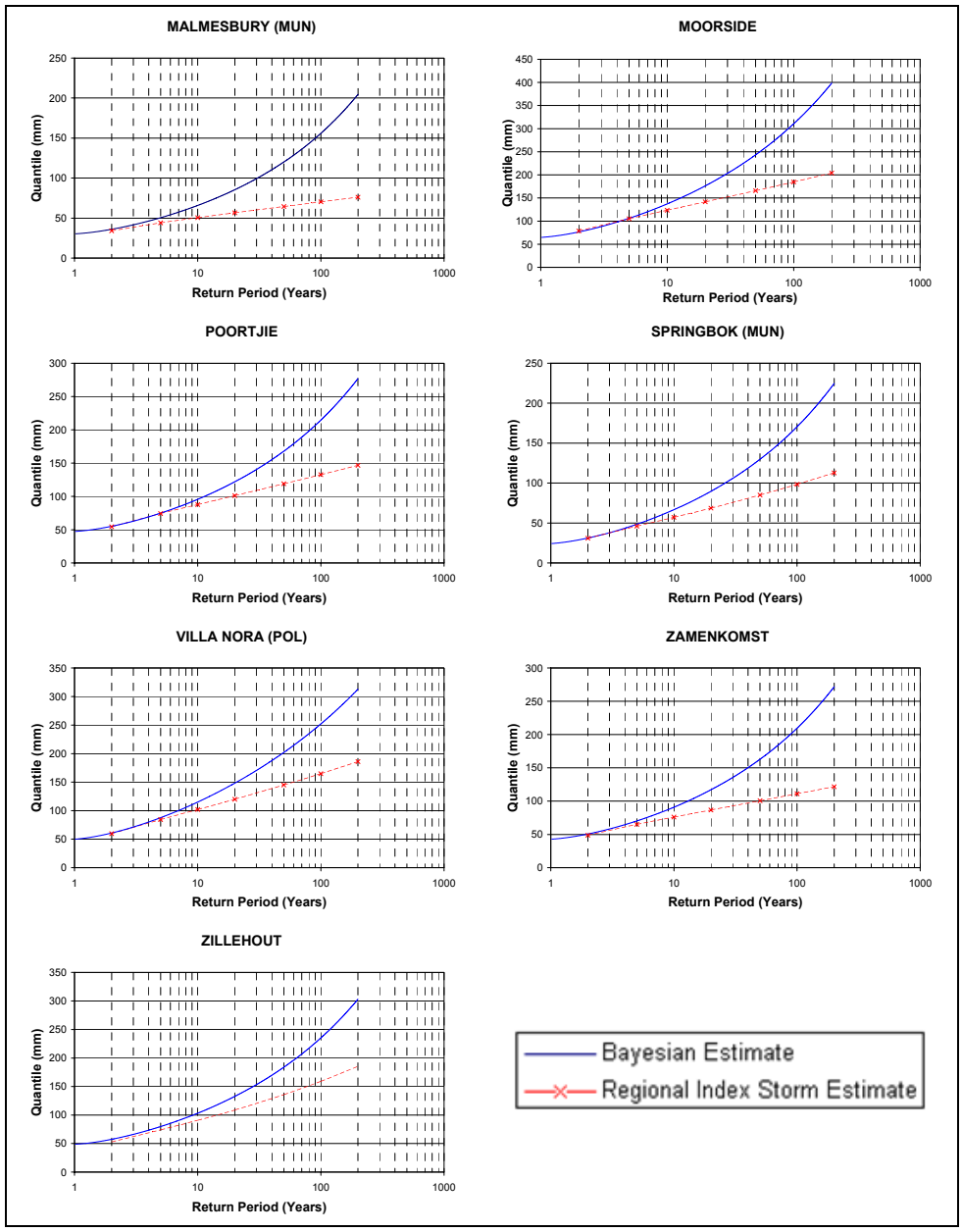

Figure 4.5b: Estimates of rainfall magnitudes (mm) at various return periods (Years) using the Bayesian approach and the Regional Index Storm approach.

4.6 Discussion

Figure 4.5 shows a comparison between the Bayesian estimates and the regional index storm estimates of rainfall magnitudes (return levels) in a plot of return level (mm) against return period (years), on a logarithmic scale, for all the stations that were considered in this study.

It is evident from Figure 4.5 that the estimates of rainfall magnitudes for the shorter return periods are reasonably similar. However, the regional index storm graphs have less curvature than the Bayesian graphs for all the stations, leading to differences in estimates that increased with the length of return periods. This is illustrated in Table 4.4 and Table 4.5 which show a comparison of rainfall magnitude estimates for the rain gauge stations for the 100-year and 200-year return periods respectively. For the 100-year return period, the average percentage increase in estimate is 63.2 % with the largest percentage increase being 121.9 % at Malmesbury (Mun) station and the smallest percentage being 10.2 % at Flagstaff (Mun) station. The average percentage increase at the 200- year return period is 87.5 % with Malmesbury (Mun) having the largest percentage increase of 168.7 % and Flafstaff (Mun), the smallest percentage increase of 15.9 %.

These results are similar to results that were obtained in other studies where the Bayesian approach to frequency analysis was used and results compared with those obtained using standard methods. For instance, using the Maximum Likelihood estimator and the Bayesian modes of inference, with the GEV distribution, Coles et al (2003) obtained a 549% difference in the estimates of the return period corresponding to the 410.4 mm rainfall value. This difference is much bigger than any obtained in this study. Kuczera (1999) obtained similar results in a flood frequency analysis with a plot of the natural log of flow against return periods resulted in the expected probability distribution curve being above the curve of the standard statistical analysis. The expected probability curve was obtained using the log-Pearson III distribution fitted to 30 years of gauged and censored data.

Table 4.4: Comparison of Rainfall Magnitude Estimates for 100-year Return Period

Rainfall Station	Rainfall magnitude (mm) for 100-yr Return Period		Percentage Increase
	Regional Index Storm	Bayesian Estimate	
SPRINGBOK (MUN)	98.5	170.6	73.20
MALMESBURY (MUN)	70.3	156.0	121.91
LADISMITH (MUN)	114	172.5	51.32
ABERDEEN (MUN)	110.6	194.2	75.59
DRIELINGSPAN	117.2	200	70.65
LEKKERVLEI	91.2	180.2	97.59
POORTJIE	132.7	214.7	61.79
HOGSBACK (BOS)	197.4	263.5	33.49
VILLA NORA (POL)	164.9	252	52.82
DE HOEK (BOS)	343.9	409.9	19.19
ZAMENKOMST	111.1	209.6	88.66
MOORSIDE	184.8	310.6	68.07
FLAGSTAFF (MUN)	226.2	249.2	10.17
ZILVERHOUT	159.1	235.2	47.83
KOFFIEFONTEIN (POL)	110.8	195.0	75.99

Table 4.5: Comparison of Rainfall Magnitude Estimates for 200-year Return Period

Rainfall Station	Rainfall magnitude (mm) for 200-yr Return Period		Percentage Increase
	Regional Index Storm	Bayesian Estimate	
SPRINGBOK (MUN)	112.9	224.8	99.07
MALMESBURY (MUN)	76.1	204.5	168.73
LADISMITH (MUN)	132.4	225.4	70.22
ABERDEEN (MUN)	126.3	253.5	100.71
DRIELINGSPAN	133.1	263.4	97.90
LEKKERVLEI	102.4	237.7	132.09
POORTJIE	146.7	277.0	88.80
HOGSBACK (BOS)	222.4	334.3	50.33
VILLA NORA (POL)	186	312.9	68.23
DE HOEK (BOS)	393	525.6	33.74
ZAMENKOMST	121.7	271.8	123.31
MOORSIDE	204.1	398.5	95.25
FLAGSTAFF (MUN)	273.7	317.1	15.86
ZILVERHOUT	185.3	303	63.52
KOFFIEFONTEIN (POL)	123.7	253.8	105.14

Bayesian analysis enables the expression of uncertainty when it expresses parameter estimators as distributions rather than a point estimate. The inclusion of this uncertainty in the analytical process i.e the use of the

distribution of parameters as opposed to the use of their point estimates may be what is responsible for higher rainfall magnitudes.

As a consequence of the higher estimates of rainfall magnitudes obtained by the Bayesian methodology for the various return periods, designs based on these estimates would have a risk of failure of structures that is greatly reduced since these structures would have to be built stronger. This is so since the design floods, which are usually considered when designing most civil engineering structures, increase as the design rainfall increases. This can be exemplified by the rational formula that relates rainfall amount to the runoff it generates. This implies a higher initial cost of erecting the structures while reducing the risk of flood damage. In addition, when planning for land-use, land is usually categorized accordingly based on risk of inundation. This is known as flood zoning (Faisal et al., 1999). This is usually followed by regulations that ensure that highly flood prone areas are spared from intensive capital investments. The use of the Bayesian estimates in the flood zoning exercise would result in larger expanses of land being considered flood prone. For instance, since the Bayesian estimate for the rainfall magnitude corresponding to any return period is larger than the corresponding estimate by the standard methodology, the estimated design flood and therefore the area of flood plain would be larger. The implication of this would be lower risk of flood damage to property and loss of lives. Since other flood mitigation measures such as evacuation of people from highly flood prone areas would extend to much wider areas.

While the Bayesian analysis used in this study enables parameter uncertainty to be accounted for, the methodology is longer, complex and involves more inputs than the standard methodology as shown in Figure 3.1. For instance, when formulating the joint prior distribution of the parameters, parameters to a gamma distribution were obtained from the median and the 90% upper confidence limit estimates of rainfall magnitudes at the 10- and 100 year return periods which in turn were estimated based on the Generalized Extreme Value (GEV) distribution using the conventional product moments for parameter estimation and yet this is just an intermediate step. Also, the gamma parameters are obtained by using the solver function in MS excel to solve two

simultaneous equations while the rest of the analysis is carried using a computer program that was developed using the Matlab programming language. While it was possible to have programmed the gamma parameter estimation routine in Matlab, the decision to use MS Excel solver to obtain the gamma parameters was because it was easy and gave estimates that were considered accurate. However, this, and the fact that the methodology is longer, may discourage the use of the Bayesian approach for routine analysis of rainfall data. The integration of all the aspects of the Bayesian methodology into a single application package, however, would make the application of the methodology easy and encourage its use.

CHAPTER 5 : CONCLUSIONS AND RECOMMENDATIONS

5.1 Conclusions

Design rainfall estimation is necessary for the planning and design of engineering projects. However, reliable estimates, especially of extreme events, remain a challenge in hydrology as the standard methods of estimation are increasingly failing to predict the scale and frequency of some extreme events. The failure can be attributed to the fact that some of these methods fail to account for parameter uncertainty in the estimation of model parameters. In this study, the effect of incorporation of uncertainties in parameter estimation on design rainfall estimates was investigated with the Bayesian approach being selected for use owing to its ability to account for parameter uncertainty in the analysis. The estimates of rainfall magnitudes corresponding to various return periods obtained using the Bayesian analysis in this study were generally higher than the corresponding regional index storm estimates of rainfall magnitudes. The estimates by the Bayesian and the regional index storm methods were reasonably similar for the shorter return periods, however, the differences between the estimates increased with the length of the return period. For instance, for the 100-year return period, the Bayesian estimates were higher by an average of 63.2 % with a range of 10.17 % to 121.9 %. At the 200-year return period the average difference was 87.5 %, with a range of 15.9 % to 168.7 %. These results are in agreement with results obtained in other studies where the Bayesian approach was compared with standard methods of frequency analysis.

5.2 Recommendations

The numerical intensity of the Bayesian approach to data analysis has contributed to lack of interest in the approach in the past. However, with the development of powerful computers and the recent advances in powerful approximation techniques, the numerical intensity of the approach in no longer a limitation. These approximation techniques, such as the Gibbs Sampler that was used in this study, are efficient and easy to implement and this in turn makes the implementation of the Bayesian approach easy. Studies that demonstrate the efficiency of these approximation techniques should be conducted as this will raise awareness about the availability of these methods.

In addition, more research should be conducted to improve these techniques as this would, further, encourage the use of the Bayesian methodology.

Although flood mitigation design is often wrought with many uncertainties, the considerable impact of incorporating uncertainties obtained here calls for a comprehensive review of the extreme rainfall estimation methods in South Africa and elsewhere in the world with more comparative studies being carried out so as to assess the reliability of the methods in use.

Having observed that the complexity of the methodology might discourage the routine use of Bayesian analysis, there is need for the development of a user-friendly package for the methodology.

This study also calls for a more comprehensive incorporation of uncertainties in engineering design especially where they are considered to be significant. An example is design flood determination using streamflow records because considerable uncertainties are encountered in measuring or determining streamflows during flood events. This study has demonstrated the effectiveness of the Bayesian approach and it is therefore recommended as a suitable method for incorporating uncertainties in design.

REFERENCES

Alexander, W. J. R., 2001, Flood Risk Reduction Measures. *Department of Civil Engineering, University of Pretoria,* ISBN 0 86979 872 3.

Behrens, C. N., Lopes, H. E., and Gamerman, D., 2004, Bayesian analysis of Extreme Events with Threshold Estimation. *Statistical Modeling. Vol. 4 No. 3 227-244*

Box, G.E.P and Tiao, G.C. 1973. Bayesian Inference in Statistical Analysis. Addison-Wesley Publishing Company.

Coles, S. G. and Tawn, J.A, 1996. A Bayesian Analysis of Extreme Rainfall Data. *J. R. Statist. Soc*. B, 4, 463-478

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

Cox, D. R., Isham V. S., and Northrop P. J. 2002. Floods: some probabilistic and statistical approaches. *Phil. Trans. R. Soc. Lond.* A (2002) 360, 1389-1408

Davison, A.C. and Smith, R.L. (1990) Model for exceedances over high thresholds. *Journal of the Royal Statistical Society, Series B*, 52, No. 3, 393-442.

Faisal, I.M., Kabir, M.R.and Nishat, A. 1999. Non-structural flood mitigation measures for Dhaka City, *Urban Water 1 (1999) 145 - 153*

Geman, S. and Geman, D., 1984 Stochastic relaxation, Gibbs distribution and Bayesian restoration of images. *IEE Transactions on Patterns Analysis and Machine Intelligence. 6: 721-741*

Grobler, R. R. 1996. A Framework for Modeling Losses arising from Natural Catastrophes in South Africa. *Magister Commercii in Actuarial Science Dissertation*, University of Pretoria.

Hastings, W. K. 1970. Monte Carlo sampling methods using Markov Chains and their applications. *Biometrika 57:* 97-109

Hosking, J. R. M., 1990. L-moment Analysis and Estimation of Distributions using Linear Combinations of Order Statistics. *J. R. Statist. Soc.* B, 52 No. 1, 105-124

Hosking, J. R. M., Wallis, J. R. and Wood, E. F., 1985, Estimation of the Generalized Extreme Value Distribution by the Method of Probability Weighted Moments. Technometrics, 27, 251-261.

Kuczera, G. 1999. Comprehensive at-site flood frequency analysis using Monte Carlo Bayesian Inference. *Water Resources Research, Vol. 35, No. 5,* 1551-1557

Kuczera, G. and Parent, E., 1998. Monte Carlo Assessment of Parameter Uncertainity in Conceptual Catchment Models: the Metropolis Algorithm. *Journal of Hydrology 211 (1998) 69-85*

Landwehr, J. M., Matalas, N. C. and Wallis, J. R. 1979, Probability weighted moments compared with some traditional techniques in estimating Gumbel parameters and quantiles. *Water Resources Research 15, 1055 - 1064*

Leadbetter, M. R. 1991. On the basis for ‘peaks over threshold modeling. *Statist. Prob. Lett.* 12, 357-362

Lynch, S. D., 2004, Development of a Raster Database of Annual, Monthly and Daily Rainfall for Southern Africa. *WRC Report No. 1156/1/04*. Water Research Commission, Pretoria, RSA

Midgley, D. C. and Pitman, W. V., 1978. A depth-duration-frequency diagram for point rainfall in Southern Africa. *HRU report 2/78. University of the Witwatersrand, Johannesburg, RSA*.

Moyeed, R. A. and Clarke, R. T., 2005. The use of Bayesian methods for fitting rating Curves, with case studies. *Advances in Water Resources 28 (2005) 807-818.*

National Environment Research Council (UK), 1975. *Flood Studies Report*.

Press, S. J., 1989. Bayesian Statistics: Principles, Models, and Applications. *John Wiley and Sons. New York.*

Smith, A. F. M. and G. O. Roberts., 1993. Bayesian computation via Gibbs sampler and related Markov chain Monte Carlo methods (with discussion). *J. Roy. Stat. Soc. Series B 55:* 3-23

Smithers, J.C. and Schulze, R.E., 2003. Design Rainfall and Flood Estimation in South Africa. *WRC Report No. 1060/01/03*. Water Research Commission, Pretoria, RSA.

Smithers, J.C. and Schulze, R.E., 2000a. Development and evaluation of techniques for estimating short duration design rainfall in South Africa. *WRC Report No. 681/1/00*. Water Research Commission, Pretoria, RSA.

Smithers, J.C. and Schulze, R.E., 2000b. Long duration design rainfall estimates for South Africa. *WRC Report No. 811/1/00*. Water Research Commission, Pretoria, RSA.

Yuan, C. and Druzdzel M. J., 2004. Importance Sampling Algorithms for Bayesian Networks: Principles and Performance.

APPENDICES

APPENDIX A

Example 1 – Simple Bayesian analysis illustration (Source: S. James Press, 1989)

A box containing eight parts is received from a supplier. In the past, 70 % of all such boxes have had zero defective parts; 20 % have had one defective part, and 10 % have had two defective parts. Therefore, it is assumed that all boxes containing eight parts will have 0, 1, or 2 defective parts. Three parts are selected at random from the box of eight, and one part is found to be defective. What is the probability the box of eight parts received from the supplier actually contained two defective parts?

Solution

Note that the number of defective parts X in a sample of size n of parts in a box approximately follows a binomial distribution, so

$$P\{X = x|\theta, n\} \equiv f(x|\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x} \qquad \text{A.1}$$

where θ denotes the probability of a defective part.

Here, for one defective part to be present, for example, when $n = 3$, the likelihood function is

$$f(1|\theta) = \binom{3}{1}(\theta)(1-\theta)^2 = 3\theta(1-\theta)^2 \qquad \text{A.2}$$

There are three possibilities: There are 0, 1, or 2 defective parts in a box of 8; or since θ denotes the probability of a defective part, then $\theta = 0, \frac{1}{8}, \frac{2}{8}$.

The prior belief, based on past experience is

Defective probability: θ	0	$\frac{1}{8}$	$\frac{2}{8}$
Probability mass function: $p(\theta)$	0.70	0.20	0.10

Moreover, for these three values of θ, the likelihood function is given by

Defective probability: θ	0	$\frac{1}{8}$	$\frac{2}{8}$
Likelihood function: $f(1\|\theta)$	0	0.287	0.42

Baye's theorem gives:

$$P\{\text{Box contains 2 defective parts} \mid \text{sample contains 1 defective}\}$$

$$= P\{\theta = 0.25 \mid X = 1\}$$

$$= \frac{f(1|0.25)p(0.25)}{f(1|0.25)p(0.25) + f(1|0.125)p(0.125) + f(1|0)p(0)}$$

$$= \frac{(0.42)(0.10)}{(0.42)(0.10) + (0.287)(0.20) + (0)(0.70)} = 0.424$$

$$= 42\%$$

In this example, while the prior probability of two defectives in the shipment is only 10 %, the posterior probability is four times as much (42 %).

APPENDIX B

Example 2 – Illustration of the sequential nature of Bayesian inference (Source: S. James Press, 1989).

There are mice of two colors, black and brown. The black mice are of two genetic kinds, homozygotes (BB) and heterozygotes (Bb), and the brown mice are of one kind (bb). It is known from established genetic theory that the probabilities associated with offspring from various matings are as follows:

Table B. 1: Probabilities for genetic character of mice offspring

Mice	BB (black)	Bb (black)	bb (brown)
BB mated with bb	0	1	0
Bb mated with bb	0	½	½
Bb mated with Bb	¼	½	¼

Suppose a 'test' mouse which is black and has been produced by a mating between two (Bb) mice. Using the information in the last line of the table, it is seen that, in this case, the prior probabilities of the test mouse being homozygous (BB) and heterozygous (Bb) are precisely known, and are $\frac{1}{3}$ and $\frac{2}{3}$ respectively. Given this prior information, suppose that the test mouse is now mated with a brown mouse and produced (by way of data) seven black offspring. Calculate the probabilities, posterior to the data, of the test mouse being homozygous (BB) and heterozygous (Bb) using Bayes' theorem.

Solution

Let θ denote the test mouse being (BB) or (Bb),

$$\theta = \begin{cases} 0 & (BB) \\ 1 & (Bb) \end{cases}$$

then the prior knowledge is represented by the distribution

$$p(\theta = 0) = \Pr(BB) = \tfrac{1}{3}, \qquad p(\theta = 1) = \Pr(Bb) = \tfrac{2}{3} \qquad \text{B. 1}$$

Also, if y denotes the offspring. This results in the likelihood of:-

$$l(\theta = 0 | y = 7 \ \ black) \propto \Pr(7 \ \ black | BB) = 1$$

$$l(\theta = 1 | y = 7 \ \ black) \propto \Pr(7 \ \ black | Bb) = \left(\tfrac{1}{2}\right)^7$$

B. 2

Recall from equations 2.8 and 2.9 that

$$Posterior \ \ (p(\theta | y)) = Likelihood \ \ (l(\theta | y)) \times \Pr ior \ \ (p(\theta))$$

It follows therefore that

$$p(\theta = 0 | y = 7 \ \ black) \propto \tfrac{1}{3}, \qquad p(\theta = 1 | y = 7 \ \ black) \propto \left(\tfrac{2}{3}\right)\left(\tfrac{1}{2}\right)^7$$

B. 3

Upon normalizing, the posterior probabilities are then

$$p(\theta = 0 | y = 7 \ \ black) = \Pr(BB | 7 \ \ black) = \tfrac{64}{65},$$

B. 4

$$p(\theta = 1 | y = 7 \ \ black) = 1 - \Pr(BB | 7 \ \ black) = \tfrac{1}{65}.$$

B. 5

Equations B. 4 and B. 5 represent the posterior knowledge of the test mouse being (BB) or (Bb).

The genetic characteristics of the offspring, the mating results of 7 black offspring changes our knowledge considerably about the test mouse being (BB) or (Bb), from a prior probability ratio of 2:1 in favor (Bb) to a posterior ratio of 64:1 against it.

As an illustration of the sequential nature of Bayes' theorem, suppose the 7 black offspring are viewed as a sequence of seven independent observations; then, letting $y' = (y_1, \ldots, y_7)$, the likelihood can be written

$$l(\theta | y = 7 \ \ black) = l(\theta | y_1 = black) \ \ \ldots \ \ l(\theta | y_7 = black)$$

B. 6

where

$$l(\theta | y_m = black) \propto \begin{cases} 1 & \theta = 0 \\ \tfrac{1}{2} & \theta = 1 \end{cases}, \qquad m = 1, \ldots, 7.$$

Applying equation (2.13), the changes in the probabilities of the test mouse being (BB) or (Bb) after the m^{th} observation, $m = 1,...,7$, are given in .

Table B.2: Probabilities for the test mouse being homozygous and heterozygous

Mice	Probabilities	
	$\theta = 0$ (BB)	$\theta = 1$ (Bb)
Initial	$\frac{1}{3}$	$\frac{2}{3}$
1st black	$\frac{1}{2}$	$\frac{1}{2}$
2nd black	$\frac{2}{3}$	$\frac{1}{3}$
3rd black	$\frac{4}{5}$	$\frac{1}{5}$
4th black	$\frac{8}{9}$	$\frac{1}{9}$
5th black	$\frac{16}{17}$	$\frac{1}{17}$
6th black	$\frac{32}{33}$	$\frac{1}{33}$
7th black	$\frac{64}{65}$	$\frac{1}{65}$

This shows the increasing certainty of the test mouse being (BB) as more and more black offspring are observed.

In most scientific applications, however, exactly known objective prior distributions are rarely available.

APPENDIX C

Derivation of the Joint Prior Distribution of the GPD Parameters.

Consider the Generalized Pareto Distribution of the form

$$G(y \mid k,\alpha,u) = \begin{cases} 1-\left(1+k\dfrac{y-u}{\alpha}\right)^{-1/k} & if \quad k \neq 0 \\ 1-\exp\{-(y-u)/\alpha\} & if \quad k = 0 \end{cases}$$

C.1

Considering a non-exceedance probability, p, then $G(y \mid k,\alpha,u)$ will be equivalent to the quantile $(1-p)$ and therefore, equation 3.1 becomes

$$(1-p) = 1-\left(1+k\frac{(y-u)}{\alpha}\right)^{-1/k}$$

C.2

What follows below is a rearrangement of equation C.2 to make y the subject.

Step 1

$$1-p-1 = -\left(1+k\frac{(y-u)}{\alpha}\right)^{-1/k}$$

Step 2

$$-p = -\left(1+k\frac{(y-u)}{\alpha}\right)^{-1/k}$$

Step 3

$$p^{-k} = \left(1+k\frac{(y-u)}{\alpha}\right)^{-k\times -1/k}$$

Step 4

$$p^{-k} = 1+k\frac{(y-u)}{\alpha}$$

Step 5

$$\left(p^{-k}-1\right) = k\frac{(y-u)}{\alpha}$$

Step 6

$$\frac{\alpha}{k}\left(p^{-k}-1\right)=\left(y-u\right)$$

Step 7

$$y=u+\frac{\alpha}{k}\left(p^{-k}-1\right)$$

C.3

Recall that the joint prior distribution of the GPD parameters is defined as

$$\pi(\alpha,k)\propto d_1\times d_2$$

C.4

where

d_i is defined as $d_i=y_i-y_{i-1}$ and $y_0=0$ since 0 is the physical lower bound of rainfall (Coles and Tawn, 1996). The differences d_i are assumed to follow a gamma distribution, hence:

$$d_1=y_1\sim Ga(a_1,b_1)$$

C.5

$$d_2=y_2-y_1\sim Ga(a_2,b_2)$$

C.6

The Gamma expressions in equations C.5 and C.6 are as written in equations C.7 and C.8 below.

$$d_1=Ga(a_1,b_1)=\frac{d_1^{a_1-1}.b_1^{a_1}.e^{-b_1d_1}}{\Gamma a_1}$$

C.7

$$d_2=Ga(a_2,b_2)=\frac{d_2^{a_2-1}.b_2^{a_2}.e^{-b_2d_2}}{\Gamma a_2}$$

C.8

Using equation C.3, d_1 and d_2 can be expressed in terms of the GPD parameters α, k, u yielding

$$d_1 = u + \frac{\alpha}{k}\left(p_1^{-k} - 1\right)$$

C.9

and

$$d_2 = \left[u + \frac{\alpha}{k}\left(p_2^{-k} - 1\right)\right] - \left[u + \frac{\alpha}{k}\left(p_1^{-k} - 1\right)\right]$$
$$= \frac{\alpha}{k}\left(p_2^{-k} - p_1^{-k}\right)$$

C.10

On substitution of the expressions for d_1 and d_2 obtained in equations C.9 and C.10 into equations C.7 and C.8, the Gamma expressions can be rewritten as

$$Ga(a_1, b_1) = \frac{b_1^{a_1}}{\Gamma a_1}\left(u + \frac{\alpha}{k}\left(p_1^{-k} - 1\right)\right)^{a_1 - 1} \times e^{-b_1\left(u + \frac{\alpha}{k}\left(p_1^{-k} - 1\right)\right)}$$

C.11

$$Ga(a_2, b_2) = \frac{b_2^{a_2}}{\Gamma a_2}\left(\frac{\alpha}{k}\left(p_2^{-k} - p_1^{-k}\right)\right)^{a_2 - 1} \times e^{\left(\frac{-b_2\alpha}{k}\left(p_2^{-k} - p_1^{-k}\right)\right)}$$

C.12

Substituting equations C.11 and C.12 into equation C.4, the joint prior distribution can be expressed in terms of the GPD parameters by following the following steps

$$\pi(\alpha, k) \propto \left[\frac{b_1^{a_1}}{\Gamma a_1}\left(u + \frac{\alpha}{k}\left(p_1^{-k} - 1\right)\right)^{a_1 - 1} .e^{-b_1\left(u + \frac{\alpha}{k}\left(p_1^{-k} - 1\right)\right)}\right]$$
$$\times \left[\frac{b_2^{a_2}}{\Gamma a_2}\left(\frac{\alpha}{k}\left(p_2^{-k} - p_1^{-k}\right)\right)^{a_2 - 1} \times e^{\left(\frac{-b_2\alpha}{k}\left(p_2^{-k} - p_1^{-k}\right)\right)}\right]$$

Since expressions $b_1^{a_1}/\Gamma a_1$ and $b_2^{a_2}/\Gamma a_2$ are constants in the above expression, it follows then that the joint prior distribution, $\pi(\alpha,k)$ can be expressed as

$$\pi(\alpha,k) \propto \left[\left(u+\frac{\alpha}{k}\left(p_1^{-k}-1\right)\right)^{a_1-1}.e^{-b_1\left(u+\frac{\alpha}{k}\left(p_1^{-k}-1\right)\right)}\right]\times\left[\left(\frac{\alpha}{k}\left(p_2^{-k}-p_1^{-k}\right)\right)^{a_2-1}.e^{\left[\frac{-b_2\alpha}{k}\left(p_2^{-k}-p_1^{-k}\right)\right]}\right]$$

C.13

APPENDIX D

Prior and posterior distributions of the GPD parameters

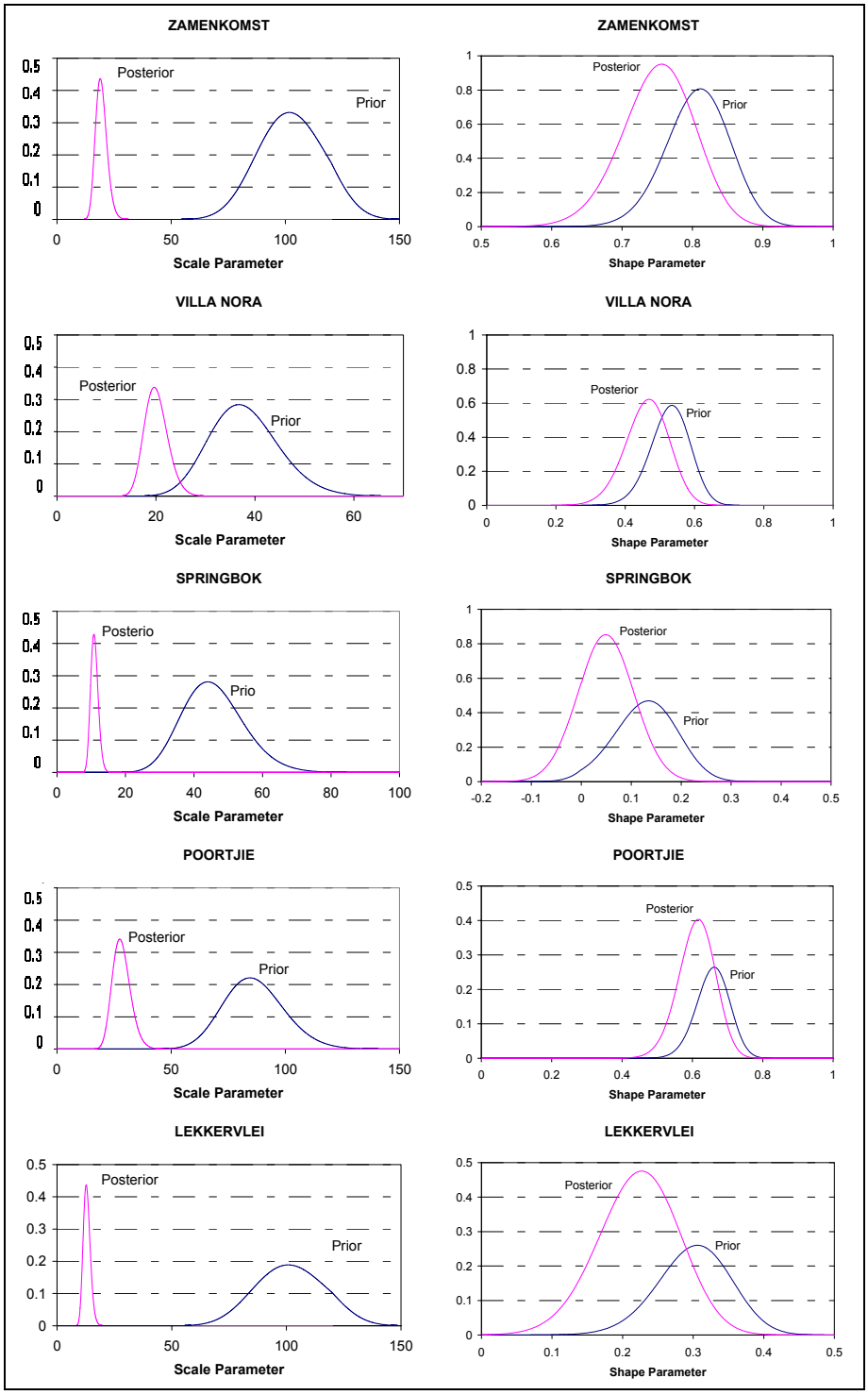

Figure D- 1a: Plots of the Prior and Posterior Distribution of the GPD Scale and Shape Parameters of the Rainfall Stations

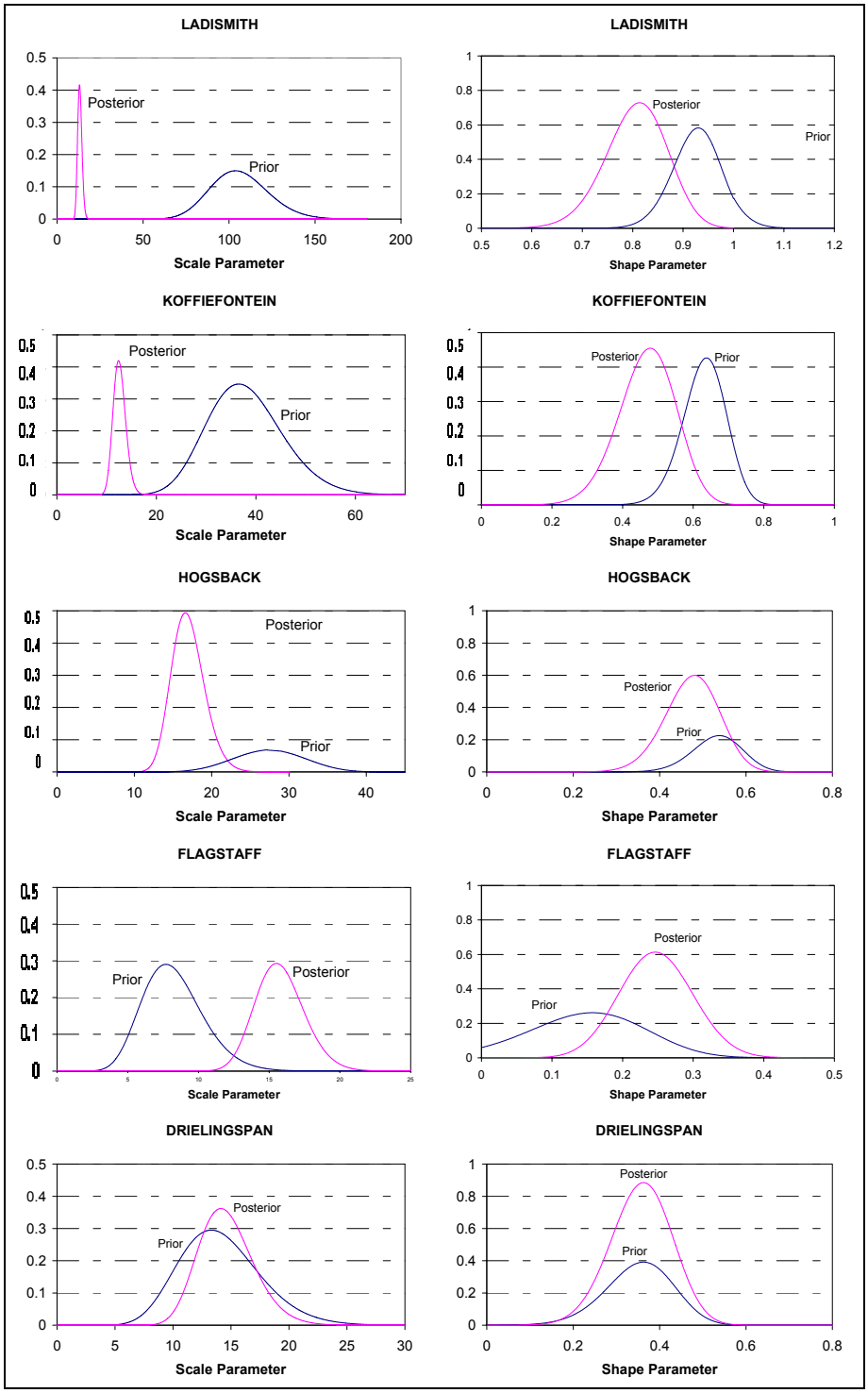

Figure D-1b: Plots of the Prior and Posterior Distribution of the GPD Scale and Shape Parameters of the Rainfall Stations

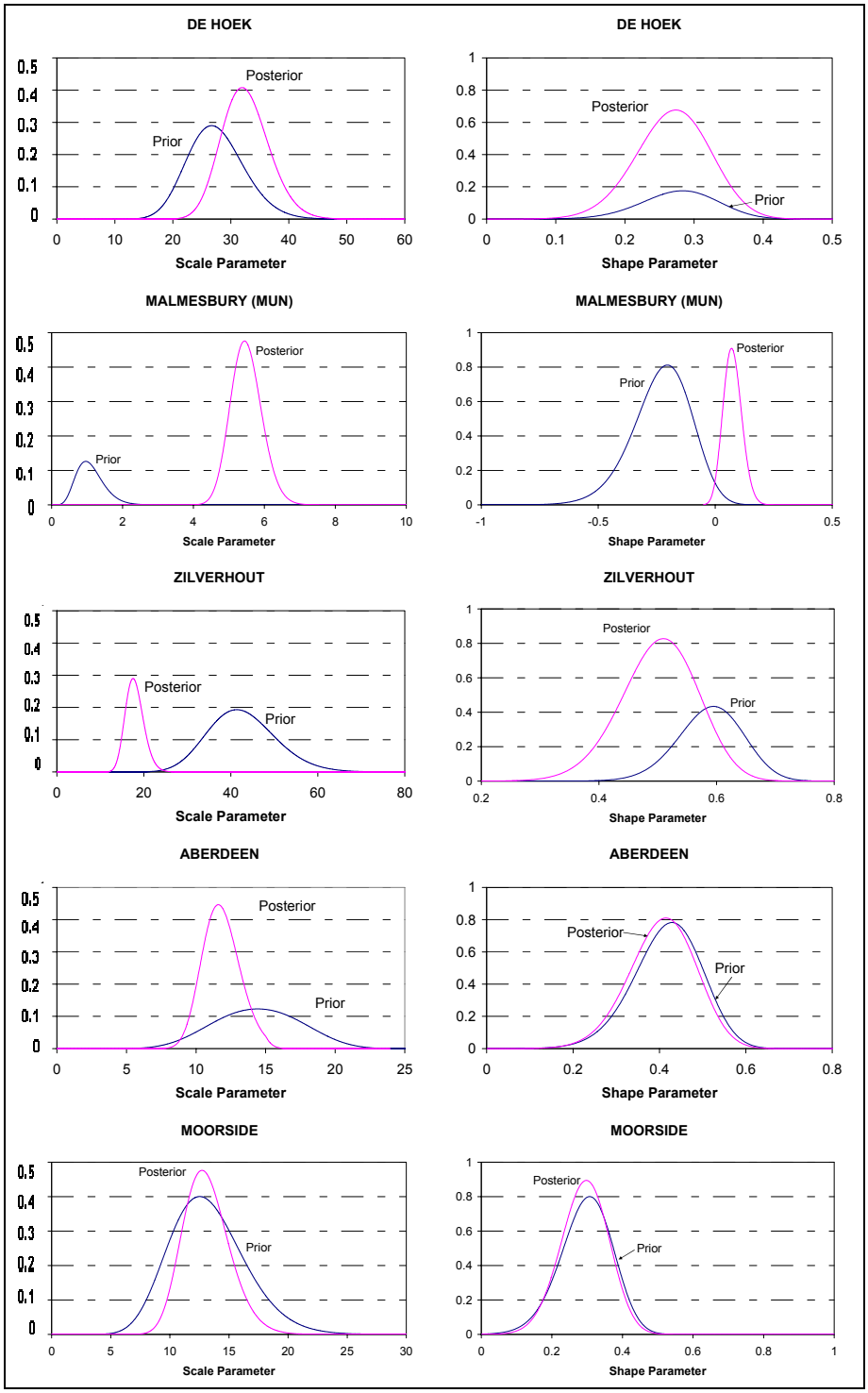

Figure D- 1c: Plots of the Prior and Posterior Distribution of the GPD Scale and Shape Parameters of the Rainfall Stations