DEVELOPMENT AND EVALUATION OF TECHNIQUES FOR ESTIMATING SHORT DURATION DESIGN RAINFALL IN SOUTH AFRICA

Jeffrey Colin Smithers

Submitted in partial fulfilment of the requirements
for the degree of Ph.D.

Department of Agricultural Engineering
University of Natal
Pietermaritzburg
South Africa

November 1998

ABSTRACT

The objective of the study was to update and improve the reliability and accuracy of short duration (≤ 24 h) design rainfall values for South Africa. These were to be based on digitised rainfall data whereas previous studies conducted on a national scale in South Africa were based on data that were manually extracted from autographic charts. With the longer rainfall records currently available compared to the studies conducted in the early 1980s, it was expected that by utilising the longer, digitised rainfall data in conjunction with regional approaches, which have not previously been applied in South Africa, that more reliable short duration design rainfall values could be estimated.

A short duration rainfall database was established for South Africa with the majority of the data contributed by the South African Weather Bureau (SAWB). Numerous errors such as negative and zero time steps were identified in the SAWB digitised rainfall data. Automated procedures were developed to identify the probable cause of the errors and appropriate adjustments to the data were made. In cases where the cause of the error could be established, the data were adjusted to introduce randomly either the minimum, average or maximum intensity into the data as a result of the adjustment. The effect of the adjustments was found to have no significant effect on the extracted Annual Maximum Series (AMS). However, the effect of excluding erroneous points or events with erroneous points resulted in significantly different AMS. The low reliability of much of the digitised SAWB rainfall data was evident by numerous and large differences between daily rainfall totals recorded by standard, non-recording raingauges, measured at 08:00 every day, and the total rainfall depth for the equivalent period extracted from the digitised data. Hence alternative techniques of estimating short duration rainfall values were developed, with the focus on regional approaches and techniques that could be derived from daily rainfall totals measured by standard raingauges.

Three approaches to estimating design storms from the unreliable short duration rainfall database were developed and evaluated. The first approach used a regional frequency analysis, the second investigated scaling relationships of the moments of the extreme events

and the third approach used a stochastic intra-daily model to generate synthetic rainfall series.

In the regional frequency analyses, 15 relatively homogeneous rainfall clusters were identified in South Africa and a regional index storm based approach using L-moments was applied. Homogeneous clusters were identified using site characteristics and tested using at-site data. The mean of the AMS was used as the index value and in 13 of the 15 relatively homogeneous clusters the index value for 24 h durations were well estimated as a function of site characteristics only, thus enabling the estimation of 24 h duration design rainfall values at any location in South Africa.

In 13 of the 15 clusters the scaling properties of the moments of the AMS were used to successfully estimate design rainfall values for duration < 24 h, using the moments of the AMS extracted from the data recorded by standard raingauges and regional relationships based on site characteristics. It was found that L-moments scaled better and over a wider range of durations than ordinary product moments.

A methodology was developed for the derivation of the parameters for two Bartlett-Lewis rectangular pulse models using only standard raingauge data, thus enabling the estimation of design values for durations as short as 1 h at sites where only daily rainfall data are available.

In view of the low reliability of the majority of short duration rainfall data in South Africa, it is recommended that the regional index value approach be adopted for South Africa, but scaled using values derived from the daily rainfall data. The use of the intra-daily stochastic rainfall models to estimate design rainfall values is recommended as further independent confirmation of the reliability of the design values.

I hereby declare that the work reported in this dissertation is my own original and unaided work except where specific acknowledgement is made.

J C Smithers

November 1998

ACKNOWLEDGEMENTS

A study of this nature is not possible without the assistance and support of many people and organisations. The author would like to thank the following people and to acknowledge their contributions to this study:

- Professor Roland Schulze, supervisor of this thesis, for the encouragement, guidance and opportunity to undertake this study,
- Professor Geoffrey Pegram, co-supervisor of this thesis, for his expert knowledge and assistance and confidence in my ability,
- Professor Peter Lyne, head of the Department of Agricultural Engineering, for his encouragement,
- the Water Research Commission for providing the funding for a project which enabled this study to be conducted,
- the University of Natal Research Fund for contributing towards the funding for this study,
- the SAWB, and in particular Mr Chris Koch, for making the digitised rainfall database available to this project,
- the CSIR, and in particular Mr Arthur Chapman, for providing the project with autographic rainfall charts and digitised rainfall data for Jonkershoek, Cathedral Peak and Mokobulaan,
- the Cape Town City Engineer's Department, for providing autographic rainfall charts for Athlone and Newlands,
- Mr Cobus Pretorius for his advice on digitising rainfall charts and to him and his team for the dedicated effort over many years of rainfall monitoring at the Cedara and Ntabamhlope research catchments,
- Mrs Pricilla Gouws for her invaluable assistance and dedicated digitising of many years of autographic rainfall charts,
- the Computing Centre for Water Research, for computing facilities and assistance,
- Dr Mark Dent for discussions on errors in the digitisation of autographic rainfall charts and his enthusiastic support,

- Mr Arne Kure for support in computing matters and for facilitating access to the Super Parallel Computer at the University of Potchefstroom,
- Mr Reinier de Vos for assistance in obtaining and manipulating rainfall data,
- Potchefstroom University for CHE and IBM for access to the Super Parallel Computing Facility for Academic Research,
- the four anonymous examiners, for their useful comments and compliments,
- Miss Kershani Chetty and Miss Marilyn Royappen for being willing helpers and for producing diagrams of station locations in South Africa, and most importantly,
- to my wife, Kary, for the love, encouragement and understanding throughout our lives together, for her assistance in the preparation of this document, for providing me with space to complete this study, for coping with being a single parent at times and for the many hours of family time that were lost during the study,
- and to our children, Jonathan and Bronwen, for their understanding when family time and play was sacrificed for "Dad's PhD", which was perhaps the most difficult part of this study.

TABLE OF CONTENTS

TABLE OF CONTENTS (continued)

TABLE OF CONTENTS (continued)

TABLE OF CONTENTS (continued)

Page

TABLE OF CONTENTS (continued)

LIST OF TABLES

LIST OF TABLES (continued)

LIST OF TABLES (continued)

LIST OF TABLES (continued)

LIST OF FIGURES

LIST OF FIGURES (continued)

LIST OF FIGURES (continued)

LIST OF FIGURES (continued)

LIST OF FIGURES (continued)

LIST OF FIGURES (continued)

LIST OF ACRONYMS

AIA	Average Intensity Adjustment
AMS	Annual Maximum Series
ANOVA	Analysis of Variance
BL	Bartlett-Lewis
BLRP	Bartlett-Lewis Rectangular Pulse
BLRPGM	Bartlett-Lewis Rectangular Pulse Gamma Model
BLRPM	Bartlett-Lewis Rectangular Pulse Model
CCWR	Computing Centre for Water Research
CSIR	Council for Industrial and Scientific Research
CV	Coefficient of Variation
DAEUN	Department of Agricultural Engineering, University of Natal
DDF	Depth-Duration-Frequency
EXEVNT	Exclude events if any errors within the event
EXPOINT	Exclude erroneous points from database prior to extraction of AMS
GOF	Goodness-of-Fit
IET	Inter Event Times
LIA	Lowest Intensity Adjustment
LM	L-Moments
MAP	Mean Annual Precipitation
MARE	Mean Absolute Relative Error
MBLRPM	Modified Bartlett-Lewis Rectangular Pulse Model
MIA	Maximum Intensity Adjustment
MNSRPM	Modified Neyman-Scott Rectangular Pulse Model
MOM	Method of Moments
NSRPM	Neyman-Scott Rectangular Pulse Model
PDS	Partial Duration Series
PWM	Probability Weighted Moment
RALM	Regional Average L-Moment
RGC	Regional Growth Curve
RI	Reliability Index
RLMA	Regional L-moment Algorithm
RS	Regional Slope i.e slope of log-transformed L-moment:duration relationship
RU	Rhodes University
SAWB	South African Weather Bureau
UZ	University of Zululand
WITS	University of the Witwatersrand
WRR	Winter Rainfall Region

CHAPTER 1

INTRODUCTION

Engineers and hydrologists involved in the design of hydraulic structures (e.g. culverts, bridges, dam spillways and reticulation for drainage systems) need to assess the frequency and magnitude of extreme rainfall events in order to generate design flood hydrographs. Many thousands of engineering and conservation design decisions involving millions of Rands of construction and which require accurate short duration ($\leq$ 24 h) design rainfall intensity information are made annually in South Africa. Depth-Duration-Frequency (DDF) relationships, which utilise recorded events in order to predict future exceedance probabilities and thus quantify risk and maximise design efficiencies are a key concept in the design of hydraulic structures (Schulze, 1984).

Estimates of design rainfall for durations shorter than one day were last comprehensively produced for South Africa in the early 1980s (Midgley and Pitman, 1978; Van Heerden, 1978; Adamson, 1981) and for selected stations in KwaZulu-Natal in the mid 1980s (Schulze, 1984). The objective of this study was to develop and apply new techniques, including regional approaches which have not been applied previously, for improving the estimates of short duration design rainfall values for South Africa. With longer available records from recording raingauges and an increased spatial density of short duration rainfall data, more reliable estimates of design storms may now be made than are currently used in practice.

Techniques used in single site frequency analysis are widely documented (e.g. Stedinger *et al.*, 1993). One of the requirements of frequency analyses is a collection of long periods of records. The short duration rainfall data available in South Africa have generally been recorded autographically and digitised into a computer compatible format. The record lengths of the available data are relatively short, with only 49 out of a total of 412 recording rainfall stations in South Africa having record lengths of 30 years or longer, and only 4

stations with record lengths of 50 years and longer. Thus the network of these stations with record lengths longer than 30 years is very sparse.

A regional approach to rainfall frequency analysis attempts to supplement the limited information available from the relatively short periods of record with regional information from surrounding stations. This approach is not new in frequency analysis, with many different techniques available. However, until recently, there has been very little consensus regarding the best technique to use. The development of a regional index-flood type approach to frequency analysis based on L-moments (Hosking and Wallis, 1993; Hosking and Wallis, 1997) has many reported benefits and has the potential of unifying current practices of regional design rainfall analysis.

The main objective of the project was to estimate short duration design rainfalls for South Africa. These were to be based on current digitised rainfall records, which were approximately 20 years longer than the manually extracted values used in previous studies conducted in the 1980s, and to utilise regional techniques to supplement the sparse distribution of recording raingauges and hence produce more reliable short duration design rainfall values than are currently available for South Africa.

A short duration rainfall database was established after a survey of the available data in South Africa. Some of the data were only available in chart form and have been subsequently digitised as part of this study. The organisation contributing the majority of the data to the database is the South African Weather Bureau (SAWB). Unfortunately the guidelines for routine digitisation spelt out by Dent and Schulze (1987) were not followed by the SAWB and numerous errors and inconsistencies in the SAWB data are evident. Thus approaches were developed to estimate short duration design rainfall values notwithstanding the limited reliability of the majority of the digitised rainfall data.

Three approaches to estimating design storms from the unreliable short duration rainfall database were evaluated. The first approach used a regional frequency analysis, the second investigated scaling relationships of the moments of the extreme events and the third

approach used a stochastic intra-daily model to generate synthetic rainfall series. A common theme in all three approaches is the development of techniques to estimate short duration design storms from the daily rainfall database, which contains rainfall data recorded manually at daily intervals, and is deemed to be more reliable than the short duration rainfall data.

The severity of the errors and the amount of missing short duration data varies from station to station. Hence the use of a regional approach will supplement information at sites which may have unreliable information with better information from within the region, assuming that not too many sites in the region have unreliable data. As part of the regional approach, homogeneous rainfall regions in South Africa were identified and a regionalised, index storm based frequency analysis using L-moments was adopted. Regionalisation was performed using site characteristics and tested independently using at-site data. For each of the homogeneous regions and for various durations, growth curves, which relate the ratio between design rainfall depths and an index storm to return period, have been developed. Regression equations, based only on site characteristics, have been derived to estimate the 24 h index storm for each region. Thus it is possible to estimate the 24 h index storm at a site which has no recorded rainfall data, and in conjunction with the regionalised growth curve, design storms may be estimated at any ungauged site in South Africa.

A second approach developed to overcome the limitations of the short duration rainfall database was to use the scaling properties of the moments of the extreme events in conjunction with the moments derived from the daily rainfall database to estimate short duration design storms at a particular location. In this respect, the use of L-moments instead of conventional moments were found to scale more linearly over a wider range of durations. Regionalised regressions to estimate the slope of the L-moment:duration relationships have been developed. Thus the L-moments for durations less than 24 h can be estimated using the L-moments computed from the daily data and regionalised regressions, thereby enabling short duration design storms to be estimated at any location in South Africa.

A third approach to estimating design storms from the generally unreliable database was to generate synthetic rainfall series using stochastic models and to estimate design storms from the synthetic series. Techniques have been developed to estimate the parameters for the models using moments and other information derived only from the daily rainfall data, thus utilising the relatively dense network of daily rainfall stations available in South Africa. Hence, at any site where a reasonable record of rainfall recorded at daily intervals is available, the parameters of the stochastic model can be derived and hence design storms for durations less than 24 h can be estimated from the synthetic rainfall series. The effect of short rainfall record lengths was investigated and the use of a stochastic rainfall model to overcome the limited available data is illustrated.

This document is divided into two parts. In Part A, the literature are reviewed and the theoretical framework is presented for the techniques used. The results from applications of the techniques and the development of new methods are presented in Part B.

Part A consists of Chapters 2 and 3. The international and South African literature pertaining to the estimation of design storms is reviewed in Chapter 2. Similarly, in Chapter 3 the use of stochastic models to generate synthetic rainfall series is reviewed.

Part B consists of Chapters 4 to 8. In Chapter 4 the establishment of a short duration rainfall database is described and the effect of the errors and unreliability of the data on the estimation of design storms is assessed. The application of the index-storm based regional frequency analysis algorithm in South Africa is described in Chapter 5. The scaling of L-moments in order to extrapolate design storms for a particular duration to another duration is discussed in Chapter 6 and results are presented for selected locations in South Africa. Similarly in Chapter 7, results are presented from the estimation of design storms at selected locations in South Africa using synthetic rainfall series generated by stochastic rainfall models. The various techniques developed and results obtained are discussed in Chapter 8 and the most appropriate techniques for estimating short duration design storms in South Africa are recommended.

PART A

LITERATURE REVIEW

In Part A the international and South African literature relevant to this study are reviewed. Techniques for the estimation of design storms are reviewed in Chapter 2 and the use of stochastic rainfall models to generate time series of rainfall, from which design storms can be estimated, are reviewed in Chapter 3.

CHAPTER 2

DESIGN STORM ESTIMATION

Estimates of high intensity rainfall are not only important for flood estimation and engineering design, but are also important in the estimation of soil loss and vegetation damage resulting from high intensity storms. It is thus desirable to express, in probabilistic terms and for different durations, the likelihood of different amounts of rain (Tomlinson, 1980). The results of under- or over-design of even small hydraulic structures such as farm dams or culverts results in considerable national waste of resources (Reich, 1961; Reich, 1963). Thus rainfall Depth-Duration-Frequency (DDF) relationships are a key concept in the design of hydraulic structures where a return period is selected according to the cost and significance of the structure. In order to minimise risk and maximise efficiency in design, statistical and probabilistic methods are thus applied to past events in order to predict the exceedance probability of future events (Schulze, 1984).

Adamson (1981) summarised the state of extreme value analysis as applied in hydrology as "copious, confusing and conflicting" and adds that many advances in extreme value analysis rarely find routine application. This results in the practising engineer relying on "well tried but often crude methodologies" (Adamson, 1981). Although much has been published on DDF studies since 1981 there still appears to be little consensus in the literature on preferred approaches to design storm estimation. However, the relatively recent developments in regional approaches to the estimation of DDF relationships at a point hold much promise for more general acceptance. Thus the objective of this chapter is to review and summarise some established and current, as well as new, procedures to estimate design storms. Both single at-site approaches (Section 2.1) and joint at-site and regional approaches (Section 2.2) to design storm estimation are reviewed. This is followed by a review in Section 2.3 of DDF studies in South Africa. Finally, a review of the use of scaling relationships is presented in Section 2.4 which includes results from both South African and international studies.

2.1 SINGLE SITE APPROACH

The objective of frequency analysis is to utilise a recorded sample of the hydrological variable in order to estimate future probabilities of occurrence (Cannarozzo *et al.*, 1995). Design rainfall values may be estimated by extracting either the Annual Maximum Series (AMS) or Partial Duration Series (PDS) from the rainfall data and then analysing the extracted series analytically or graphically (Hershfield, 1984). Both methods require the selection of a suitable probability distribution to be fitted to the extracted series. The analytical method requires a curve-fitting procedure and the graphical method requires the selection of an appropriate plotting position formula which assigns a probability of exceedance (P_e) to each value in the extracted series. By definition the relationship between the return period (T) and P_e is:

$$T = \frac{1}{P_e} \qquad \ldots 1$$

The estimation of design storms over a catchment commonly involves all or some of the following steps (Tomlinson, 1980; Canterford *et al.*, 1987a; Alexander, 1990; Griffiths and Pearson, 1993):

- DDF relationships are developed at each site by fitting probability distributions to the primary data series.
- Procedures are developed to determine short duration intensities from the daily raingauge network and thus to supplement the recording raingauge network.
- Relationships are developed to extrapolate from and interpolate between defined durations.
- Methods are deduced for interpolating between stations.
- Point to area relationships are derived to predict areal distribution of extreme rainfall.
- Procedures are developed to specify the temporal sequences of the design hyetograph.

- Guidelines are recommended to try and account for future climate change.

In order to develop the DDF relationships at each site, the following principal steps are commonly used (Cunnane, 1989; Nathan and Weinmann, 1991):

- A data set to be analysed is selected. This may either be the AMS or PDS.
- An appropriate probability distribution is selected.
- A parameter and quantile estimation method is selected.
- A scheme is chosen for joint use of at-site and, where available, regional data.

The above methods involve choices which are both descriptive, with the shape of the distribution resembling the observed sample's distribution, and predictive where quantile estimates are robust with small bias and standard error (Cunnane, 1989). Bias is defined as the difference in the estimated quantile and the population value. The above four steps are expanded on in the following sections.

2.1.1 Data Series

2.1.1.1 Annual maximum vs partial duration series

In order to perform an extreme value analysis, Sevruk and Geiger (1981) list necessary assumptions about the data as follows:

- the data are correct or, where necessary, have been corrected,
- the data series is consistent, homogeneous, stationary and independent,
- the length of record is sufficient to represent the population,
- the AMS or PDS series follow a particular distribution, and
- the estimates of the parameters of the distribution are unbiased.

According to Cunnane (1989) either one of the AMS or PDS may be used to derive the magnitude-return period relationship. The design values estimated using the two series converge beyond the 10 year return period (Reich, 1963), although Schulze (1998) has found that the convergence between the two series can occur at return periods as low as 5 years. The theoretical relationship between the return period from the AMS (T_{AMS}) and PDS (T_{PDS}) is

$$T_{AMS} = \frac{1}{1 - \exp(-1/T_{PDS})} \qquad ...2$$

Various opinions regarding the use of the AMS and PDS have been expressed in the literature. An advantage of using the AMS as compared to the PDS is that AMS are statistically independent if care is taken in the selection of events occurring over the end of the year, whereas statistical independence is not as easily achieved using the PDS (Cunnane, 1989). However, Adamson (1981) expressed the view that the popular use of the AMS rather than the PDS was due to the ease of use of the AMS and not on the theoretical efficiency in characterising extreme value time series. The use of the AMS may, in the case of short records, result in a considerable loss of information for the estimation of rainfall probabilities.

Stedinger *et al.* (1993) report that the use of PDS overcomes the objection that large events may be excluded when they are not the largest event in a year and design estimates based on the PDS should, if the arrival rate of events is large enough, yield more accurate estimates of quantiles than estimates based on the AMS. A disadvantage of the PDS is that the events selected have to be independent and the PDS analysis is more complicated than analysis using the AMS (Stedinger *et al.*, 1993).

2.1.1.2 Record length

Limited length of available records makes it impossible to conclusively select a distribution that could consistently provide adequate rainfall frequency estimates for return periods much greater than the period of record (Richards and Wescott, 1987) and a small sample may define a distribution which is markedly different to the parent population (Schulze, 1980; Oyebande, 1982). The lengths of record used in some rainfall frequency studies reported in the literature are listed in Table 1. As evident in Table 1, the minimum record length of 10 years suggested by Viessman *et al.* (1989) has generally been adhered to in most studies.

Schulze (1984) questioned the significance of the period of available record on the extreme events recorded and hence the design values. This issue was addressed by Hogg (1991; 1992) who used a moving window ranging from 10 to 40 years to estimate the 100 year return period event and compared the results to the 100 year return period event computed from the entire data set. In addition, Hogg (1991) used an expanding window which used a window from the starting point to the year in question. The expanding window estimate of the 100 year event showed some trends at particular stations in Canada, but Hogg (1991) concludes that these trends reflect natural climate variations and sampling variability, as the trends were not spatially (i.e. between stations) consistent. Using the moving window approach Hogg (1991) demonstrated that 20 years of data are not stable enough to estimate the 10 year return period event, while Hogg (1992) concluded that even a 40 year period of record is insufficient to estimate the 100 year return period event. Thus, Hogg (1992) postulates that the assumptions of stationarity and homogeneity of the AMS of rainfall are seldom valid and suggests that a regional approach may improve the frequency analysis of extreme rainfall events.

Table 1 Record lengths used in some rainfall frequency studies

Reference	Location	Record Length (years)
Van Wyk and Midgley (1966)	South Africa	5-26
Canterford and Pierrehumbert (1977)	Australia	> 12
Midgley and Pitman (1978)	South Africa	5 - 38
Oyebande (1982)	Nigeria	5 - 30
Sendil and Sahil (1987)	Saudia Arabia	10 - 20
Schaefer (1990)	USA	mean ≈32
Kothyari and Garde (1992)	India	10 - 53
Cannarozzo *et al.* (1995)	Sicily	10 - 45 (mean=23)

2.1.1.3 Errors and missing data

Raingauge malfunctioning and rainfall processing errors are inherent in rainfall data. The volume of raw data often precludes the manual editing of the data and missing data may be in-filled using relationships previously established at the site (Aron *et al.*, 1987), or rules may be established to exclude the data from the analysis should defined thresholds of allowable missing data be exceeded (Canterford and Pierrehumbert, 1977).

Weddepohl (1988) discusses problems associated with short duration rainfall data and their availability in South Africa. Some of the common errors in digitised data include inherent raingauge malfunctions, raingauge operator errors, errors in transposition of data from charts into computer compatible format and unrealistically lumped station data when a station is relocated within a period of record. Other problems associated with the data are the spatial density and distribution of raingauges, the fact that the standard rain day ends at 08:00 whereas the digitised data are continuous, the length of available records and the presence of outliers.

Errors are apparent when different rainfall depths are recorded at the same site using different types of raingauges. Differences are common between rainfall recorded at daily intervals and rainfall recorded continuously and aggregated to the same period as the daily rainfall. Thus the New Zealand Meteorological Service and the National Water and Soil Conservation Organisation have similar data editing procedures which contain internal consistency checks and inter-site comparisons and recording raingauges are scaled to bring them into agreement with total rainfall recorded by the check gauges (Tomlinson, 1980).

Guttman (1993), in a probabilistic analysis of monthly totals of rainfall in the USA using L-moments, recognised and accepted that there were still possible errors in the data, but did not attempt to correct or in-fill the missing data. This decision was based on Hosking's (1990) assertion that asymptotic biases of L-moments ratios are negligible for sample sizes greater that 20.

2.1.1.4 Outliers

It is generally accepted that outliers in rainfall data are the result of:

- the occurrence of a meteorological phenomenon different to those which caused all the other events, or
- a rare occurrence of a meteorological phenomenon similar to which has occurred previously, or
- incorrect observations or keying in of data (Tomlinson, 1980).

The phenomenon that data may not arise from the same population (distribution) has led to the use of the two-component extreme value distribution by, *inter alia,* Rossi *et al.* (1984), Versace and Rossi (1985), Arnell and Beran (1987), Pegram and Adamson (1988) and Cannarozzo *et al.* (1995).

Outliers are commonly identified by the degree of deviation from their plotted positions on the frequency curve, by their ratio to the mean, by comparison to other records in the region of study or if the equivalent return period assigned to an event is much longer than the length of the series (Wang, 1987). Statistical tests, such as those used by Pilgrim and Doran (1987), can be developed to identify high and low outliers. These generally relate deviations about the mean in log-space to identify an outlier. Tomlinson (1980) suggested three approaches to dealing with outliers:

- Exclude the event and recalculate the parameters of the probability distribution.
- If the event is found to be drawn from a non-homogeneous population, then exclude the event.
- Include the event and select a more appropriate distribution, fitting technique or plotting formula.

Cunnane (1989) expressed the opinion that outliers should be retained if an efficient parameter estimation method is used, as the effect of the outliers would then not be significant. In Australia, guidelines for the treatment of outliers is subjective and the probable cause of the event, the prior belief and statistical evidence are taken into account. The omission or deletion of a data point is taken as an extreme step (Pilgrim and Doran, 1987). According to Stedinger *et al.* (1993) the thresholds used to define high (X_H) and low outliers (X_L) in log space are

$$X_{H,L} = \overline{X} \pm K_n S \qquad ...3$$

where

$\overline{X}$ = mean of the log-transformed data,

S = standard deviation of log-transformed data,

n = sample size, and

$$K_n = 3.345\sqrt{\log(n)} - 0.4046\log(n) - 0.9043\,. \qquad ...4$$

2.1.1.5 Conversion of fixed time interval value to true maxima

When converting values calculated at specific times of the day to independent durations of the same length, conversion factors have to be used (Alexander, 1990). The conversion factors are dependent on the duration in question and various values have been proposed. For example, the factors recommended to convert the 1 day (fixed time) to 24 h continuous maxima are 1.13 in the USA (Hershfield, 1962), 1.06 in the UK (NERC, 1975), 1.13 (Alexander, 1978) and 1.11 (Adamson, 1981) in South Africa. Schulze (1984), using a digitised database, showed that in South Africa the conversion factor varies regionally and, at some locations, with return period with variations of up to 20% evident. More recently, Dwyer and Reed (1995) show that, based on theoretical considerations, the correction factor should be 1.33, but recommend a value of 1.16, which is based on rainfall data from the United Kingdom and Australia.

2.1.2 Selection of a Probability Distribution

The question of which probability distribution to adopt and methods of selecting the most appropriate distribution has received considerable attention in the literature, particularly for flood frequency estimation and to a lesser extent for rainfall frequency estimation. The choice is particularly important when estimating extreme events with return periods greater than the length of record (Canterford and Pierrehumbert, 1977; Chow *et al.*, 1990; Karim and Chowdhury, 1995). Cunnane (1989) reports that the choice is often based on factors such as the probability distribution being widely accepted, simple, easy to apply, consistent, theoretically well founded and documented, but concedes that theoretical arguments alone cannot identify the best distribution. Schulze (1984) postulates that the choice of distribution may be less important than other factors such as whether manually extracted or digitised data are used, the stationarity of the data and the method of fitting the distribution to the data. Cunnane (1989) expresses the opinion that the consequence of using the wrong form of the distribution is over and under design of hydraulic structures.

Since the exact probability distribution of the population is not known, it is required to select a reasonable and simple distribution to describe the phenomenon of interest (Stedinger *et al.*, 1993). The choice of distribution should take into account both descriptive abilities, to ensure that the shape of the distribution resembles the observed sample's distribution, and predictive abilities, which implies that the quantile estimates of possible candidate distributions are robust with small bias and standard errors (Cunnane, 1989; Cannarozzo *et al.*, 1995). This view was also expressed by Pegram and Adamson (1988), who advocate using a "theoretically and intuitively correct model" rather than a best-fit model, which may be a hazardous strategy for extrapolation. Chow and Watt (1990) express the opinion that no deductive reasoning or goodness-of-fit tests can arrive conclusively at a single correct/appropriate distribution. In addition, much uncertainty is inherent in the estimation of parameters and hence quantile estimates. Therefore Chow and Watt (1990) believe that it is necessary to use an expert system which mimics heuristics used by experts. In the light of the instability of design rainfall events, Hogg (1991) questions the selection of the "best" probability distribution to use.

The probability distributions investigated and used in selected rainfall frequency studies both in South Africa and internationally are listed in Table 2. From Table 2 and as reported by Stedinger *et al.* (1993) the EV1, LP3 and GEV probability distributions are commonly used for short-duration rainfall probability analysis. In South Africa the EV1 distribution has been extensively used in rainfall DDF studies, even though Adamson (1978) notes that the fixed skew of 1.13 inherent in the EV1 distribution is "a considerable limiting assumption". Although limited use of the GEV distribution in rainfall frequency analysis is reported in Table 2, the GEV distribution is extensively used in flood frequency analyses (Cunnane, 1989) and the use of the EV1, EV2 and EV3 distributions and the integrated GEV distribution is growing in the application of frequency analysis (Raynal-Villasenor and Acosta, 1995). According to Wallis and Wood (1985) the GEV distribution outperformed the LP3 in a regional analysis even when the samples used were generated by an LP3 distribution. The selection of an appropriate frequency distribution for South Africa is described in Chapter 5, Section 5.5.

Table 2 Summary of probability distributions used in selected rainfall frequency studies (See Table 3 for explanation of abbreviations)

Reference	Location	Probability Distribution	
		Investigated	Recommended/Used
South Africa			
Reich (1963)	SA		EV1
Van Wyk and Midgley (1966)	SA		EV1
Bergman and Smith (1973)	Western Cape		EV1
Midgley and Pitman (1978)	SA		LEV1
Adamson (1978)	SA		EV1
Schulze (1980)	SA		EV1
Adamson (1981)	SA		LN3
Schulze (1984)	KwaZulu-Natal		EV1, LN2, LP3
Pegram and Adamson (1988)	KwaZulu-Natal		TCEV
Weddepohl (1988)	SA		LN2
Smithers (1996)	SA	LN2, LN3, LP3, PE3, LP3, EV1, LEV1, GEV, GPA, GLO, WAK	GEV
International			
NERC (1975)	UK		GEV
Canterford and Pierrehumbert (1977)	Australia	LN2, EV1, GEV, double LN2 , mixed distribution	mixed distribution
Tomlinson (1980)	New Zealand		EV1
Hershfield (1982)	USA		EV1
Oyebande (1982)	Nigeria		EV1
Pescod and Canterford (1985)	Australia		LN2
Aron *et al.* (1987)	USA		LP3
Richards and Westcott (1987)	USA	PE3, LP3, GAM, EV1	EV1

Reference	Location	Probability Distribution	
		Investigated	Recommended/Used
Canterford *et al.* (1987a)	Australia		LP3
Canterford *et al.* (1987b)	Australia		LP3 / LN2
James *et al.* (1987)	India		EV1
Sendil and Salih (1987)	Saudia Arabia		EV1
Ferreri and Ferro (1990)	Sicily		EV1
Schaefer (1990)	USA		GEV
Shuy (1990)	Singapore		EV1
Buishand (1991)			GEV
Griffiths and Pearson (1993)	New Zealand		EV1 (local) KAP (regional)
Naghavi *et al.* (1993)	USA		LP3
Guttman (1992)	USA	LP3, GEV, LN3	LP3
Cannarozzo *et al.* (1995)	Sicily		TCEV

Table 3 Abbreviations used for probability distributions

Abbreviation	Probability Distribution	Abbreviation	Probability Distribution
EV1	Extreme Value Type I (Gumbel)	LN3	3 parameter Log-Normal
GAM	Gamma	LP3	Log-Pearson Type III
GEV	General Extreme Value	PE3	Pearson Type III
GPA	Generalised Pareto	LEV1	Log-EV1
GLO	Generalised Logistic	TCEV	Two Component Extreme Value
KAP	Kappa	WAK	Wakeby
LN2	2 parameter Log-Normal		

2.1.3 Parameter Estimation

The fitting of a distribution to a data set provides a compact and smoothed representation of the frequency distribution revealed by the limited available data and enables the systematic extrapolation to frequencies beyond the range of the data set (Stedinger *et al.*, 1993).

2.1.3.1 Fitting procedures

Some approaches available for estimating the parameters of a selected distribution are listed, with some comments, in Table 4. The use of L-moments to fit distributions has received extensive coverage in the recent literature (e.g. Wallis, 1989; Hosking, 1990; Pearson *et al.*, 1991; Gingras and Adamowski, 1992; Guttman, 1992; Pilon and Adamowski, 1992; Guttman, 1993; Guttman *et al.*, 1993; Lin and Vogel, 1993; Vogel and Fennessy, 1993; Vogel *et al.*, 1993a; Vogel *et al.*, 1993b; Wallis, 1993; Gingras and Adamowski, 1994; Zrinji and Burn, 1994; Hosking, 1995; Hosking and Wallis, 1995; Karim and Chowdhury, 1995; Hosking and Wallis, 1997). In addition, L-moments are reported to have advantages when compared to other techniques and hence are reviewed in the following section.

2.1.3.2 L-moments

While being similar to ordinary product moments, the purpose of L-moments and Probability Weighted Moments (PWMs) is to summarise theoretical probability distributions and observed samples (Vogel *et al.*, 1993a). Hence L-moments can be used for parameter estimation, interval estimation and hypothesis testing.

L-moments have several important advantages over ordinary product moments (Vogel *et al.*, 1993b). In order to estimate the sample variance and sample skew, ordinary product moments require the squaring and cubing of the observations respectively. Sample

estimators of L-moments are linear combinations of the ranked observations and do not require squaring and cubing of the observations. Thus L-moments are subject to less bias than ordinary product moments (Wallis, 1989; Pearson *et al.*, 1991; Vogel *et al.*, 1993a; Karim and Chowdhury, 1995).

Table 4 Summary of methods used for parameter estimation (Cunnane, 1989; Lin and Vogel, 1993; Stedinger *et al.*, 1993)

Method	Comment
Moments (MOM)	• easy to apply and simple to use • not suitable for distributions with more than 3 parameters
Maximum Likelihood Procedure (MLP)	• good statistical properties in large samples • often cannot be reduced to simple formulae, so are estimated using numerical methods • solution not always possible
L-Moments (LM) / Probability Weighted Moments (PWM)	• easy to apply • almost as efficient as MLP, particularly in small samples • easily used in regional analysis • LM more reasonable and reliable than MOM
Bayesian Inference (BI)	• combines prior information and regional hydrological information with the likelihood function • allows explicit modelling of uncertainty in parameters
Non-Parametric	• an advantage is that they do not assume a particular family of distributions • more robust, but less efficient than parametric methods • have not seen much use in practice and are rarely used officially

L-moments, as defined by Hosking (1990), are linear combinations of PWMs. Greenwood *et al.* (1979) summarise the theory of PWMs. Unbiased sample estimates for the first four PWMs can be computed from Equation 5 (Stedinger *et al.*, 1993; Vogel and Fennessy, 1993).

$$b_0 = \frac{1}{n}\sum_{j=1}^{n} x_j \qquad ...5a$$

$$b_1 = \frac{1}{n}\sum_{j=1}^{n-1}\left[\frac{(n-j)}{n(n-1)}\right]x_j \qquad ...5b$$

$$b_2 = \frac{1}{n}\sum_{j=1}^{n-2}\left[\frac{(n-j)(n-j-1)}{n(n-1)(n-2)}\right]x_j \qquad ...5c$$

$$b_3 = \frac{1}{n}\sum_{j=1}^{n-3}\left[\frac{(n-j)(n-j-1)(n-j-2)}{n(n-1)(n-2)(n-3)}\right]x_j \qquad ...5d$$

where

b_r = r-th order PWM sample estimate,

n = number of observations in the sample, and

x_j = ranked observations, with x_1 being the largest observation and x_n the smallest observation.

The first four L-moments for a sample can be computed from the first four PWMs using

$$\lambda_1 = b_0 \equiv \text{L-location (mean)} \qquad ...6a$$

$$\lambda_2 = 2b_1 - b_0 \equiv \text{L-scale} \qquad ...6b$$

$$\lambda_3 = 6b_2 - 6b_1 + b_0 \qquad ...6c$$

$$\lambda_4 = 20b_3 - 30b_2 + 12b_1 - b_0 \qquad ...6d$$

where

λ_r = r-th L-moment

Hosking (1990) defines the L-moment ratios as:

$$\tau = \frac{\lambda_2}{\lambda_1} \equiv \text{L} - \text{CV (coefficient of L - variation)} \quad \text{...7a}$$

$$\tau_3 = \frac{\lambda_3}{\lambda_2} \equiv \text{L} - \text{skewness} \quad \text{...7b}$$

$$\tau_4 = \frac{\lambda_4}{\lambda_2} \equiv \text{L} - \text{kurtosis} \quad \text{...7c}$$

Hosking (1990) shows that λ_2, τ_3 and τ_4 can be thought of as measures of a sample's scale, skewness and kurtosis respectively.

In order to select an appropriate distribution and parameter estimation procedure, tests are required to evaluate the distribution and parameter estimation method.

2.1.3.3 Goodness-of-fit tests

Probability plots are useful to reveal the character of the data set and to determine if a fitted distribution appears consistent with the data. Analytical Goodness-Of-Fit (GOF) criteria provide insights as to whether the lack of fit is due to sample variability, or whether the model and data are significantly different (Stedinger *et al.*, 1993). Generally GOF tests will identify more than one distribution which is statistically acceptable and are more valuable in identifying which distributions appear to be inconsistent with the data (Cunnane, 1989; Stedinger *et al.*, 1993).

Cunnane (1989) categorises GOF tests into tests for descriptive ability and predictive ability, both of which should complement each other. When testing for descriptive ability the best fitting distribution is sought from known distributions based on one or more of the following:

- Graphical/Visual inspection.

  Although graphical methods have traditionally been used and are a useful check of reasonable fit, there is a distinct possibility of error when choosing a distribution using an inspection of a probability plot.
- GOF tests such as Chi-squared, Kolmogorov-Smirnov, Anderson-Darling statistical tests.

  These test the null hypothesis that the sample could have been drawn from the parent population and generally have little statistical power and cannot discriminate between acceptable distributions.
- Tests based on skewness and moment-ratio diagrams.

  It is difficult to attribute the scatter of points in moment-ratio diagrams to sampling error or to genuine differences between parent populations, particularly when only short records are available.
- Numerical indices of agreement calculated from probability plots.

  These tests do not account for the greater natural sampling variation of the largest elements in a sample and usually select the 3-parameter distributions.
- Regional pooling of data, and applying the above GOF tests to the pooled data.
- Behaviour analysis by simulation study or theoretical analysis to determine if the sample could have been drawn from a candidate distribution.

Tests for predictive ability involve testing how well candidate distributions can estimate quantiles when the population distribution is not identical to that of the candidate distribution and may utilise:

- split sample tests, and/or

- tests for robustness by testing whether a distribution and method of parameter estimation are insensitive to departure from assumptions made.

One relatively recent innovation for visual interpretation of GOF is the L-moment diagram. L-moment diagrams have been used extensively in recent studies to select appropriate probability distributions (e.g. Hosking and Wallis, 1987; Vogel *et al.*, 1993a; Vogel *et al.*, 1993b). L-moment diagrams are similar to conventional product moment diagrams and compare sample estimates of τ_2, τ_3 and τ_4 with a range of different theoretical distributions. An advantage of L-moment diagrams is that a range of distributions can be plotted on the same diagram and it is thus useful for evaluating which distribution provides a satisfactory approximation to the distribution of a particular hydrological variable. Vogel and Fennessey (1993) advocate the replacement of product moment diagrams by L-moment diagrams because, unlike product moment diagrams, L-moment ratios are nearly unbiased for all underlying distributions.

The theoretical relationships between τ_3 and τ_4 for the probability distributions shown in Figure 1 are summarised by Hosking (1991a) and Stedinger *et al.* (1993). The two parameter distributions in an L-moment diagram are represented by a single point, and the 3 parameter distributions by a continuous curve.

Regional rainfall frequency estimation methods have been favoured over conventional at-site methods in recent years (Nandakumar, 1995) and are hence reviewed in the following section. Four generic approaches to frequency analysis are listed by Cunnane (1989) and Nathan and Weinmann (1991) as:

- At site analysis

  Hydrometric data at the site are used to estimate the quantiles.
- At site/regional analysis

  Quantile estimates are based on both the data of the site under consideration and the data from other sites in the region.

- Regional analysis only

  Quantiles are derived from data from other sites in the region.

- Transposition of information from other sites.

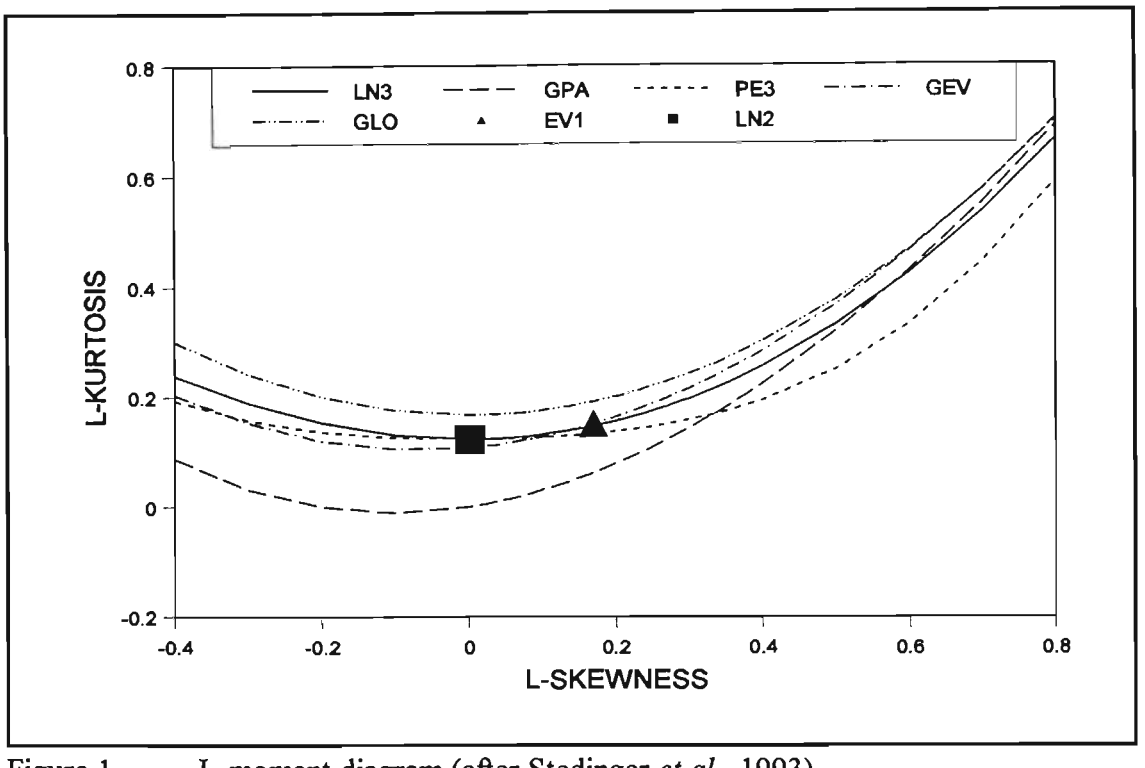


Figure 1 L-moment diagram (after Stedinger *et al.*, 1993)

2.2 JOINT AT-SITE AND REGIONAL APPROACHES

Given that the data at a site of interest will seldom be sufficient or available for frequency analysis, it is necessary to use data from similar and nearby locations (Stedinger *et al.*, 1993). This approach is known as regional frequency analysis and utilises data from several sites to estimate the frequency distribution of observed data at each site (Hosking and Wallis, 1987; Hosking and Wallis, 1997). Thus the concept of regional analysis is to supplement the time limited sampling record by the incorporation of spatial randomness using data from different sites in a region (Schaefer, 1990; Nandakumar, 1995).

Regional frequency analysis assumes that the standardised variate has the same distribution at every site in the selected region and that data from a region can thus be combined to

produce a single regional flood or rainfall frequency curve that is applicable anywhere in the region with appropriate site-specific scaling (Cunnane, 1989; Gabriele and Arnell, 1991; Hosking and Wallis, 1997). This approach can also be used to estimate events if no information exists (ungauged) at a site (Pilon and Adamowski, 1992).

2.2.1 Advantages

In nearly all practical situations a regional method will be more efficient than the application of an at-site analysis (Potter, 1987). This view is also shared by both Lettenmaier (1985; cited by Cunnane, 1989) who expressed the opinion that "regionalisation is the most viable way of improving flood quantile estimation" and by Hosking and Wallis (1997) who, after a review of recent literature, advocate the use of regional frequency analysis based on the belief that a "well conducted regional frequency analysis will yield quantile estimates accurate enough to be useful in many realistic applications". When regions are "slightly" heterogenous (i.e.$1<H<2$, as defined in Section 2.2.3.2), regional analysis yields more accurate design estimates than at-site analysis (Lettenmaier and Potter, 1985; Lettenmaier *et al.*, 1987; Hosking and Wallis, 1988). Even in heterogenous regions, regional frequency analysis may still be advantageous for estimation of extreme quantiles (Cunnane, 1989; Hosking and Wallis, 1997).

The extrapolation to return periods beyond the record length introduces much uncertainty which can be reduced by regionalisation procedures which relate the observed flood or rainfall at a particular site to a regional response (Ferrari *et al.*, 1993). Nathan and Weinmann (1991) illustrate the effect of record length on quantile estimates and show that the at-site/regional estimates are far more robust in relation to length of record than those based only on at-site data, particularly when only short record lengths are available.

The advantages of regionalisation are thus evident from previous studies. The next section briefly reviews some methods of regionalisation.

2.2.2 Methods

Frequency analyses estimate how often a specified event is likely to occur and is applicable to many environmental variables such as rainfall and runoff (Hosking and Wallis, 1997). Hence the methods described could be applied to both rainfall, flood and low flow frequency analyses (Stedinger *et al.*, 1993). General approaches to regional frequency analysis are categorised by Nathan and Weinmann (1991) as:

- station year methods,
- record extension,
- region averaging methods, and
- Bayesian methods.

Regional averaging of at-site statistics of the data is the best known alternative to the station year method (Buishand, 1991). Hosking and Wallis (1997) summarise approaches to regionalisation using regional averaging as listed in Table 5. At-site estimation, where all the parameters of the distribution are estimated from at-site estimates, is included for reference in Table 5.

The regional shape approach estimates the mean and dispersion from at-site statistics and the shape parameters are estimated from the mean of the at-site shape measure for the sites in the region. The method is intermediate between the regional shape estimation procedures and the index value procedures. Some justification for this approach is that the accuracy of the higher order moments may be better estimated using a regionalised approach. The regional shape estimation method may be preferred to the index value method if there are:

- doubts about the homogeneity of extreme rainfall events in the region,
- the main interest is in the estimation of quantiles in the extreme upper tail, or
- if the at-site records are fairly long, but the regional estimate of L-skewness is still more accurate than the at-site estimate (Hosking and Wallis, 1997).

Table 5 Estimates of distribution parameters used by different variants of regional frequency analysis (after Hosking and Wallis, 1997)

Variant	Mean	Dispersion	Shape
At-site	at-site	at-site	at-site
Regional shape estimation	at-site	at-site	regional average
Index value	at-site	regional average	regional average
Hierarchical regions	at-site	regional average for subregion	regional average for full region
Fractional membership	at-site	weighted average of regional estimates	
Region of influence	at-site	weighted average of regional estimates, for stations in a site's region of influence	
Mapping	at-site	estimated function of site characteristics	

For index-value procedures the mean is estimated from at-site estimates, while the dispersion and shape statistics are both estimated by regional averaging.

The hierarchical regional approach is an index value procedure in which relatively large regions are used to define the shape parameter. These regions are then subdivided into smaller regions over which the dispersion is assumed to be constant. A disadvantage of this method is that estimated parameters and quantiles may change abruptly between adjacent regions (Hosking and Wallis, 1997). This approach has been used, *inter alia*, by Gabriele and Arnell (1991) and Cannarozzo *et al.* (1995).

Fractional membership entails a site having fractional membership in several regions, and not only in a single region. The use of fractional membership does not allow any relaxation of the criteria for homogeneous regions, but does enable a smooth transition between regions.

Using the region of influence approach, parameters and quantiles at the site of interest are based on a regional frequency analysis in which a region is chosen to consist of sites that are expected to have a similar distribution to the site of interest. The sites are considered to be the "region of influence". Smooth transitions between regions are possible. This approach has been used, *inter alia*, by Burn (1990a), Burn (1990b) and Zrinji and Burn (1994). A disadvantage of the method is that appropriate site characteristics have to be chosen and weights have to be assigned to the characteristics (Hosking and Wallis, 1997).

Mapping involves constructing a map that can be used to estimate the parameters at a particular site and is applicable when the parameters of a regional frequency analysis vary smoothly and hence can be mapped (Hosking and Wallis, 1997). For example, Schaefer (1990) mapped the CV and skewness of a fitted GEV distribution as a function of at-site Mean Annual Precipitation (MAP). A similar approach has also been used by McKerchar and Pearson (1990) and McConachy (1995).

Hosking and Wallis (1997) recommend that the following concepts and principles should be incorporated in a regional frequency analysis:

- Frequency analysis should be robust.

  Modelling of environmental variables is extremely complex and hence exact representations of the physical processes are not feasible. Therefore the procedure should be such that even when the model's assumptions deviate from the true physical process, the quantile estimates yielded by the model would not be seriously degraded.
- Simulation should be used to assess a frequency distribution.

  Monte Carlo simulation is recommended to evaluate the properties of a frequency analysis procedure or to compare two or more procedures. Synthetic series can be generated to simulate real world data, and the adequacy of the proposed modelling procedure can be assessed for such series, since the true quantiles of the frequency distribution are known.

- Regionalisation is valuable.

  Based on the assumption that the sites form a homogeneous region, more information is available from a regional analysis than from an at-site only analysis, and hence quantile estimates are, potentially, more accurate.

- Regions need not be geographical.

  Station proximity is not necessarily an indicator of the similarity of the frequency distributions. It is proposed that groupings are formed based on variables or site characteristics which are thought to influence the frequency distribution, such as latitude, longitude, altitude or MAP.

- Frequency distributions need not be "textbook" type distributions.

  Environmental variables are generally "heavy tailed" (i.e. quantiles increase rapidly with return period) and usually have a relative short length of record. Hence it is often not possible to unequivocally identify a particular distribution. Therefore, distributions other than "standard" distributions should be considered.

- L-moments provide useful summary statistics.

  Fitting a distribution to the data involves assuming a particular distribution and estimating a finite number of parameters. Sample moment statistics such as skewness and kurtosis are often used to judge the goodness-of-fit between a sample and a postulated distribution. However, it has been shown that these statistics are algebraically bounded with bounds dependent on sample size. In addition, it has been found that the sample skewness and kurtosis, particularly in small samples, seldom approximate population statistics well. Therefore L-moments are recommended, as they are able to characterise a wider range of distributions and, when estimated from a sample, are more robust to the presence of outliers in the data. When compared to conventional moments, L-moments are less subject to bias in estimation.

A regional index value based procedure which incorporates the above guidelines has been developed and has been shown in recent studies to yield suitably robust and accurate quantile estimates (Guttman, 1993; Hosking and Wallis, 1993; Hosking and Wallis, 1997).

2.2.3 An Index Value Procedure Based on L-moments

Hosking and Wallis (1993) presented a procedure to estimate the parameters of the regional frequency distribution by combining the at-site L-moments to give regional values. Assuming the region to be homogeneous, the regional average L-moment ratios are computed from observations scaled by an index value. The regional average L-moment ratios are computed by weighting according to an individual site's record length. These regional average L-moment ratios are equated to the population L-moment ratios and used to fit the distribution. This distribution, after appropriate re-scaling by the at-site index value, is used at each site to estimate quantiles. This procedure has been termed the regional L-moment algorithm (Hosking and Wallis, 1997). The strength of regional frequency analysis using the regional L-moment algorithm is that it is useful even when not all of its assumptions are satisfied (Hosking and Wallis, 1997).

An index value approach assumes that the region is homogeneous, i.e. the frequency distributions of values of all the sites in the region are identical, apart from a site-specific scaling factor. If data are available from N sites in a region and the record length at site i is n_i, and if $Q_i(F)$ is the quantile of non-exceedance probability F at site i, then

$$Q_i(F) = \mu_i q(F), \quad i = 1,.....N \qquad ...8$$

where

μ_i = index value, and

$q(F)$ = regional quantile of non-exeedance probability F.

The index value (μ_i) may be taken as the mean of the at-site frequency distribution or any other location parameter (Hosking and Wallis, 1997). The regional quantiles ($q(F)$) define a dimensionless regional frequency distribution common to all sites, known as a *regional growth curve*, i.e. the common distribution of Q_{ij}/μ_i, where Q_{ij} is the j-th observation at site i. The mean ($\overline{Q}$) is commonly used as the index value, although other location parameters could be used.

The dimensionless values ($q_{ij} = Q_{ij} / \mu_j$, j=1,...n_i, i = 1,....N) may be rescaled to estimate $q(F)$. If the form of $q(F)$ is known, then it is necessary to estimate the p parameters, $\theta_1 ... \theta_p$.

In the regional L-moment algorithm (Hosking and Wallis, 1993; Hosking and Wallis, 1997) the p parameters are estimated separately at each site, and if the site i estimate of θ_k is denoted $\hat{\theta}_k^{(i)}$, then the at-site estimators are combined to give regional estimates as

$$\hat{\theta}_k^R = \sum_{i=1}^{N} n_i \hat{\theta}_k^{(i)} / \sum_{i=1}^{N} n_i \quad . \qquad ...9$$

This is a record length weighted average, with the estimate at site i given weight proportional to n_i. The quantile estimates at site i are then obtained by combining the estimates of μ_i and $q(F)$ as

$$\hat{Q}_i(F) = \hat{\mu}_i \hat{q}(F) \; . \qquad ...10$$

The results of statistical analyses are inherently uncertain and require an assessment of the magnitude of the uncertainty. Hosking and Wallis (1997) point out that the accuracy of the assessment is a function of the assumptions made and recommend that the method used to assess the uncertainties should be robust enough to be useful even when the assumptions are not all satisfied. For example, the region may be slightly heterogenous, the incorrect distribution may have been chosen, or statistical dependence of the data may exist. Hosking and Wallis (1997) recommend that Monte Carlo simulations be used to estimate the accuracy of the estimated quantiles.

Monte Carlo simulation techniques were used by Hosking and Wallis (1997) to investigate the performance of the regional L-moment algorithm under a wide range of conditions and concluded:

- Regionalisation is valuable.

  Regional estimation is more accurate than at-site estimation, even if the region is slightly heterogenous, or if the incorrect distribution is selected, or if inter-site dependence is evident. This is particularly so in the estimation of quantiles far into the tail of the frequency distribution.

- There is little gain in using regions containing more than 20 stations.

  This is a result of the errors in quantiles and errors in growth curves decreasing slowly as a function of the number of sites in a region.

- Regional estimates are less valuable relative to at-site estimates as record lengths increase.

  Regions should thus contain fewer sites when the at-sites record lengths are long.

- The use of 2-parameter distributions are not recommended in regional frequency analyses.

- Mis-specification of the correct frequency distribution is only important for quantiles far into the tail of the distribution (F>0.99).

- Certain robust distributions such as the Kappa and Wakeby distributions yield reasonably accurate estimates over a wide range of at-site frequency distributions.

- Heterogeneity introduces bias into estimates which are not typical of the region, and can be the major source of error in estimated quantiles and growth curves.

- Small amounts of inter-site dependence should not be a concern in regional estimation.

  Inter-site dependence has little effect on bias, but does increase the variability of estimates.

- The advantage of regional estimates over at-site estimates is greatest at extreme quantiles (F>0.999), where mis-specification of the frequency distribution is more important than heterogeneity.

In order to implement the index value procedure as outlined above, which has been termed the *Regional L-Moment Algorithm* (RLMA), Hosking and Wallis (1993; 1997) proposed the following stages in a regional frequency analysis and developed statistics, based on L-moments, that provide objective support in this process.

2.2.3.1 Screening of data

Initial screening of the data should aim at verifying that the data collected at a site are a true representation of the quantity being measured and that all the data are drawn from the same frequency distribution. Two kinds of important and plausible errors occur in environmental data:

- data values may be incorrect (incorrect recording/transcription), and/or
- circumstances under which data were collected may have changed over time (e.g. moving of measuring device).

Gross error checks for outlying values and repeated values should be performed (Hosking and Wallis, 1997). In addition, checks in levels and trends are useful and comparisons between sites should be performed to check for any irregularities. The above errors are reflected in the L-moments of the sample and the use of a convenient amalgamation of the L-moment ratios into a single measure of discordancy (D) is recommended. Hence sites whose L-moments are markedly different from those of the other sites in the data set can be identified as being discordant. The D statistic is based on the "cloud of points" when plotted in three-dimensional space (L-CV, L-skewness, L-kurtosis). A site is flagged as being discordant if it is far from the centre of the cloud containing the other points.

Assuming that a region comprises of N sites with $\mathbf{u}_i = [t^{(i)}, t_3^{(i)}, t_4^{(i)}]^T$ the vector of sample L-moments for the i-th site in the region i.e. L-CV, L-skewness and L-kurtosis respectively, which are analogous to the population τ, τ_3, and τ_4 in Equation 7, and T denotes the

transposition of a matrix. Hosking and Wallis (1997) define the discordancy index for site i as

$$D_i = \tfrac{1}{3} N(\mathbf{u}_i - \bar{\mathbf{u}})^{\mathrm{T}} \mathbf{A}^{-1}(\mathbf{u}_i - \bar{\mathbf{u}}) \qquad ...11$$

where

$$\bar{\mathbf{u}} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{u}_i \text{ , and} \qquad ...12$$

$$\mathbf{A} = \sum_{i=1}^{N} (\mathbf{u}_i - \bar{\mathbf{u}})(\mathbf{u}_i - \bar{\mathbf{u}})^{\mathrm{T}} \qquad ...13$$

The critical value of D is determined as a function of the number of sites in the region and is 3 for $N \geq 15$. It is envisaged that the D statistic could initially be used to identify gross errors within a large group of sites within a defined geographical area. When tentative homogeneous regions have been identified, the discordancy measure can then be calculated for each site in a proposed homogeneous region. The use of the discordancy measure in this study is explained in Section 5.1.

2.2.3.2 Identification of homogeneous regions

The identification of homogeneous regions is usually the most difficult of all the stages in a regional frequency analysis and requires the most subjective judgment (Hosking and Wallis, 1997). This step aims to form groups of sites that approximate the homogeneity condition, i.e. the site's frequency distributions are identical apart from a site-specific scale factor.

Data available for the formation of regions are site statistics (quantiles calculated from measurements) and site characteristics (e.g. latitude, longitude, elevation, MAP and other physical properties). Hosking and Wallis (1997) recommend that the site characteristics,

and not the site statistics, be used for regionalisation. The at-site statistics should be used for independent testing of proposed homogeneous regions. Some statistics (e.g. MAP, rainfall seasonality) which are estimated from measurements may be included in the site characteristics, provided that the statistics are not too highly correlated with the variable of interest. This approach would enable the estimation of quantiles at ungauged sites.

In a homogeneous region all sites will have the same population of L-moments. Owing to sampling variability, the sample L-moments will be different. Hence it is necessary to evaluate whether the between-site variation in sample L-moments is what the variation would be expected to be in a homogeneous region.

Hosking and Wallis (1993) developed a heterogeneity test statistic (H) which compares the between-site variability (dispersion) of L-moments with what would be expected for a homogeneous region. Dispersion is measured as the distance on a plot of L-skewness vs L-CV from a site's plotted point to the group's average point, weighted according to record length of individual sites.

Assume that a proposed region consists of N sites with the i-th site having a record length of n_i and sample L-moment ratios of $t^{(i)}$, $t_3^{(i)}$, $t_4^{(i)}$. The regional average L-CV, L-skewness and L-kurtosis, denoted by t^R, t_3^R, t_4^R respectively, are weighted proportionally to the sites n_i. For example

$$t^R = \sum_{i=1}^{N} n_i t^{(i)} \Big/ \sum_{i=1}^{N} n_i \qquad \ldots 14$$

The weighted standard deviation of the at-site sample L-CVs are calculated as

$$V = \sqrt{\sum_{i=1}^{N} (n_i - t^R)^2 / \sum_{i=1}^{N} n_i} \qquad \ldots 15$$

The 4-parameters Kappa distribution, which includes as special cases the generalised logistic, generalised extreme value and generalised Pareto distributions, is fitted to the regional average L-moment ratios ($1, t^R, t_3^R, t_4^R$) and a large number (N_{sim}, generally ≥ 500) realisations of a homogeneous region with N sites are simulated using this Kappa distribution as its frequency distribution. This approach is less restrictive than other commonly applied homogeneity tests (Hosking and Wallis, 1997). For each simulated region, V is calculated and thus the mean (μ_v) and standard deviation (σ_v) of the N_{sim} values of V may be estimated. The H test statistic is computed as

$$H = \frac{(V - \mu_v)}{\sigma_v} \qquad ...16$$

If this test statistic has a large positive value, then the hypothesis of homogeneity is not true. If $H<1$, the region is considered "acceptably homogeneous"; if $1<H<2$, the region is claimed "possibly heterogeneous" and for $H>2$ the region is "definitely heterogeneous" (Hosking and Wallis, 1997). Despite these guidelines, Hosking and Wallis (1997) recommend that the H test statistic not be used as a significance test, as the criteria are somewhat arbitrary.

Hosking and Wallis (1997) review methods of forming groups of similar sites to be used in a regional frequency analysis and categorise procedures used in previous studies as:

- geographical convenience,
- subjective partitioning,
- objective partitioning,
- cluster analysis, and
- other multivariate methods of analysis.

Hosking and Wallis (1997) regard cluster analysis as "the most practical method of forming regions from large data sets". The reciprocal of the Euclidian distance in a space of site-characteristics is used to measure similarity. The site characteristics should be re-scaled such that all the characteristics have similar variability, i.e. the ranges or standard deviations

are similar for all sites in the data set. If equal weighting for each site characteristic is not required, then subjective weighting may be introduced. As mentioned above, the use of the site characteristics in the cluster analysis enables the independent testing of clusters for homogeneity using site statistics. Subjective adjustments of the cluster analysis may reduce the heterogeneity and improve the physical coherence of regions. For a homogeneous region, simulation experiments by Hosking and Wallis (1997) indicated that little additional accuracy is gained by having more than 20 sites per cluster . The use of cluster analysis to identify homogeneous rainfall regions in South Africa, in conjunction with the H test statistic, is detailed in Section 5.2.

2.2.3.3 Choice of regional frequency distribution

After initial regionalisation has been performed, regions may still be slightly heterogeneous (i.e. $1<H<2$) and the aim when selecting a suitable distribution is not to identify the "true" distribution, but to select a distribution which provides accurate estimates of quantiles at all sites in the region and which will give accurate estimates of quantiles of the distribution from which future events will arise. It is not necessary to seek the distribution that fits the observed data best, but to select a robust distribution which fits the data adequately. Using this approach to selection of a distribution will ensure that, even if the selected distribution is not the true distribution, or if future events come from a slightly different distribution, reasonably accurate quantiles will still be estimated (Hosking and Wallis, 1997).

In regions with slight heterogeneity, even though no distribution will adequately fit the data at all sites, a single distribution may still lead to more accurate estimates of the quantiles. In such cases, robust distributions such as the Kappa and Wakeby distribution should be used (Hosking and Wallis, 1997).

The choice of distribution may be affected by the intended application and the properties of the distribution such as the upper bound, upper tail, shape, lower bound and whether zero values are handled by the distribution.

Hosking and Wallis (1997) argue against using distributions that have an upper or lower bound which may impose a physical limit or may compromise the accuracy of estimates for large return periods. When an unbounded distribution is used, it is assumed that the upper bound of the distribution cannot be estimated with sufficient accuracy and that over the range of return periods of interest an unbounded distribution would better approximate the true distribution. Hosking and Wallis (1997) recommend using a set of candidate distributions that covers a range of different tail weights, as usually insufficient data are available to estimate the shape of the tail of the distribution with any accuracy. Most probability distributions are single peaked, but where observations have qualitatively different causes, such as when the extreme events arise from different meteorological conditions, a mixture of two distributions could be used. This approach was used by Pegram and Adamson (1988) in a risk analysis of extreme storms and floods in KwaZulu-Natal, South Africa. If estimates of quantiles in the lower tail are of interest, a distribution that allows for a non-zero proportion of zero values should be considered (Hosking and Wallis, 1997).

Hosking and Wallis (1997) advocate using distributions with three or more parameters in a regional frequency analysis, as sufficient data are usually available to accurately estimate the parameters of the distribution. Two parameter distributions are not robust enough for application in regional frequency analyses and may give rise to large biases in the tails of the distribution if the selected candidate distribution is not the correct one.

Given a homogeneous region, a GOF test statistic (Z) was developed by Hosking and Wallis (1993) to test whether a region's average L-moments are consistent with those of the fitted distribution. In a homogeneous region, the scatter of the sample's L-moments represent no more than sampling variability and therefore the L-moments are well summarised by the regional average values. The GOF test statistic is derived by the difference between the L-kurtosis of the fitted distribution and observed data, scaled by the standard deviation of the L-kurtosis of the fitted distribution, which is estimated by simulation. The selection of an appropriate probability distribution for rainfall in South Africa is detailed in Section 5.5.

Assume that a proposed region consists of N sites with the i-th site having a record length of n_i and sample L-moment ratios of $t^{(i)}$, $t_3^{(i)}$, $t_4^{(i)}$. The regional average L-CV, L-skewness and L-kurtosis, denoted by t^R, t_3^R, t_4^R respectively, are weighted proportionally to the sites record length (n_i). A Kappa distribution is fitted to the regional average L-moment ratios 1, t^R, t_3^R, t_4^R and N_{sim} realisations of a region with N sites are simulated, each with this Kappa distribution as its frequency distribution. For the m-th simulated region with regional average L-skewness t_3^m and L-kurtosis t_4^m, the bias (B_4) of t_4^R is calculated as

$$B_4 = \frac{1}{N_{sim}} \sum_{m=1}^{N_{sim}} (t_4^m - t_4^R) \qquad ...17$$

and the standard deviation of t_4^R as

$$\sigma_4 = \sqrt{\frac{1}{N_{sim} - 1} \times \left[\sum_{m=1}^{N_{sim}} (t_4^m - t_4^R)^2 - N_{sim} B_4^2 \right]} . \qquad ...18$$

For each candidate distribution, the goodness-of-fit measure is calculated as

$$Z^{DIST} = \frac{(\tau_4^{DIST} - t_4^R + B_4)}{\sigma_4} \qquad ...19$$

where

τ_4^{DIST} = L-kurtosis of a candidate 3-parameters distribution (*DIST*) fitted to the regional average L-moments 1, t^R, t_3^R.

The fit of a candidate distribution is deemed to be adequate if $|Z| \leq 1.64$.

2.2.3.4 Estimation of regional frequency distribution

Assuming that N sites form a homogeneous cluster, with site i having a record length n_i, sample mean $l_1^{(i)}$ (analogous to the population λ_1 in Equation 6), and sample L-moment ratios $t^{(i)}$, $t_3^{(i)}$, $t_4^{(i)}$, ..., analogous to the population τ, τ_3 and τ_4 in Equation 7, then the regional average L-moment ratios t^R, t_3^R, t_4^R, ..., which are weighted proportionally to the sites' record length, are computed as:

$$t^R = \sum_{i=1}^{N} n_i t^{(i)} / \sum_{i=1}^{N} n_i \qquad ...20$$

$$t_r^R = \sum_{i=1}^{N} n_i t_r^{(i)} / \sum_{i=1}^{N} n_i \;,\; r = 3,4,.... \qquad ...21$$

The regional average mean is set to 1 ($l_1^{(R)} = 1$) and the selected distribution is fitted by equating the theoretical L-moment ratios to $l_1^{(R)}$, t^R, t_3^R, t_4^R calculated in Equations 20 and 21. As shown in Equation 22, the quantile, with non-exceedance probability F, may be estimated by combining the quantile function of the fitted distribution ($\hat{q}$) with the at-site mean.

$$\hat{Q}_i(F) = l_1^{(i)} \hat{q}(F) \qquad ...22$$

Slightly more accurate quantile estimates are obtained in most cases if, as above, L-moment ratios and not L-moments are averaged (Hosking and Wallis, 1997).

This index value based region frequency analysis approach using L-moments has been termed the Regional L-Moment Algorithm (RLMA) by Hoksing and Wallis (1997). As discussed above, the RLMA has many reported advantages, including robustness, and is relatively simple to apply. Routines obtained from Hosking (1996) were utilised for the

calculation of the D and H test statistics and for the implementation of the RLMA in South Africa, as described in Chapter 5. A procedure for the assessment of the accuracy of the quantiles estimated using the RLMA is described in the following section.

2.2.3.5 Assessment of accuracy of estimated quantiles

The inherent uncertainty in statistical analysis requires that an assessment of the uncertainty should be made. Traditionally, this has be done by constructing confidence intervals for estimated parameters and quantiles, assuming that the statistical models assumptions are satisfied. Such confidence intervals are of limited use as rarely are all the assumptions regarding the data valid and uncertainty concerning the "correct" model selection is generally present (Hosking and Wallis, 1997). In particular for the RLMA, the possible of heterogeneity in the region, mis-specification of the frequency distribution and statistical dependence between the data should all be taken in account, in a way consistent with the data, in order to obtain realistic assessments of the accuracy of the quantiles.

Hosking and Wallis (1997) propose that Monte Carlo simulation is a reasonable approach to estimate the accuracy of the quantiles. The simulated regions should have the same number of sites, record lengths at each site and regional average L-moments as the actual data, and should include appropriate combinations and levels of heterogeneity, inter-site dependence and mis-specification of model. Inter-site dependence is accounted for by assuming that if each site's frequency distribution were transformed into the Normal distribution, then the joint distribution of all N site would be multivariate Normal. The algorithm for the proposed Monte Carlo simulation procedure is:

(i) For each of the specified N sites, with individual record lengths n_i, calculate the at-site L-moments from the observed data.

(ii) Estimate the parameters of the at-site frequency distribution given the at-site L-moment ratios. The at-site frequency distribution should be chosen using

goodness-of-fit measures or if several or no distributions are suitable, then the flexible Wakeby or Kappa distributions may be used.

(iii) Generate the matrix **R** of inter-site correlations.

(iv) For M repetitions of the simulation procedure a random sample of length n_i is generated from the selected frequency distribution for each site in the region. For sites that have inter-site dependence:

- Generate a realisation of a random vector y_k, for each time point k=1,.... max(n_i), with elements $y_{i,k}$, i=1,... N, that have a multivariate Normal distribution with mean vector zero and covariance matrix **R**.
- Calculate data values $Q_{ik} = Q_i(\Phi(y_{i,k}))$, where Q_i is the quantile function for site i and Φ is the cumulative distribution function of the standard Normal distribution i.e. each $y_{i,k}$ is transformed to the required marginal distribution.

(v) Apply the RLMA to the sample of regional data.

- Calculate the at-site and regional average L-moment ratios.
- Fit the chosen distribution.
- Calculate estimates of the regional growth curve and at-site quantiles.

(vi) Calculate the measures of accuracy for example as:

$$R_i(F) = \sqrt{\frac{1}{M}\sum_{m=1}^{M}\left(\frac{\hat{Q}_i^{\,m}(F) - Q_i(F)}{Q_i(F)}\right)^2} \qquad \ldots 23$$

where

$R_i(F)$ = RMSE,

$\hat{Q}_i^{\,m}(F)$ = quantile estimate at i-th site of m-th repetition for non-exceedance proabability F,

$Q_i(F)$ = quantile at i-th site for non-exceedance proabability F estimated using regional growth curve, and

M = number of repetitions of simulation procedure.

An estimate of the accuracy of the quantiles over all the sites in the region may be defined as the regional average relative RMSE, $R^R(F)$:

$$R^R(F) = \frac{1}{N}\sum_{i=1}^{N} R_i(F) \qquad ...24$$

In the following section a review of DDF studies in South Africa is presented. None of the studies reviewed has adopted a regional approach to design storm estimation in South Africa.

2.3 REVIEW OF DESIGN STORM ESTIMATION STUDIES IN SOUTH AFRICA

Vorster (1945) applied regionalised relationships adopted from the USA and identified six rainfall regions in South Africa which were similar to the regions which had been identified in the USA. The relationships were modified to fit local conditions based on 24 h rainfall totals and similarities in vegetation cover. Owing to a paucity of recording raingauges at the time of the study, he combined data from different sites within a region to produce 5, 10, 30, 60, 120, 240, 480, 960 and 1440 min rainfall intensity maps in SA for return periods of 5, 10, 20, 40 and 80 years. Weddepohl (1988) points out that the regions in SA and USA displayed dissimilarities and the practice of combining records into a single record (station year approach) is now considered a poor procedure. Woolley (1947) stated that Vorster's regions were too broad and investigated the use of MAP as a predictor variable for design storms. Bergman and Smith (1973) found that Vorster's (1945) work generally overestimated the magnitude of extreme events.

The SAWB (1956) used the EV1 distribution to produce 1 day design rainfalls for return periods of 5, 10, 15, 20, 30, 40, 60, 80 and 100 years for 253 stations in South Africa. Maps of 1 day : MAP ratios for 5, 10, 20, 30, 60 and 100 year return periods were also

presented. Weddepohl (1988) refers to possible errors in the data and the short record used in this study.

Reich (1961) used autographic data from 12 stations in South Africa and the EV1 distribution to estimate the 2, 5, 10, 25, 50 and 100 year return period rainfall intensities for durations of 30 min, 1 and 24 h. Reich (1963) determined and mapped the 2 year return period, 1 h design storm ($P_{2,1}$) using data from 12 autographic and 210 daily raingauges in South Africa and modified USA depth-frequency relationships, after showing that the USA relationships underestimated intermediate frequencies. Hershfield's (1962) relationships were modified to enable the T year return period, D h design storm ($P_{T,D}$) to be predicted from T, D, average number of days per year on which thunder was heard and average 24 h annual maximum precipitation.

The depth-duration relationships from the USA were extended by Reich (1963) to include the estimation for 15 min intervals in South Africa. Maps of the ratio $P_{100,24} / P_{2,24}$ were derived in order to predict the 100 year return period event. Thus from $P_{2,1}$ and $P_{2,24}$, and the depth-duration relationship, the 2 year return period design storm for any intermediate duration can be derived. Then using the depth-frequency relationship, $P_{100,D}$ is obtained from the $P_{100,D} / P_{2,D}$ ratio. The 2 and 100 year return period intensities are then used with the depth-frequency relationship to obtain the $P_{T,D}$ value.

The Californian plotting position (i.e. $T=N/m$) has been used to compute the probabilities of extreme rainfalls, as used, for example, by Vorster (1945). Bergman and Smith (1973) recognised the limitations of using this approach as the relative frequencies were based on short record lengths and cannot be extrapolated. Based on a review of previous work, Bergman and Smith (1973) adopted the EV1 distribution for use in the Western Cape. Data from 14 autographic stations in the Western Cape were used with record lengths ranging from 6 to 30 years. With outliers excluded, the extreme magnitudes obtained were approximately half of the values estimated by Reich (1963). When the outliers were included, the design rainfalls were similar to, but generally less than those obtained by Vorster (1945). Bergman (1974) generalised the design rainfall values for the winter rainfall

region and introduced a "K-factor", related to MAP and number of raindays, which is used to estimate $P_{10,1}$.

The SAWB (1974) published data from 64 autographic raingauges and used the EV1 distribution to estimate the 15, 30, 45 and 1440 min duration events for return periods of 25, 50 and 100 years. Sinske (1982) points out the difficulty of transferring these data to a desired location and of interpolating between durations and return periods, but recognises the pioneering work done. Adamson(1978) used the database from the SAWB and the EV1 distribution to estimate design storm depths for return periods ranging from 5 to 500 years and considered durations of 15, 30, 45 and 60 min as well as 1 day rainfalls.

Alexander (1978) presented Reich's (1963) graphical relationships in equation form as:

$$P_{T,D} = (0.35 \times \ln(T) + 0.76) \times (0.54D^{0.25} - 0.50) \times (1.83M^{0.67}R^{0.33}) \qquad ...25$$

where

$P_{T,D}$ = T year return period, D hour design storm (mm),
D = duration (min), with maximum allowable value = 120 min,
M = mean of the 24 h annual maximum daily rainfall in the range 50-115 mm, and
R = average number of days per year on which thunder is heard.

Alexander (1978) used the $P_{5,1}$ value as an predictor variable and developed the following relationship:

$$P_{5,1} = 1.55M^{0.63}R^{0.20} \qquad ...26$$

Equation 26 is very similar to the equation proposed by Hershfield (1962) and in the light of Reich's work, Alexander (1978) proposed the following equation:

$$P_{T,D} = 1.13 \times (0.27\ln(T) + 0.56) \times (0.54D^{0.25} - 0.50) \times (1.55M^{0.63}R^{0.20}) \qquad ...27$$

Midgley and Pitman (1978) derived a generalised Depth-Duration-Frequency (DDF) relationships using MAP and locality (i.e. inland vs coastal) as input variables. Adamson (1981) postulates that storms of less than 2 h duration are likely to be independent of MAP. The co-axial diagram of Midgley and Pitman (1978), which uses MAP as a predictor for durations of 15, 30, 45, 60 and 1440 min, accounts to some extent for this by introducing a locality factor which demarcates rainfall regimes. Sinske (1982) refers to the practical difficulties of reading off the diagram and on deciding whether an inland or coastal estimate is applicable to the site of interest. Schulze (1984) highlights some anomalies in the database used by Midgley and Pitman (1978), is critical of the use of LEV1 distribution and points out the physically impossible rainfall values that are estimated by the distribution and which are contained in the report by Midgley and Pitman (1978).

Op Ten Noort (1983) re-analysed the data used by Midgley and Pitman (1978) and by a least squares regression analysis derived the following two equations.

$$\text{Inland region}: I = \frac{(7.5 + 0.034\,MAP)T^{0.3}}{(0.24 + D)^{0.89}} \qquad ...28$$

$$\text{Coastal region}: I = \frac{(3.4 + 0.023\,MAP)T^{0.3}}{(0.20 + D)^{0.75}} \qquad ...29$$

where

I = point rainfall intensity ($mm.h^{-1}$),

MAP = mean annual precipitation (mm),

D = storm duration (h), and

T = recurrence interval (years).

Van Heerden (1978) produced standard intensity curves for eight intensity classes for durations up to 2 h and return periods up to 100 years. The classes were based on the 60 min intensity values and hence do not form geographic regions. Hence Sinske (1982) points out the practical difficulty of knowing which of the eight classes are applicable to the

site of interest. Adamson (1981) is critical of the subjective nature of the grouping scheme and the lack of any meaningful reference to meteorological or physical parameters.

Henderson-Sellers (1980) used the data from Midgely and Pitman (1978) to compute the parameters in Equation 30.

$$I = \frac{a}{(D+b)^n} \qquad ...30$$

where D is the duration (h) and I is the intensity ($mm.h^{-1}$). The optimum solution was found by holding the value of b=1/3. Values of a varied widely and n was found to have distinct regional differences. The four regions subsequently delineated were found to coincide closely with previous climatological classifications of precipitation regimes. Henderson-Sellers (1980) concluded that the value of n could be assumed to be constant within regions and not to vary with return period. Thus the T year return period rainfall for a duration D ($P_{T,D}$) can be derived as function of daily rainfall $P_{T,1d}$.

$$P_{T,D} = \frac{D}{24}\left[\frac{24+b}{D+b}\right]^n P_{T,1d} \qquad ...31$$

Henderson-Sellers (1980) only considered return periods of 2, 5 and 10 years in the derivation of regional values of n in Equation 31, and hence Equation 31 should not be considered for return periods > 10 years. Although Henderson-Sellers (1980) considers that the use of Equation 31 would extend the hydrological database by the use of $P_{T,1d}$ values, no adjustment was made to reflect the difference between $P_{T,1d}$ and $P_{T,24}$.

Schulze (1980) used the EV1 distribution to estimate the 1, 2 and 7 day duration rainfalls for the 2, 10, 25 and 50 year return periods. Data from 396 raingauges were used in the analysis and record lengths ranged from 30 to100 years.

Adamson (1981) estimated the 1, 2, 3 and 7 day extreme rainfalls for return periods of 2, 5, 10, 20, 50, 100 and 200 years and used approximately 8000 stations in his analysis. A censored log-N model of PDS was used in the analysis. Adamson (1981) expressed doubts as to the availability and accuracy of estimating both M and R in Equation 27 and hence replaced these values with the mean annual value of lightning flash density (L, in flashes.km^{-1}.annum^{-1}) and $P_{2,1d}$ respectively, as shown in Equation 32.

$$P_{T,D} = 1.13 \times (0.27 \ln(T) + 0.56) \times (0.54 D^{0.25} - 0.50) \times (4.53 + 0.55 P_{2,1d} + 1.893 L) \quad ...32$$

Schulze (1984) lists the most widely (as of 1983) used direct methods of estimating short duration DDF relationships in SA as:

- the Midgley and Pitman (1978) co-axial diagram,
- the modification of Reich's (1961) equations by both Alexander (1978) and Adamson (1981),
- the tabulated design values by the SAWB (1974), and
- the generalised ratios of short duration to 24 h rainfall for summer and winter rainfall/coastal regions as published by Adamson (1981).

Schulze (1984) used a digitised rainfall database to calculate D:24 h ratios and showed marked divergence between these ratios and values computed from Midgley and Pitman (1978), Adamson (1981) and from the SCS type I and II distributions. Schulze's (1984) study also showed that intensities calculated from the digitised database are generally higher than when the intensities were manually extracted from autographic rainfall charts.

Weddepohl *et al.* (1987) and Weddepohl (1988) expanded on concepts used previously by Schulze (1984) and developed four synthetic extreme storm temporal distributions from design relationships in South Africa. Hence daily design rainfall values can be disaggregated to obtain the temporal distribution of the design storms for four different regions in South Africa.

More recently, Smithers (1996) used L-moments to fit various distributions to data from 38 sites in South Africa, each of which had more than 30 years of record. Using both parametric and non-parametric GOF tests, Smithers (1996) recommended that the GEV distribution is the most appropriate distribution to use in South Africa for 24 h duration events, but concedes that this recommendation may change at a local scale.

Sinske (1982) illustrates the discrepancies between the different methods, and highlights the lack of methods to estimate design rainfall beyond the 100 year return periods. Adamson (1981) concludes from a review of previous short duration rainfall studies that regionalisation has met with little success in South Africa.

The search for generalised DDF relationships in South Africa has concentrated on linear associations between selected recurrence interval, short duration rainfall depth and other readily available predictor values (Adamson, 1981). Selected studies, both in South Africa and internationally, which have used this approach are reviewed in the next section. In addition, summaries of depth-duration and depth-frequency ratios, which are extracted directly or derived from the literature reviewed, are presented.

2.4 SCALING OF FREQUENCY RELATIONSHIPS

A number of studies have mapped predictor values such as design storms for a particular duration or return period and used regionalised ratios to estimate design storms for other durations or return periods. Some studies have assumed that these ratios are independent of return period and others have assumed that the ratios are independent of duration.

2.4.1 Depth-Duration Relationships

Many studies, both in South Africa (e.g. Bergman, 1974; Alexander, 1978) and internationally (e.g. Chen, 1983; Ferreri and Ferro, 1990; Blodgett and Nasseri, 1995), have

investigated the estimation of design storms for a required duration from an index storm. A ratio, commonly termed a depth-duration ratio, is used to convert the index storm to the design storm for the required duration. The advantage of developing D:24 h ratios and thus utilising the relatively large daily rainfall database in order to estimate shorter duration events at sites where no short duration data are available, is expanded on by Schulze (1984).

Bergman (1974) computed depth-duration ratios for durations of 15, 30, 120 and 1440 min in relation to the 60 min duration and for return periods of 5, 10, 20, 40, 50 and 100 years for the Winter Rainfall Region (WRR) in South Africa. No differences in the ratios were noted for given durations or different return periods and hence average ratios, which are independent of return period, were computed. Bergman (1974) presented a comparison of $P_{T,D} : P_{T,1}$ ratios (Table 6) with the results published by Bell (1969). Included in Table 6 are results derived from Henderson-Sellers (1980), using Equation 31 for inland (n=0.92) and the WRR (n=0.86) in South Africa, as well as results derived from Adamson (1981) for the WRR and inland regions in South Africa. Some similarities are evident for different regions in Table 6, particularly for shorter durations. However, differences in the depth-duration relationships are noted within SA for longer durations.

Froehlich (1995) and Froehlich and Tufail (1995) report on four general forms of intensity-duration relationships, listed in Table 7, which have been used in the USA. Chen (1983) derived a generalised rainfall intensity-duration-frequency relationship for use in the USA and utilised the $P_{10,1}$, $P_{10,24}$, $P_{100,1}$ and $P_{100,24}$ as index values. The depth-duration ratio ($P_{T,1} / P_{T,24}$) was assumed to be independent of return period and varied spatially in the USA with the values varying from 0.1 - 0.6. From the literature reviewed by Hargreaves (1988), there is considerable agreement that depth-duration rainfall amounts vary with a 1/4 power function of duration ($D^{0.25}$).

Table 6 Examples of $P_{T,D} / P_{T,1}$ ratios

Duration (h)	Bergman (1974)	Bell (1969)			Derived from Henderson-Sellers (1980)		Derived from Adamson (1981)		Derived from Midgley and Pitman (1978)		
	Winter Rainfall Region, SA	USA	Australia	USSR	WRR, SA	Inland, SA	WRR, SA	Inland, SA	Cape Town, SA	Durban, SA	Johannesburg, SA
0.083		0.29	0.30	0.32	0.23	0.24					
0.250	0.67	0.57	0.57	0.55	0.51	0.53	0.56	0.53	0.53	0.45	0.49
0.500	0.82	0.79	0.78	0.79	0.75	0.77	0.78	0.77	0.77	0.72	0.76
2.000	1.26	1.25	1.24	1.30	1.24	1.20	1.29	1.20			
24.000	3.60				1.97	1.66	2.44	1.67	1.97	1.97	1.71

Table 7 Generalised forms of rainfall intensity equations (after Froehlich, 1995)

Equation Type	Equation Form	Equation Parameters
I	$I=a_1/(D+b_1)$	a_1, b_1
II	$I=a_2/(D^{c2})$	a_2, c_2
III	$I=a_3/(D+b_3)^{c3}$	a_3, b_3, c_3
IV	$I=a_4/(D^{c4}+b_4)$	a_4, b_4, c_4

In order to estimate design storms for durations and return periods other than those available from isopluvial maps published for regions in the USA, Froehlich (1995) and Froehlich and Tufail (1995) used Equation 34 to express

$$P_{T,D} = P_{T,1} + f_D(P_{T,24} - P_{T,1}) \qquad ...34$$

where

f_D = D h rainfall duration factor that applies to all return periods.

Equation 34, which does not assume that the depth-duration ratio is constant for different return periods, may be expressed as a ratio of $P_{T,1}$, as shown in Equation 35 such that

$$\frac{P_{T,D}}{P_{T,1}} = 1 + f_D\left(\frac{P_{T,24}}{P_{T,1}} - 1\right). \qquad ...35$$

Ferreri and Ferro (1990) computed depth-duration ratios for data from Sicily and Sardinia and compared the ratios to those computed from Bell's (1969) depth-duration equation. The ratios were very similar for durations from 30 - 55 min, but Bells ratios were slightly smaller for durations less than 30 min. Ferreri and Ferro (1990) conclude that the small differences in the ratios confirms the independence of short duration depth-duration ratios

from geographic factors and confirms the applicability of Bell's relationship for these durations.

These findings by Ferreri and Ferro (1990) contradict those of Canterford *et al.* (1987b) who found in Australia that the use of constant ratios to interpolate to durations of less than 1 h from the1 h intensity varied significantly and could be explained on a geographical, meteorological and return period basis.

The depth-duration ratio has also been assumed to be independent of return period in some studies (e.g. Adamson, 1981; Chen, 1983). However, as shown in Table 8 using data from Midgley and Pitman (1978) and illustrated for stations in KwaZulu-Natal by Schulze (1984), the depth-duration ratios do appear to be dependent on return period. In the example shown in Table 8 for Johannesburg there are distinct trends of the $P_{T,D} / P_{T,1}$ ratio varying as a function of return period for all durations shown.

Table 8 $P_{T,D} / P_{T,1}$ ratios for Johannesburg (derived from Midgley and Pitman, 1978)

Duration (min)	Return Period (years)						Mean
	2	5	10	20	50	100	
15	0.55	0.52	0.50	0.48	0.45	0.43	0.49
30	0.80	0.78	0.77	0.76	0.74	0.73	0.76
60	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1440	1.84	1.77	1.72	1.68	1.63	1.59	1.71

2.4.2 Depth-Frequency Relationships

Bergman (1974) used the 10 year return period value as the denominator in the computation of depth-frequency ratios for the WRR in South Africa. No significant differences in the ratios were found for different durations and averaged values were compared to those presented by Bell (1969), as listed in Table 9. Also included in Table 9 are depth-duration ratios derived from results published by Midgley and Pitman (1978). Again the depth-frequency ratios appear to vary regionally in SA, particularly for longer durations.

Table 9 Comparison of $P_{T,D} / P_{10,D}$ ratios

Return Period (years)	Bergman (1974)	Bell (1969)		Derived from Midgley and Pitman (1978)		
	WRR	USA	Australia	Cape Town	Johannesburg	Durban
2	0.66	0.63	0.65	0.63	0.57	0.54
5	0.86	0.85	0.85	0.83	0.80	0.78
20	1.13			1.20	1.24	1.27
25		1.17	1.18			
50	1.30	1.31	1.33	1.51	1.64	1.73
100	1.44	1.46	1.50	1.80	2.01	2.18

Hargreaves (1988) concurs with Bell (1969) that depth-duration and depth-frequency ratios are approximately constant for diverse countries and regions. However, as shown in Table 9 and illustrated by Schulze (1984) using digitised data from 9 stations in KwaZulu-Natal, the depth-frequency ratios do appear to vary considerably from location to location.

2.4.3 Depth-Duration-Frequency Relationships

"Strict sense simple scaling" describes the assumption that storm rainfall is characterised by the property of scale invariance (Gupta and Waymire, 1990). This implies that the probability distributions of rainfall depth is the same at different time scales. According to Burlando and Rosso (1996) this can be written as

$$Z_{\lambda T}(t) \overset{d}{=} \lambda^{n} Z_{T}(t) \qquad \ldots 36$$

where $\overset{d}{=}$ denotes equality in the probability distribution and

$Z_T(t)$ = measured rainfall depth in time span of length T,
λ = scale factor and
n = scaling exponent.

If the assumption that the equality of distributions of maxima for a certain period (e.g. annual), observed at different time scales, also holds true, then both the quantiles and raw moments of any order are also scale invariant as shown in Equations 37 and 38 (Burlando and Rosso, 1996).

$$\xi_{q}(\lambda T) \overset{d}{=} \lambda^{n} \xi_{q}(T) \qquad \ldots 37$$

where

$\xi_q(T)$ = q-th quantile of H_T,
H_T = max $[Z_T(t_0), Z_T(t_0+\tau)]$,
t_0 = point on time axis (e.g. beginning of rainy season), and
τ = length of period (e.g. 1 year for AMS).

$$E[H_{\lambda T}^{l}] \overset{d}{=} \lambda^{n} E[H_{\lambda T}^{l}] \qquad \ldots 38$$

where

l = order of the moment, and

n = scaling exponent of mean.

The assumptions of scale invariance are based on trends noted in observed data. For example, as shown in Figure 2, data from raingauge CP6 at Cathedral Peak in KwaZulu-Natal, South Africa, are used to illustrate the scaling concepts.

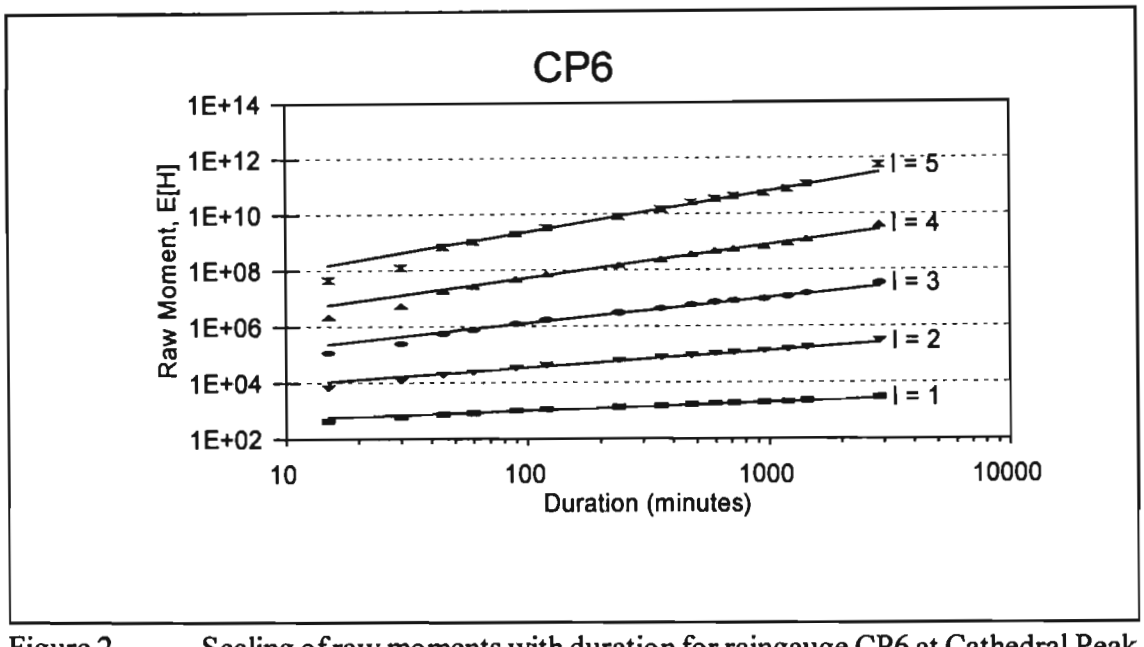


Figure 2 Scaling of raw moments with duration for raingauge CP6 at Cathedral Peak, KwaZulu-Natal, South Africa

The slope of the straight line fitted for duration ≥ 1 h and ≤24 h to the double logarithmic plot of raw moments against duration, as shown in Figure 2, is the scaling exponent α_l, for each l-th order moment. Simple scaling is said to hold true if $\alpha_l = n.l$, where n is the scaling exponent of the mean. Multiple scaling is defined as $\alpha_l \neq n.l$ (Burlando and Rosso, 1996). Simple scaling is illustrated in Figure 3 using data from raingauge CP6 at Cathedral Peak, KwaZulu-Natal, South Africa.

Burlando and Rosso (1996) explored the scaling properties of the rainfall depth-duration-frequency relationship in order to interpolate design storms for durations other than those commonly published. Menabde *et al.* (1998) tested the scaling concepts on rainfall data from two stations, one in New Zealand and the other in South Africa, and concluded that simple scaling was applicable at both sites and postulated that the scaling exponent was related to local climate. Burlando and Rosso (1996) investigated the scaling of rainfall depth while Menabde *et al.* (1998) used rainfall intensity in their investigations. Menabde *et al.* (1998) found that the extreme rainfall intensity relationships scaled for durations ranging from 0.5 - 48 h, while Burlando and Rosso (1996) showed that the range could be from as little as 2 min and up to 48 h or longer.

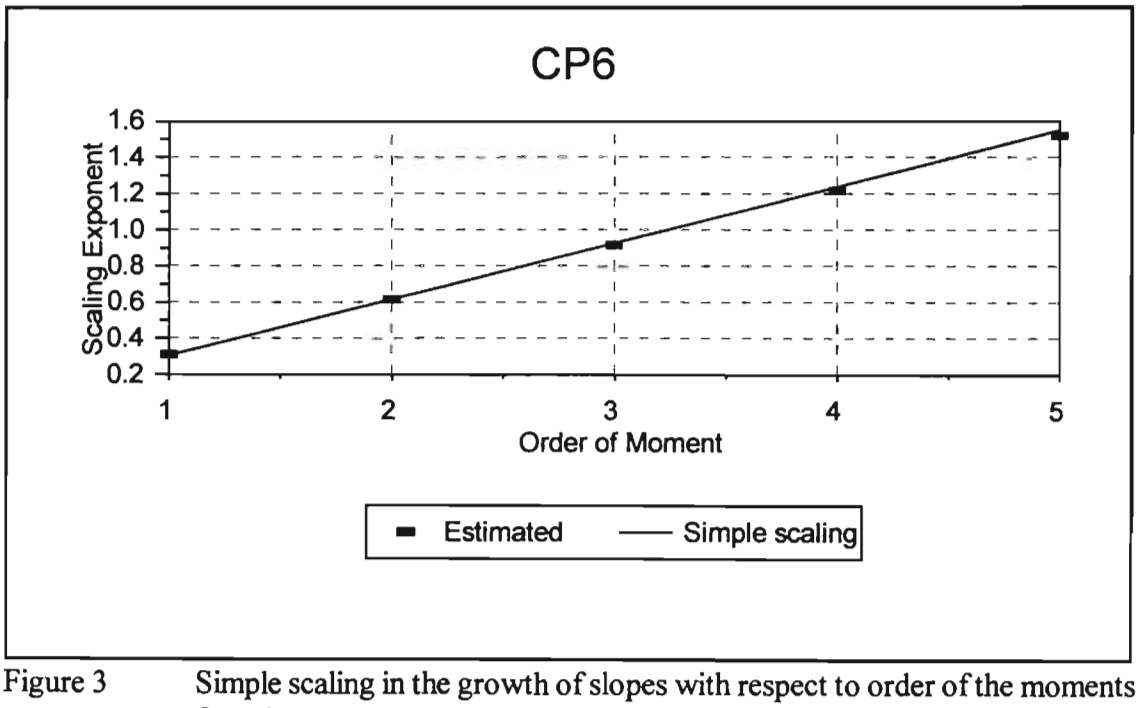


Figure 3 Simple scaling in the growth of slopes with respect to order of the moments for raingauge CP6 at Cathedral Peak, KwaZulu-Natal, South Africa

2.5 CHAPTER CONCLUSIONS

In this chapter both at-site and regional techniques of design rainfall estimation have been reviewed. Substantial benefits of using a regional approach have been reported in the literature, assuming that relatively homogeneous regions can be identified. In particular, the relatively recently developed RLMA appears to be a robust procedure and has been applied successfully in a number of studies. These techniques have been applied to short duration rainfall data from South African and the results are presented in Chapter 5.

The limited number and relatively short record lengths of reliable, observed short duration rainfall data available in South Africa are highlighted in Chapter 4. A much denser network of standard daily raingauges, which are manually recorded at 24 h intervals ending at 08:00 every day, and which have relatively longer record lengths than the recording raingauges, are available in South Africa. A number of studies reported in the literature have demonstrated the successful use of stochastic rainfall models to estimate design rainfall values. Hence the literature on modelling rainfall using stochastic Bartlett-Lewis type models are reviewed next in Chapter 3. The potential thus exists to use the stochastic rainfall models, with parameters determined from daily rainfall data, to estimate short duration design rainfall values and thus increase the spatial density at which short duration design rainfall estimates can be made in South Africa.

CHAPTER 3

MODELLING POINT RAINFALL AS A CLUSTER PROCESS

In the light of the relatively few recording rainfall stations in South Africa which have reliable short duration rainfall data, as illustrated in Chapter 4, three approaches for estimating design rainfall values have been explored. The first is a regional approach, with techniques discussed in Chapter 2 and results presented in Chapter 5, where the information at sites not having reliable data is supplemented or replaced by information from the region. In order to estimate short duration design storms at locations which do not have reliable short duration rainfall data, the second approach, with results presented in Chapter 6, attempts to utilise the scaling properties of the moments of the extreme digitised rainfall events as described in Chapter 2, in conjunction with moments derived from the daily rainfall data. The third approach, which is discussed in this chapter with results presented in Chapter 7, investigates stochastic, cluster-based rainfall models for use in the estimation of design rainfall values.

The use of stochastic processes, which consist of point events occurring in time and which have characteristics derived from sampling probability density functions, is increasing in hydrology (Entekhabi *et al.*, 1989). The modelling of rainfall using stochastic techniques has a wide range of potential hydrological applications ranging from hydrological design to the disaggregation of large time interval data into shorter durations (Onof and Wheater, 1993; Onof and Wheater, 1994a). One such application could be the disaggregation of daily rainfall into shorter time intervals for use in time dependent infiltration modelling (Bo *et al.*, 1994). Another potential application could be in flood frequency analysis where the use of a long synthetic rainfall series, generated using appropriate mathematical techniques, can provide insight and further aid in the extrapolation of the data when estimating design storms from a limited time series of historical observations (Cowpertwait *et al.*, 1996b; Verhoest *et al.*, 1997).

Rainfall models range from complex dynamic meteorological models to empirical statistical models with stochastic models, which have a modest number of parameters, representing intermediate complexity (Chandler *et al.*, 1995). While Foufoula-Georgiou and Krajewski (1995) report on a recent shift from stochastic point process models to models based the concepts of scale invariance, the use of point process models and, in particular, the use continuous time cluster based point process models are widely reported in the recent literature (e.g.Onof and Wheater, 1993; Bo *et al.*, 1994; Velghe *et al.*, 1994; Cowpertwait *et al.*, 1996; Khaliq and Cunnane, 1996; Verhoest *et al.*, 1997).

In cluster-based rainfall models, events are modelled as clusters of rain cells and each cell is a pulse with a random duration and random intensity, which is constant for the duration of the pulse. Poisson processes are used to model the distribution in time of both the storm origins and the clusters of cells. Cluster-based models thus combine the rainfall occurrence, or frequency, and depth process (Khaliq and Cunnane, 1996). One of the main advantages of rectangular pulse, cluster-based rainfall models is that the parameters are independent of the time scale used (Verhoest *et al.*, 1997).

It has been shown in the recent literature that cluster models have built into their structure the capability of representing rain cells and preserving the rainfall statistics over a range of the time scales (Rodriguez-Iturbe *et al.*, 1987a; Rodriguez-Iturbe *et al.*, 1987b; Cowpertwait, 1991). Rodriguez-Iturbe *et al.*(1987b) postulated that the range of temporal scales over which cluster based rainfall models could achieve aggregation and disaggregation was likely to be of the order of 1 to 48 h. Bo *et al.* (1994) showed that cluster based models are capable of preserving hourly statistics when only 24 and 48 h moments, computed from historical data, are used in parameter determination. The potential of using cluster-based models in the estimation of design rainfall events has been demonstrated *inter alia* by Onof and Wheater (1993; 1994b), Khaliq and Cunnane (1996), Cowpertwait *et al.*(1996a) and Verhoest *et al.* (1997).

The Bartlett-Lewis Rectangular Pulse Model (BLRPM) and the Neyman-Scott Rectangular Pulse Model (NSRPM) are examples of cluster-based models which have been shown to be

able to model rainfall characteristics over a range of time scales ranging from 1 to 24 h (Rodriguez-Iturbe *et al.*, 1987a; Rodriguez-Iturbe *et al.*, 1987b; Entekhabi *et al.*, 1989; Onof and Wheater, 1993; Bo *et al.*, 1994; Velghe *et al.*, 1994; Cowpertwait *et al.*, 1996a; Khaliq and Cunnane, 1996; Verhoest *et al.*, 1997).

3.1 BARTLETT-LEWIS AND NEYMAN-SCOTT RECTANGULAR PULSE MODELS

In cluster-based models events are represented as clusters of rain cells, with each cell a pulse of random duration and intensity which is constant throughout the duration. The Poisson distribution, which has a random number of cells or cluster size, is used to model the storm origins. A cell arrival distribution is assigned to each storm. The Bartlett-Lewis model assumes that the number of cells are geometrically distributed, whereas the Neyman-Scott model allows any convenient form of distribution to be assumed, in addition to the geometric distribution. The depth and duration of each cell are modelled by an exponential distributions (Onof and Wheater, 1993; Khaliq and Cunnane, 1996). Thus the rainfall occurrence process and rainfall depth are described independently and are then superimposed to form the rainfall model, as shown schematically in Figure 4.

In the NSRPM the cell arrival times are independent, identically distributed exponential random variables which are measured from the storm origin and have no cell at the storm origin. The BLRPM has a cell located at the storm origin with the interval between successive cells independent and exponentially identically distributed. Overlap between and within storms can occur (Entekhabi *et al.*, 1989; Khaliq and Cunnane, 1996).

Using the NSRPM and BLRPM as described above, Rodriguez-Iturbe *et al.* (1987b) found that the models were able to preserve the rainfall depth statistics and extreme values of rainfall at Denver, USA, but did not reproduce the proportion of periods with no rainfall (dry level states) satisfactorily. The BLRPM was modified by Rodriguez-Iturbe *et al.* (1988) to allow random variation from storm to storm of the exponential parameter of the

distribution of cell duration. This Modified version of the original BLRPM, or MBLRPM, enabled the model to reproduce the proportion of dry states for various time intervals. The NSRPM was similarly modified by Entekhabi *et al.* (1989) to created the Modified NSRPM (MNSRPM).

Entekhabi *et al.* (1989) expressed the opinion that the differences between the BLRPM and NSRPM are subtle and it is unlikely that empirical analysis will be able to distinguish between them. An advantage of these two cluster-based models is the efficiency of their parameter estimation procedures (Entekhabi *et al.*, 1989).

Rodriguez-Iturbe *et al.* (1987b) found that the BLRPM gave slightly more satisfactory results than the NSRPM. Khaliq and Cunnane (1996) found good agreement between observed and extreme events simulated by the MBLRPM. Hence further discussion is focussed on the BLRPM and adaptions thereof.

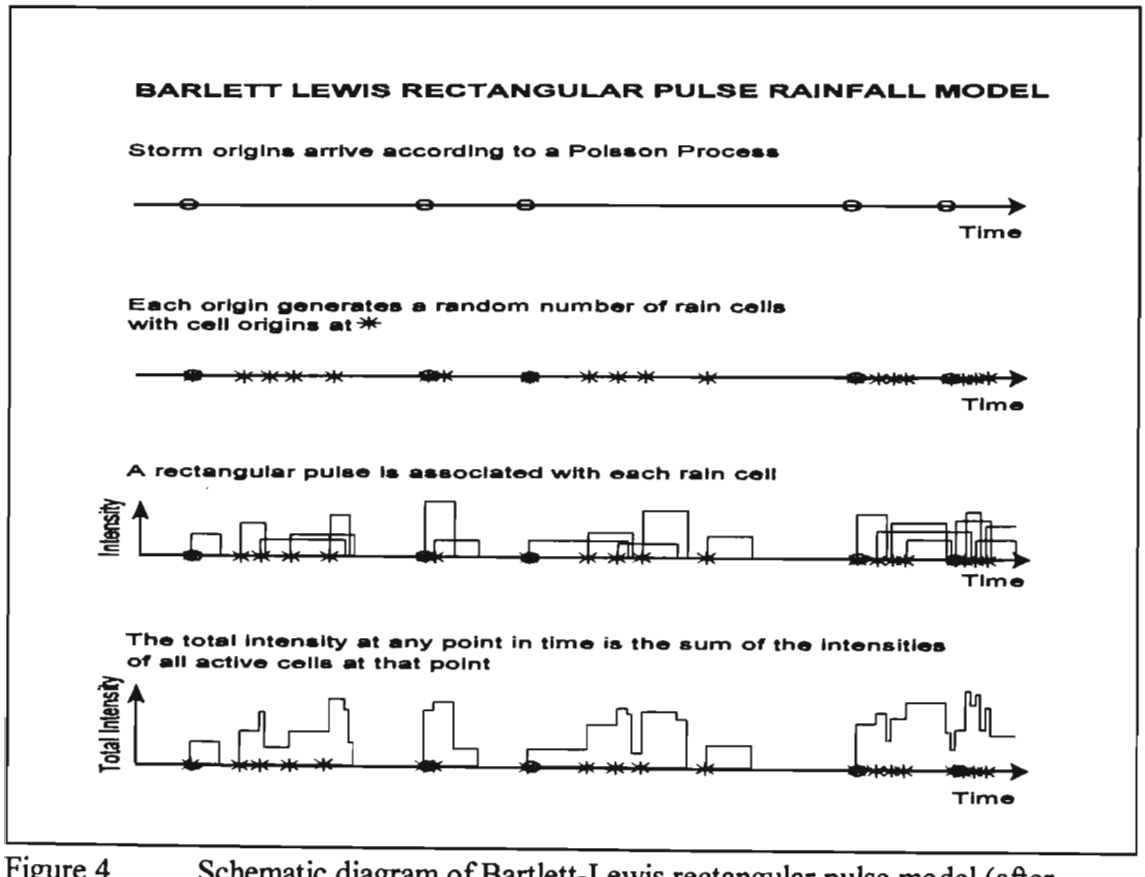


Figure 4 Schematic diagram of Bartlett-Lewis rectangular pulse model (after Cowpertwait *et al.*, 1996a)

3.2 MODIFIED BARTLETT-LEWIS RECTANGULAR PULSE MODEL

3.2.1 Procedure

The algorithm for the MBLRPM (Rodriguez-Iturbe *et al.*, 1988), which is an extension to the BLRPM (Rodriguez-Iturbe *et al.*, 1987a), is described by Entekhabi (1989), Onof and Wheater (1993), Bo *et al.* (1994), Onof and Wheater (1996) and Khaliq and Cunnane (1996) as:

- a Poisson process (parameter λ) used to model arrival rate of storm origins, and
- storm origins are followed by a Poisson process of rain cell origins with rate parameter β.
- The process of new rain cell origins terminates after a time that is exponentially distributed with parameter γ, i.e the storms have an exponentially distributed duration with parameter γ.
- The duration of the rectangular pulse of each rain cell is exponentially distributed with parameter η, and for distinct storms are assumed to be independent variables from a gamma distribution with index α and scale parameter ν, i.e $E(\eta) = \alpha/\nu$ and $var(\eta)=\alpha/\nu^2$.
- Each rain cell intensity is a random constant, exponentially distributed with mean μ_x, and
- the number of rain cells per storm C has a geometric distribution with a mean of

$$\mu_C = 1+\kappa/\phi, \qquad ...39$$

where κ and ϕ are dimensionless parameters and

$$\kappa = \beta/\eta \qquad ...40$$

$$\phi = \gamma/\eta. \qquad ...41$$

By keeping κ and ϕ constant, the mean and variance of different storms change randomly from storm to storm. Hence the mean inter-arrival interval time of cells (β^{-1}) and mean storm duration (γ^{-1}) also change randomly.

3.2.2 Characteristic Variables

The six parameters of the MBLRPM ($\lambda, \kappa, \phi, \nu, \alpha, \mu_x$) are estimated by equating the analytical expressions of certain statistical features of the rainfall process with their numerical historical counterparts (Entekhabi *et al.*, 1989). Hence at least six equations are needed.

The equations used in the derivation of the model parameters are the mean, variance, autocorrelation and dry probability. These equations, as given in Equations 42 - 47, are reproduced from Khaliq and Cunnane (1996). For the MBLRPM the mean depth of rainfall in the i-th interval of length h hours is computed as shown in Equation 42 and the variance is computed using Equation 43.

$$E\left[Y_i^h\right] = \frac{\lambda h \nu \mu_x}{\alpha - 1}\left(1 + \frac{\kappa}{\phi}\right) \qquad ...42$$

$$\begin{aligned} \mathrm{var}\left[Y_i^h\right] = {} & 2A1\left[(\alpha - 3)h\nu^{2-\alpha} - \nu^{3-\alpha} + (\nu + h)^{3-\alpha}\right] \\ & -2A2\left[\phi(\alpha - 3)h\nu^{2-\alpha} - \nu^{3-\alpha} + (\nu + \phi h)^{3-\alpha}\right] \end{aligned} \qquad ...43$$

For a lag $k \geq 1$ the covariance is

$$\begin{aligned} \mathrm{cov}\left[Y_i^h, Y_{i+k}{}^h\right] = {} & A1\left\{\left[\nu + (k+1)h\right]^{3-\alpha} - 2(\nu + kh)^{3-\alpha} + \left[\nu + (k-1)h\right]^{3-\alpha}\right\} \\ & - A2\left\{\left[\nu + (k+1)\phi h\right]^{3-\alpha} - 2\left(\nu + k\phi h\right)^{3-\alpha}\right. \\ & \left. + \left[\nu + (k-1)\phi h\right]^{3-\alpha}\right\} \end{aligned} \qquad ...44$$

where

$$A1 = \frac{\lambda \mu_c \nu^{\alpha}}{(\alpha-1)(\alpha-2)(\alpha-3)}\left[2\mu_x{}^2 + \frac{\kappa\phi\mu_x{}^2}{\phi^2-1}\right]$$

and

$$A2 = \frac{\lambda \mu_c \kappa \mu_x{}^2 \nu^{\alpha}}{\phi^2(\phi^2-1)(\alpha-1)(\alpha-2)(\alpha-3)}.$$

The probability that a period, of length h, is dry is shown in Equation 45 as

$$P\left(Y_i^h = 0\right) = \exp\left\{-\lambda h - \lambda\mu_T + \lambda G_p^*(0,0)\frac{\phi+\kappa\left[\frac{\nu}{\nu+(\kappa+\phi)h}\right]^{\alpha-1}}{\phi+\kappa}\right\} \qquad ...45$$

where μ_T is the expected duration of a single cell storm and can be approximated by

$$\mu_T \approx \frac{\nu}{\phi(\alpha-1)}\left[1+\phi(\kappa+\phi)-\frac{1}{4}\phi(\kappa+\phi)(\kappa+4\phi) + \frac{1}{72}\phi(\kappa+\phi)\left(4\kappa^2+27\kappa\phi+72\phi^2\right)\right] \qquad ...46$$

and the function

$$G_p^*(0,0) \approx \frac{\nu}{\phi(\alpha-1)}\left(1-\kappa-\phi+\frac{2}{3}\kappa\phi+\phi^2+\frac{1}{2}\kappa^2\right). \qquad ...47$$

Further characteristics describing the inter-event properties, the number and average event duration for the MBLRPM were developed by Onof and Wheater (1993) and expressed in

an easier computational form by Onof and Wheater(1994). The mean inter-event (dry) number of periods is

$$m_d^h = \frac{P(Y^h = 0)}{P(Y^h = 0) - P(Y^{2h} = 0)} \qquad ...48$$

and the average number of hourly events per month is

$$m_n = \frac{P(Y^1 = 0) \times 24 \times NM}{m_d^1} \qquad ...49$$

where

m_n = average number of hourly events in month, and

NM = number of days per month.

3.3 BARTLETT-LEWIS RECTANGULAR PULSE GAMMA MODEL

In order to improve the overestimation of daily autocorrelations and extreme events noted by Onof and Wheater (1993), Onof and Wheater (1994b) replaced the exponential distribution of *cell rainfall intensity* in the MBLRPM by a two parameter gamma distribution which would give greater flexibility in the simulation of extreme events. This modified version of the BLRPM is termed the Bartlett-Lewis Rectangular Pulse Gamma model (BLRPGM).

3.3.1 Procedure

The algorithm for the BLRPGM, a seven parameter model, is the same as that described previously for the MBLRPM, with the exception that the expressions for some of the characteristics are changed to reflect the gamma distributed cell rainfall intensity.

3.3.2 Characteristic Variables

The index and scale parameters for the gamma distributed rainfall intensity are ρ and δ respectively and the mean is calculated as shown in Equation 50. For completeness, equations for the entire set of characteristic variables for the BLRPGM are presented.

$$\mu_x = \frac{\rho}{\delta} \qquad ...50$$

The mean amount of rain in the i-th interval of length h hours is

$$E\left[Y_i^h\right] = \frac{\lambda h \mu_x \mu_c \nu}{\alpha - 1} \qquad ...51$$

and the variance is

$$\begin{aligned} \mathrm{var}\left[Y_i^h\right] = \; & 2A1\left[(\alpha - 3)h\nu^{2-\alpha} - \nu^{3-\alpha} + (\nu + h)^{3-\alpha}\right] \\ & -2A2\left[\phi(\alpha - 3)h\nu^{2-\alpha} - \nu^{3-\alpha} + \left(\nu + \phi h\right)^{3-\alpha}\right] \end{aligned} \qquad ...52$$

and for lag k ≥ 1 the covariance is

$$\begin{aligned} \mathrm{cov}\left[Y_i^h, Y_{i+k}^h\right] = \; & A1\left\{\left[\nu + (k+1)h\right]^{3-\alpha} - 2(\nu + kh)^{3-\alpha}\right. \\ & \left. +\left[\nu + (k-1)h\right]^{3-\alpha}\right\} - A2\left\{\left[\nu + (k+1)\phi h\right]^{3-\alpha}\right. \\ & \left. -2\left(\nu + k\phi h\right)^{3-\alpha} + \left[\nu + (k-1)\phi h\right]^{3-\alpha}\right\} \end{aligned} \qquad ...53$$

where

$$A1 = \frac{\lambda \mu_c \nu^{\alpha}}{\delta^2 (\alpha - 1)(\alpha - 2)(\alpha - 3)} \left[\rho(\rho + 1) + \frac{\kappa \phi \rho^2}{\phi^2 - 1} \right]$$

and

$$A2 = \frac{\lambda \mu_c \kappa {\mu_x}^2 \nu^{\alpha}}{\phi^2 (\phi^2 - 1)(\alpha - 1)(\alpha - 2)(\alpha - 3)}$$

The time distribution properties of rainfall events for the BLRPGM are not affected by the change in rainfall cell depth distribution and hence remain the same as for the MBLRPM.

3.4 PARAMETER ESTIMATION

The estimation of the six parameters for the MBLRPM is difficult, and becomes more acute for the BLRPGM, which has seven parameters (Verhoest *et al.*, 1997). Different procedures have been used to estimate the model parameters.

3.4.1 Methodology

The use of a formal statistical technique to determine parameters for rectangular pulse stochastic rainfall models, such as the maximum likelihood procedure, is not practical and probably would not be the best procedure to use (Rodriguez-Iturbe *et al.*, 1988). Rodriguez-Iturbe *et al.* (1988) suggested equating characteristic features computed from the historical data with corresponding model values, preferably computed theoretically, but failing that, by simulation. The method of moments approach, which has been frequently adopted when fitting time series models to historical data (Rodriguez-Iturbe *et al.*, 1987b; Entekhabi *et al.*, 1989; Cowpertwait, 1991; Onof and Wheater, 1993; Bo *et al.*, 1994; Onof and Wheater, 1994a; Cowpertwait *et al.*, 1996a; Verhoest *et al.*, 1997), solves a set of

simultaneous equations which relate model parameters to sampled moments (Cowpertwait *et al.*, 1996b).

The resulting set of non-linear equations can be solved simultaneously to derive parameters for the model. Different approaches can be used to solve the set of non-linear equations. Where possible, unique roots of the equations may be obtained (Rodriguez-Iturbe *et al.*, 1988; Khaliq and Cunnane, 1996). In cases where a unique solution of the non-linear equations is not possible, a scheme to minimise a defined objective function may be used. The generic format of a commonly used least squares objective function that has been used *inter alia* by Bo *et al.* (1994), Entekhabi *et al.* (1989), Cowpertwait (1991), Velghe *et al.* (1994) and Verhoest *et al.* (1997) to estimate the parameters for the models is

$$Z = \min\left[\sum_{i=1}^{N} W_i \left(\frac{F_i(X)}{F_i'} - 1\right)^2\right] \qquad ...54$$

where

$F_i(X)$ = model expression for statistic i computed using parameter vector X,

F_i' = statistic i estimated from historical data at various levels of aggregation,

N = number of statistics used in parameter determination, and

W_i = weight assigned to statistic i.

Velghe *et al.* (1994) and Verhoest *et al.* (1997) used W_i=1 for all statistics while Cowpertwait (1991) and Cowpertwait *et al.* (1996a) placed emphasis on almost exact modelling of the mean and thus set W_i=100 for the mean and used W_i=1 for all other moments.

In deriving model parameters, seasonality was taken into account by deriving parameters for each month, thus assuming data stationarity for each calendar month (Cowpertwait,

1991; Bo *et al.*, 1994). In computing the moments of the historical data, Cowpertwait (1991) and Cowpertwait *et al.* (1996a) pooled all available data for each calendar month.

Velghe *et al.* (1994) and Verhoest *et al.* (1997) used Powell's quadratically convergent algorithm to minimise the objective function (Z) while Onof and Wheater (1993) used a modified version of the Powell hybrid method.

The BLRPGM is a seven parameter model ($\lambda, \kappa, \phi, \nu, \alpha, \rho, \delta$) and Onof and Wheater (1994b) recommend fixing the δ parameter of the model owing to the difficulty in estimating the seven parameters. Despite conceding that estimating the parameters for the BLRPGM was difficult, Verhoest *et al.* (1997) did not fix any parameters and still managed to obtain a relatively good fit to the moments computed from the historical observations.

Using a different approach, Chandler (1995) developed a spectral method for estimating the parameters of point process models, which include the cluster Bartlett-Lewis cluster type models. The effect of initial conditions and the presence of many local optima necessitate that the optimisation procedure should be started from several different starting points. A general problem when estimating parameters of point type rainfall models is the lack of identifiability of model parameters (Chandler *et al.*, 1995). The disadvantages of estimating the model parameters using the method of moments is the arbitrary selection of the properties to be used and the use of only summary statistics of the data, whereas the spectral method makes more objective use of all the data and not only the summary statistics (Chandler *et al.*, 1995).

3.4.2 Moments Used

The set of characteristic variables, or moments, chosen to determine model parameters should have relatively small sampling errors and not be highly mutually correlated. Most features should be sensitive to the effects of time scale on a single cell and at least one feature should correspond to the timescale between storms (Rodriguez-Iturbe *et al.*, 1988).

The sets of variables should thus include features of both the depth process and the proportion of dry periods (Onof and Wheater, 1993). The better the estimates of analytical moments used in the parameter estimation, the better the analytical statistics at other levels of aggregation (Velghe *et al.*, 1994). The moments used in selected applications of rectangular pulse rainfall models are summarised in Table 10.

Cowpertwait *et al.* (1996a), using the NSRPM, felt that instead of fitting five moments exactly, it was better to fit more moments approximately. Khaliq and Cunnane (1996) found that the MBLRP best resembled the historical observations when more statistics than necessary (i.e. an over-determined system) were used to determine model parameters and hence suggest using 16 statistics to determine the 6 model parameters. As evident in Table 10, most applications have used short duration (hourly) resolution data in the derivation of model parameters and hence the aggregation properties of the models have been validated. Only Bo *et al.* (1994) and Cowpertwait *et al.* (1996a) have tested the disaggregation properties of the models by using longer duration data only ($\geq$ 24-h) in the derivation of parameters. This aspect was highlighted by Entekhabi *et al.* (1989), who identified the need for further research into the robustness of parameter estimation using only large aggregation periods (12 to 24-h).

In order to utilise daily rainfall data, which is much more widely available than shorter duration data, Cowpertwait *et al.* (1996a) determined parameters for the NSRPM using only daily rainfall data. The poor performance of the NSRPM when fitted using daily moments resulted in Cowpertwait *et al.* (1996b) concluding that the higher aggregation levels are unlikely to contain enough information from which the properties of the cells can be determined. Thus Cowpertwait *et al.* (1996a) developed regionalised empirical relationships between hourly variance and daily variance, thus enabling the estimation of hourly variance when only daily data were available.

Table 10 Moments used in parameter determination in selected studies

Reference	Model	Data		Fitting (MoM=Method of moments MLS= Minimisation of least squares)	Aggregation Level of Moments					
		Location	Input resolution		Mean (h)	Variance (h)	Auto-covariance (Lag-1)	Auto-correlation (Lag-1)	Dry Probability	Other
Rodriguez-Iturbe *et al.* (1987b)	NSRPM/ BLRPM	Denver, USA	Hourly	MoM Unconstrained MLS	1	1, 6		1, 6		
					1	1, 12		1, 12		
					1	1, 24		1, 24		
					6	6, 12		6, 12		
Rodriguez-Iturbe *et al.* (1988)	MBLRPM	Denver, USA Boston, USA	Hourly	Roots	1	1, 24		1	1, 24	
					1	1		1, 24	1, 24	
Entekhabi *et al.* (1989)	MNSRPM	Denver, USA	Hourly	MoM MLS	1	1, 12		1, 6, 12		
					1	1, 24		1, 6, 12		
					1	1		1, 6, 12, 24		
Onof and Wheater (1993)	MBLRPM	Birmingham, UK	Hourly	MoM 2-stage optimisation	1	1		1, 6	1, 24	m^{l}_{w} m^{l}_{d}
					1	1, 6		1	1, 24	

Reference	Model	Data		Fitting (MoM=Method of moments MLS= Minimisation of least squares)	Aggregation Level of Moments					
		Location	Input resolution		Mean (h)	Variance (h)	Auto-covariance (Lag-1)	Auto-correlation (Lag-1)	Dry Probability	Other
Onof and Wheater (1994a)	BLRPM	Birmingham, UK	Hourly	MoM 2-stage optimisation	1	1,6	1,6			m^l_w m^l_d
					1	1	1,6,12			
Onof and Wheater (1994b)	BLRPGM	Birmingham, UK	Hourly	MoM 2-stage optimisation	1	1,6	1		1,24	m^l_w m^l_d
Bo *et al.* (1994)	MBLRPM	Paducah, USA Arno, Italy	Hourly	MoM MLS	Not reported					
Velghe *et al.* (1994)	MBLRPM	Denver, USA	Hourly	MoM MLS	1	1		1, 24	1, 24	
					1	1, 24		1	1, 24	
					6	6,24		6,24	6	
					1	1,14		1,24	12	
					1	1,24		1,12,24		
Chandler (1995)	Various	South-West England	15 min	Spectral analysis	n/a					

Reference	Model	Data		Fitting (MoM=Method of moments MLS= Minimisation of least squares)	Aggregation Level of Moments					
		Location	Input resolution		Mean (h)	Variance (h)	Auto-covariance (Lag-1)	Auto-correlation (Lag-1)	Dry Probability	Other
Cowpertwait *et al.* (1996a)	NSRPM	Manston, UK	Hourly	MoM MLS	1	1,24		1,6,24	24	
					1	1,24			24	dry:dry wet:wet
Khaliq and Cunnane (1996)	MBLRPM	Valentia & Shannon Airport, Ireland	Hourly	MoM Roots/MLS	1	1,6		1	1,24	
					1	1,24		1	1,24	
					1	1		1,6	1,24	
					1	1		1,24	1,24	
					1,6, 12,24	1,6,12,24		1,6,12,24	1,6,12,24	
Verhoest *et al.* (1997)	MBLRPM	Uccle, Belgium	10 min	MoM MLS	1/6	1/6		1/6, 24	1/6,24	
Verhoest *et al.* (1997)	BLRPGM	Uccle, Belgium	10 min	MoM MLS	1/6	1/6,24		1/6,24	1/6,24	

The use of minimisation schemes and the different possible combinations of moments which may be used to determine model parameters results in non-unique parameter sets which usually all result in adequate model performance. Hence it is important to identify which model parameters are most sensitive to the scheme and moments used in parameter determination.

3.4.3 Sensitivity

The magnitude of variations between the parameter sets derived using moments from different levels of aggregation were similar to the variations obtained when changing the initial "guess" vector in the nonlinear minimisation (Rodriguez-Iturbe *et al.*, 1987b). Also using the BLRPM, Onof and Wheater (1994a) found "considerable differences" in the parameter sets determined using two different sets of moments.

Rodriguez-Iturbe (1988) reports that when two different sets of moments were used to derive MBLRPM parameters, the two sets of parameters were different, particularly the α and ν parameters. This was confirmed by Onof and Wheater (1993), who showed that with the exception of μ_x and λ, the parameters of the MBLRPM determined by two different sets of moments were very different, particularly the α and ν parameters, but that both sets of parameters could yield characteristics on the rainfall process to within 5% of historical values. In contrast to these findings, Velghe *et al.* (1994) used five different sets of moment equations and noted that there were "no striking changes in the parameter values from set to set", but perusal of their tabulated parameters indicate that large differences do occur, in particular the ν parameter.

Khaliq and Cunnane (1996) performed a sensitivity/stability analysis of parameters for the MBLRPM. As shown in Table 10, five different sets of statistics were used to estimate five sets of model parameters. The magnitude of the model parameters determined using the five different sets of statistics varied considerably. Khaliq and Cunnane (1996) concluded that μ_x and λ were the most stable and α and ν the least stable parameters. This led Khaliq and

Cunnane (1996) to use 16 moments to derive the 6 model parameters and to suggest that different starting values of α and ν should be used during optimisation.

Cowpertwait *et al.* (1996a) also noted that the parameter estimates for the NRPRM are dependent on the choice of moments used in the fitting procedure and concluded that the choice of moments "needs to be made with some discretion".

3.4.4 Optimisation

In order to improve the distribution and duration of events simulated by the BLRPM and MBLRPM and to enhance the identification of appropriate model parameters, Onof and Wheater (1993; 1994a) used a two-stage optimisation procedure with the objective function as shown in Equation 55.

$$d(i) = \sqrt{\left(1 - \frac{m_d^1(i)}{o_d^1}\right)^2 + \left(1 - \frac{m_n^1(i)}{o_n^1}\right)^2} \qquad \ldots 55$$

where

$d(i)$ = deviation at i-th iteration,
$m^1_d(i)$ = modelled mean hourly inter-event time at i-th iteration,
o^1_d = mean inter-event time of historical hourly data,
$m^1_n(i)$ = modelled mean number of hourly events at i-th iteration, and
o^1_n = mean number of hourly events in historical data.

By determining the remaining parameters for a fixed value of a poorly defined parameter, and then varying the value of the poorly selected parameter, an optimum value of the poorly defined parameter may be determined. For the BLRPM Onof and Wheater (1994a) obtained solutions for different values of β. Onof and Wheater (1993) used the MBLRPM and found that when the autocovariances were used in determining the parameters and either α or ν were kept constant, no optimum solution was found. When autocorrelations instead of

autocovariances were used, the convergence of the solution was difficult when α was fixed, but optimum solutions were obtained when ν was kept constant. The optimal values of ν obtained by an analytical solution or by simulation were very similar, but an optimum solution was not obtained for all months (Onof and Wheater, 1993). The optimised parameter set improved the simulation of inter-event and duration characteristics, but the optimised parameters showed no improvement in the simulation of extreme events.

A similar two stage optimisation procedure was used by Onof and Wheater (1994b) to optimise the parameters for the BLRPGM. The parameter δ (the scale parameter for the Gamma distribution) was incremented until an optimum (δ_1) solution was determined. A very good reproduction of extremes was obtained when the δ was optimised (δ_2) such that the mean of the 1 h and daily AMS of the simulated series best approximated the historical values. Although δ_2 was determined by simulation, as no analytical expressions are possible, the optimised values δ_1 and δ_2 were very similar for most months. This led Onof and Wheater (1994b) to conclude that the optimised δ_1 data set would provide a good simulation of the extreme values at the hourly and daily levels.

Onof and Wheater (1994a) noted that although there were some discrepancies between analytical and simulated values of the characteristic variables, the use of analytical values in the optimisation procedure was acceptable.

3.4.5 Daily Parameters

Onof and Wheater (1993) investigated whether a smoother representation of the parameters over the year was possible and if the coefficients of this representation could be used for regionalisation of parameters. The use of a polynomial produced very satisfactory results and could thus be used to yield more realistic results for periods which are not calendar months.

3.5 GOODNESS-OF-FIT CRITERIA

Various tests have been used to assess model performance. Generally both analytical and simulated values of certain characteristics of the rainfall process are compared with historical values (Onof and Wheater, 1993; Onof and Wheater, 1994a). Bo *et al.* (1994) used the mean sum of squares of the difference between the model estimated and observed mean, variance, autocorrelation and dry probability statistics for various levels of accumulation as shown in Equation 56.

$$F(j) = \frac{1}{N_L}\sum_{i=1}^{N_L}\left[Fit_{(i,j)} - His_{(i,j)}\right]^2 \qquad ...56$$

where

$F(j)$ = measure of goodness of fit for j-th statistic, e.g. mean (j=1), variance (j=2), autocorrelation (j=3), dry probability (j=4),

$Fit_{(i,j)}$ = value of model computed j-th statistic at aggregation level (duration) i,

$His_{(i,j)}$ = value of j-th statistic computed from historical data at aggregation level i, and

N_L = number of different aggregation levels used.

Verhoest *et al.* (1997) used the goodness-of-fit statistic (S) defined by Velghe *et al.*(1994) as shown in Equation 57.

$$S = \frac{100}{m \times N_L}\sum_{j=1}^{m}\sum_{i=1}^{N_L}\left|\left(1 - \frac{Fit_{(i,j)}}{His_{(i,j)}}\right)\right| \qquad ...57$$

where

m = number of moments or statistics considered.

Cowpertwait *et al.* (1991) generated a 20 year series of hourly rainfall and used t-tests to compare simulated and historical moments. Cowpertwait *et al.* (1996a) validated the NSRPM by:

- visual comparison of historical data and simulated time series,
- the crossing properties of the time series; and
- the hourly and daily extremes.

Although not explicitly detailed in the literature, the "model computed statistic" can be either an analytic or simulated value. The theoretical expressions for the moment, if available, can be computed for a given set of parameters and compared to the equivalent moment computed from the historical data. The alternative, and the only option if the theoretical expression for the statistics are not available, is to compute the statistics from a synthetic time series generated by the model. Both of these options were used by Khaliq and Cunnane (1996). Analytical moments were identified, at different levels of aggregation, which differed from the historical moments by more than ± 2SE, where SE is the estimated standard error. In addition, properties computed from a 200 year record simulated by the model, with a particular parameter set, were compared to those computed from the historical data. However, no estimate was made of the variation in the synthetic series as a result of the stochastic rainfall generation process, i.e. the sampling variation of historical data was not compared to the variation due to the stochastic process.

Features not used in the determination of parameters can be used to determine the goodness-of-fit (Rodriguez-Iturbe *et al.*, 1988). Other characteristics used by Khaliq and Cunnane (1996) to assess the performance of the model include probabilities of observing small rainfall amounts, distributions of rainfall depth and intensity for given durations, event profiles and distributions of monthly number of rainfall events, dry durations and wet durations. Rainfall events were defined as a sequence of wet hours, preceded and followed by at least one dry hour.

3.6 REGIONALISATION OF PARAMETERS

Cowpertwait *et al.* (1996a) derived linear regressions between h-hourly (h<24) and daily variance for 27 stations in the UK. Of the 27 stations which had hourly rainfall data, 66% had record lengths of between 5-10 years and the remainder had record lengths less than 30 years (Cowpertwait *et al.*, 1996b). Using both daily moments and variances for durations < 24 h, derived from the regressions, when fitting the NSRPM resulted in a reasonable simulation of hourly data (Cowpertwait *et al.*, 1996a). It was concluded that the regionalised model could estimate rainfall properties that were within the sampling error expected in a 20 year historical record of daily rainfall data.

Cowpertwait *et al.* (1996b) developed regressions at 112 sites in the UK between NSRPM parameters and both location dependent variables that influence rainfall and harmonic variables. At sites where no short duration data were available, four of the NSRPM parameters were estimated using these regressions and the fifth parameter was estimated using the mean of a nearby daily rainfall station and the four derived parameters. In order to simulate durations as short as 5 minutes, a stochastic disaggregation model was developed which used hourly time series as input.

3.7 MODEL VALIDATION

Model performance can be assessed by checking the model's ability to reproduce rainfall properties not used in the fitting procedure, but which are considered important (Rodriguez-Iturbe *et al.*, 1988; Cowpertwait *et al.*, 1996a).

3.7.1 Neyman-Scott Rectangular Pulse Model

Cowpertwait *et al.* (1996a) compared the means and standard deviations of the proportions of time that the historical and simulated rainfall exceeded various depths. The NSRPM was

found not to simulate the mean proportion of events less than 1 mm well. Using an exponential distribution for cell intensity, Cowpertwait *et al.* (1996a) found that the NSRPM under-simulated historical extreme events for return periods greater than 5 years. The use of a Weibull distribution to model cell intensity did not necessarily improve the simulation of extreme events. Cowpertwait *et al.* (1996a) conclude that the inconsistent simulation of extreme events by the NSRPM "may be due to an over-simplification in the parameterisation of the model" and that consequently a "good fit to the extreme values is unlikely to be achieved consistently using the present form of the model".

3.7.2 Original and Modified Bartlett-Lewis Rectangular Pulse Models

Rodriguez-Iturbe *et al.* (1987b) applied the BLRPM to a 27 year record of hourly rainfall data for one month from Denver, USA, and found that the model was able to preserve the rainfall depth statistics and extreme values of rainfall, but did not reproduce the proportion of dry level states satisfactorily. The MBLRPM, which allowed random variation from storm to storm of the exponential parameter of the distribution of cell duration, enabled the model to reproduce the proportion of dry states for various time intervals (Rodriguez-Iturbe *et al.*, 1988).

Rodriguez-Iturbe *et al.* (1988) found that the MBLRPM underestimated the hourly and 24-hourly extremes for return periods greater than the record length. By plotting the cumulative distribution of the modelled and historical extreme values for both the 1 and 24 h aggregation levels, it was apparent that the MBLRPM underestimated the extreme values for return periods greater than approximately 10 years.

Onof and Wheater (1993) used the MBLRPM to improve the simulation of rainfall in the UK. Generally the second-order properties of the data were well reproduced by the model. In addition, dry periods for all time scales (hourly to daily) and daily rainfall depths were also well reproduced by the model. The MBLRPM improved the autocorrelations for lags > 12 h, inter-event intervals (dry periods), the duration and number of events when

compared to the BLRPM, but autocorrelations were still not adequately simulated by the MBLRPM. In addition, the design rainfall for return periods longer than the length of the data set were not reproduced well.

Bo *et al.* (1994) showed that both the aggregation and disaggregation of rainfall using the MBLRPM were satisfactory and that, using readily available daily rainfall data to determine model parameters, statistics for finer time scales of up to 1 h could be reproduced using the MBLRPM.

Khaliq and Cunnane (1996) used the MBLRPM to successfully model point rainfall with parameters derived from a 45 year record from Valentia, Ireland and from a 38 year record from Shannon Airport, Ireland. Two hundred years of synthetic data were simulated. Generally the autocorrelations for lags ranging from 1 to 24 in the hourly data and for lags from 1 to 10 in the 24 h data were adequately simulated. Probabilities of no rain for accumulation periods great than 24 h were generally over-simulated by the model. Khaliq and Cunnane (1996) found that, whilst the simulation of extreme events by the MBLRPM was dependent on the moment set used in the derivation of the model parameters, the model generally under-simulated hourly extreme events for return periods greater then 5 years. However, the model generally reproduced the 24 h extreme values well for most months.

Velghe *et al.* (1994) compared the performance of the NSRPM, MNSRPM, BLRPM and MBLRPM for the Denver, USA data used by Rodriguez-Iturbe *et al.* (1987b). The analytical (theoretical) and simulated statistics were compared to the statistics computed from the historical data. The NSRPM model was found to perform better than the BLRPM. This was partially attributed by Velghe *et al.* (1994) to the better fit of the analytical values (lower Z) for the NSRPM. Similar to the finding by Rodriguez-Iturbe *et al.* (1988), Velghe *et al.* (1994) found that the modified versions of the NSRPM and BLRPM gave better estimates of dry (zero depth) probabilities at higher levels of aggregation and better estimates of extreme values, but that the correlation structure of the original models fitted the historical values better. The MBLRPM was found by Velghe *et al.* (1994) to differ more from the historical statistics than the NSRPM, and the MBLRPM was also more sensitive

to the sets of moment equations used in parameter estimation. When zero depth probabilities were used at more than one level of aggregation in the moment equations used to determine parameters, it was found the zero depth probabilities were well preserved at all levels of aggregation, but due to the limited number of moments used in the estimation, the second order moments were not fitted well. When only one or no zero depth probabilities were used in the moment equations, the zero depth probabilities were overestimated and the second order moments were better represented at all levels of aggregation. Velghe *et al.* (1994) found that the simulation of extreme values by the MBLRPM was not sensitive to different moment equations, but concluded that the major drawback of applying the MBLRPM was the sensitivity of the performance to the selected moment equations used in the determination of the model parameters. For all models, hourly design rainfall depths were generally underestimated for longer return periods but, for corresponding return periods, were better simulated for longer durations.

3.7.3 Bartlett-Lewis Rectangular Pulse Gamma Model

Onof and Wheater (1994b) used a 38.5 year record of hourly rainfall from Birmingham, UK and found that after optimising the δ parameter, the BLRPGM simulated the extreme events well at both the hourly and daily time scales. The difficulty in estimating the seven parameters for the model, and the success of the BLRPGM, led Onof and Wheater (1994b) to conclude that future research effort should concentrate on widespread applications of the models and regionalisation of the parameters of the model, and not on developing models with more parameters.

Verhoest *et al.*(1997) compared the BLRPM, MBLRPM and BLRPGM using a 27 year period of record of 10 min rainfall data from Uccle, Belgium. Based on first and second order moments computed from 100 years of generated synthetic rainfall series, it was shown that all three models performed adequately and that the MBLRPM best simulated the second order moments of the historical data. It was found that none of the three models were able to satisfactorily model the extreme value behaviour of the data, particularly for

short duration (10 to 200 min) events where the extreme events were under-simulated. However, Verhoest *et al.*(1997) used a 24 h period of no rain to extract storms and showed a good agreement between the mass curves generated by the MBLRPM and the observed rainfall mass curves. The mean length of the synthetic storm was generally found to be shorter than for the historical series. This led to the conclusion that the cluster-based models produce individual rain cells more clustered than the historical series.

3.8 CHAPTER CONCLUSIONS

It is apparent from the literature that cluster based rectangular point rainfall models that use a Poisson process to simulate storm and cell arrival times can adequately reproduce most of the properties of historical rainfall data. Varied performances of the simulation of extreme events, which is of most interest to this study, have been reported in the literature. Cowpertwait *et al.* (1996a) report that performance of the NSRPM was inconsistent. For the Bartlett-Lewis based models, the simulated design rainfall values were generally poor for shorter durations ($\pm \leq$3-h) and for return periods longer than the historical record, but encouraging for longer durations and return periods up to the record length (Rodriguez-Iturbe *et al.*, 1988; Onof and Wheater, 1993; Velghe *et al.*, 1994; Khaliq and Cunnane, 1996; Verhoest *et al.*, 1997). However, Onof and Wheater (1994b) obtained satisfactory results using the BLRPGM, after optimising the δ parameter, for both hourly and daily durations and return periods up to 200 years. Hence the results, presented in Chapter 7, of using stochastic cluster-based rainfall models in South Africa to estimate design rainfalls, are focussed exclusively on the MBLRPM and BLRPGM.

Most of the studies reported in the literature used data from only one station and, in some cases, used only data from individual months, e.g. Rodriguez-Iturbe *et al.* (1988). It is assumed that the limited amount of data used are from selected, well maintained stations with good, well checked records. Hence, some of the conclusions pertaining to the performance of the models are only applicable to the site and data used, and may not be

generally applicable to different locations and with the use of "operational" data, which may not be as error free as those stations used in the studies reported in the literature.

The inherent stochastic variability of the cluster-based rainfall models has not been demonstrated explicitly in the literature reviewed. Most studies have generated a long sequence of synthetic rainfall (e.g. 200 years) and have estimated design rainfall values from this series. In the application of the stochastic rainfall models to data from South Africa, presented in Chapter 7, the stochastic variability of design rainfall values computed from the synthetic rainfall series is shown explicitly.

This chapter concludes Part A, in which the theoretical framework is set for the remainder of the thesis, with results presented in Chapters 4 - 7. Chapter 4 following in Part B, details the compilation of a short duration rainfall database for South Africa and highlights errors and inconsistencies in the data. The database is used both to estimate short duration design rainfalls using the techniques presented in Chapter 2, with results presented in Chapters 5 and 6, and to estimate the parameters of the cluster-based models discussed in this chapter. The results of estimating design rainfalls from the synthetic rainfall series generated by the stochastic cluster-based rainfall models are presented in Chapter 7.

PART B

APPLICATION AND DEVELOPMENT OF TECHNIQUES

In Part B, the results of the study are presented in Chapters 4, 5, 6, and 7. In Chapter 4, the compilation of a short duration rainfall database is described and techniques are developed and assessed for identifying and removing errors such as zero and negative time steps from the data. The consistency of the digitised data are evaluated by comparing daily rainfall totals computed from the digitised and standard daily rainfall databases. A case study on the effect of missing data on the estimation of design rainfall depths is also presented.

In Chapter 5 relatively homogeneous regions of design rainfall frequency distribution in South Africa are identified and the results of a regional index storm based approach to design rainfall estimation is presented. Regional regression equations are developed to estimate the index storm for 24 h duration events as a function of site characteristics, thus enabling the index storm based approach to be applied at any ungauged site in South Africa. In order to estimate short duration design storms from daily rainfall data, hypotheses were proposed which combine the properties of homogeneous regions, where the distribution of the scaled Annual Maximum Series (AMS) is assumed to be the same at each site in the region, with the scaling characteristics of the AMS. The hypotheses and results of applying the hypotheses at selected regions and sites in South Africa are presented in Chapter 6.

In Chapter 7 the results from generating stochastic rainfall time series with Bartlett-Lewis Rectangular Pulse rainfall models and estimating design storms from the synthetic rainfall series are presented. Both parameter optimisation techniques and a procedure for determining the model parameters using only daily rainfall data are developed and evaluated. In addition, the stochastic variability is used to estimate confidence limits for the design storms and the temporal distribution of synthetic storms estimated at selected sites are presented. Two interesting case studies are also presented which evaluate two approaches that can be adopted to estimate short duration design storms at sites which only have a short period of observed data available.

CHAPTER 4

ESTABLISHMENT OF A SHORT DURATION RAINFALL DATABASE FOR SOUTH AFRICA

In order to establish a short duration (≤ 24 h) rainfall database for South Africa it was necessary to assess the availability of automatically recorded rainfall data. Questionnaires were distributed to numerous organisations, which included Government Departments, Universities and local authorities, requesting information on rainfall data collected by the organisations. The organisations which responded positively with relevant information were requested to provide the data which were included in the database. In numerous cases the rainfall data were still in chart form and had to be manually digitised for entry into a computer. The organisations which contributed relevant and useable data to the database, and the number of stations which were made available, are listed in Table 11. In total data from 412 stations were obtained. The distribution of record lengths of the 412 stations in the database is shown in Figure 5 and the locations of stations with record lengths ≥ 10 years is shown in Figure 6.

Table 11 Organisations which contributed short duration rainfall data

Organisation	Number of stations
Department of Agricultural Engineering, University of Natal (DAEUN)	24
Council for Scientific and Industrial Research (CSIR)	4
Rhodes University (RU)	28
South African Sugar Association Experiment Station (SASEX)	4
University of the Witwatersrand (Wits)	3
South African Weather Bureau (SAWB)	334
Cape Town City Engineer's Department (CTCE)	2
University of Zululand (UZ)	13

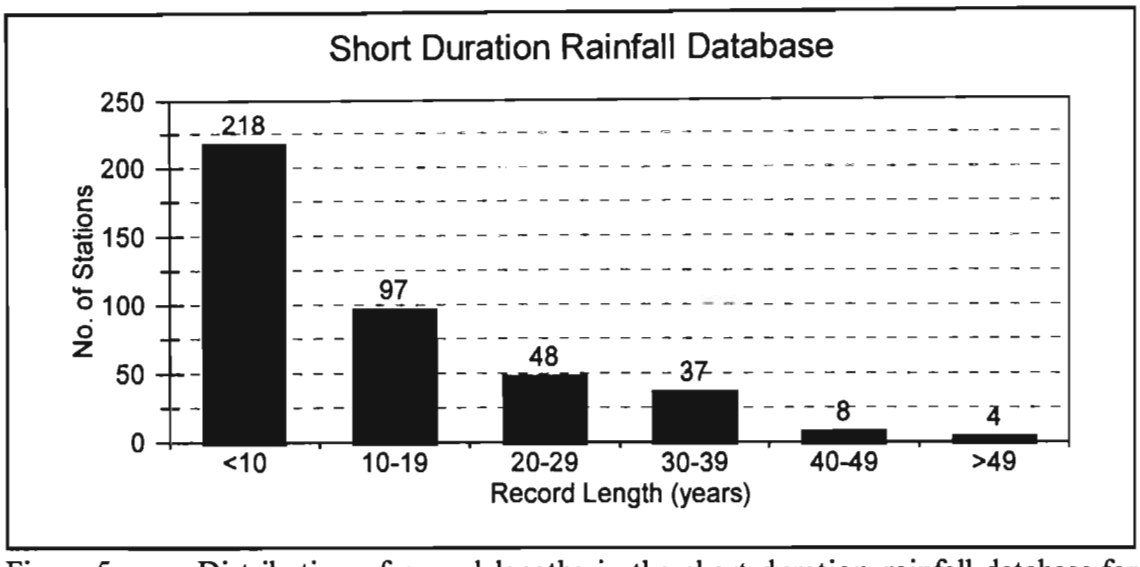


Figure 5 Distribution of record lengths in the short duration rainfall database for South Africa

As shown in Table 11 the majority of stations in the database were contributed by the SAWB. Errors, such as negative and zero time steps, were found in the data from most of the organisations which contributed processed rainfall data to the database. A zero time step occurs when consecutive data points are assigned the same time of day while having an increase in rainfall and thus create an infinite intensity. With the exception of the SAWB data, these errors were relatively few, with usually only one or two errors in the entire data set for a particular station. However, numerous errors were encountered in the data obtained from the SAWB. Hence the cause of these errors had to be established and procedures developed in order to correct the errors and allow the continuous processing of data. The term "correction of errors" used in this chapter refers to the adjustment of data points in order to eliminate the errors and allow continuous processing of the data and does not refer to the correction of data in the sense of infilling missing data points.

An analysis of the probable causes and suggested procedures to correct errors in the SAWB digitised rainfall database are investigated in the following section. This is followed by some consistency checks on the digitised data, which include sections on comparing the digitised and manually extracted extreme events, the frequency and magnitude of differences between digitised and standard, non-recording raingauge daily rainfall totals and an analysis of the impact of incomplete data on the estimation of design rainfall.

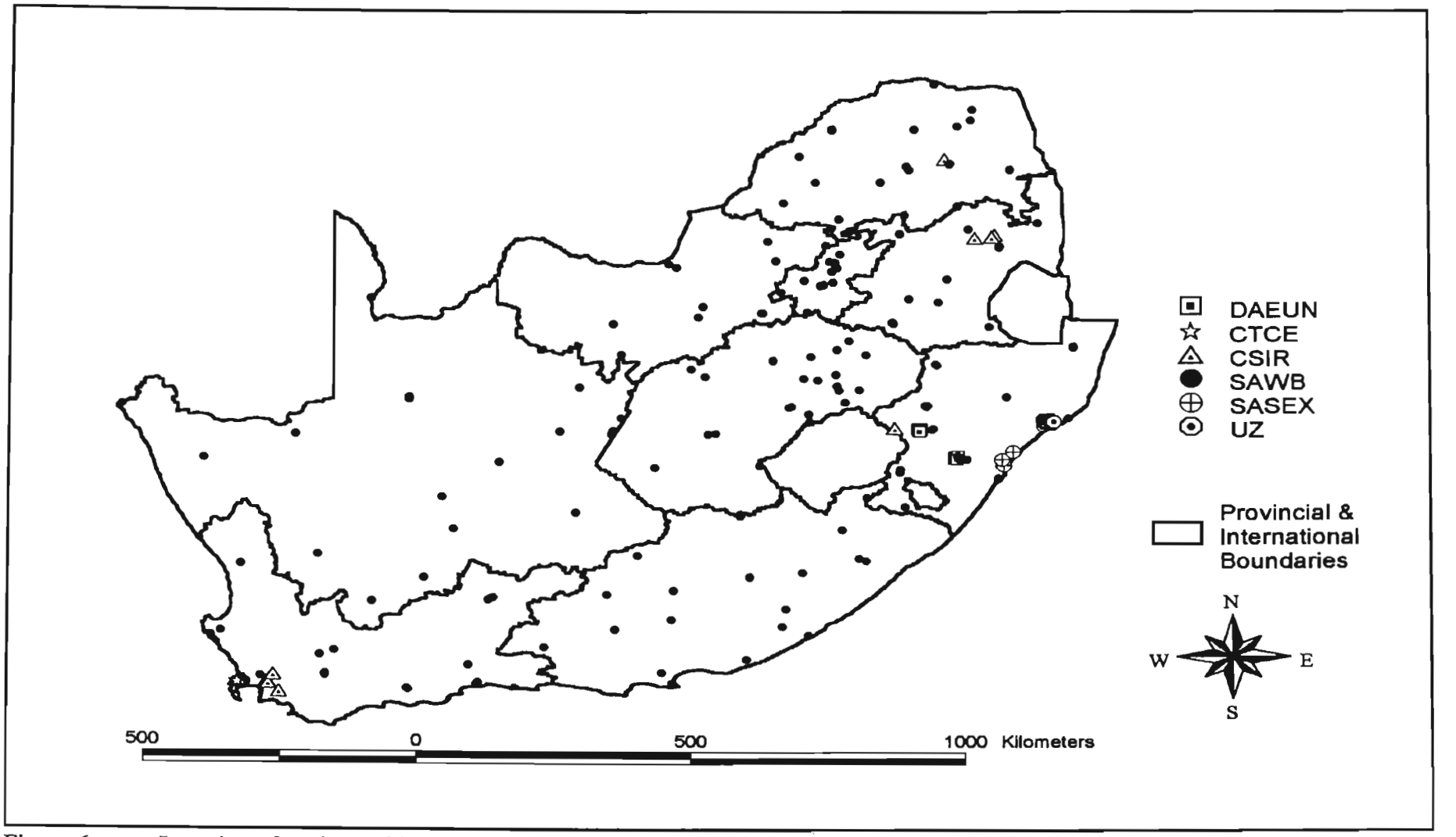


Figure 6 Location of stations with record lengths ≥ 10 years in the short duration rainfall database for South Africa (Acronyms detailed in Table 11)

4.1 ERRORS IN SAWB DATA AND DATA ADJUSTMENT PROCEDURES

The most common errors found in the database consist of rainfall events with negative or zero time steps. As indicated in Figure 7 the majority of errors found in the database are a result of negative time steps.

4.1.1 Sources of Errors

As shown in Figure 7(a), the most frequently occurring negative time step errors are those associated with a decrease in the digitised depth of rainfall (labelled negative & less), followed by those associated with raingauge siphons (negative siphon), equal rainfall depth (negative & equal) and increasing digitised rainfall depth (negative and increase). It is concluded from the intra-daily temporal distribution of the occurrences of the negative time step errors associated with decreasing digitised rainfall depths, as shown in Figure 7(b), that the majority of these errors are a result of not synchronising the time at the end of one daily chart with the beginning time of the following chart. The possible causes of the negative time step errors which occur at chart changes may be incorrect digitising, autographic raingauge clock errors and possible incorrect setting or failing to record the time at which the chart was placed on and removed from the gauge. An analysis of the magnitude of the time differences of negative time step errors is given in Figure 7(c), with the majority of negative time step errors being less than 30 minutes. Examination of the intra-daily temporal distribution of the occurrences of zero time step errors showed that these errors occurred randomly throughout the day and were thus probably a result of incorrect chart digitisation. From Figure 7(d), it is seen that the magnitude of the differences in the rainfall amounts associated with the majority of the zero time step errors is less than 2 mm. The large number of errors contained in the database makes the task of manually correcting the database extremely time consuming, which prompted the development of automatic correcting procedures.

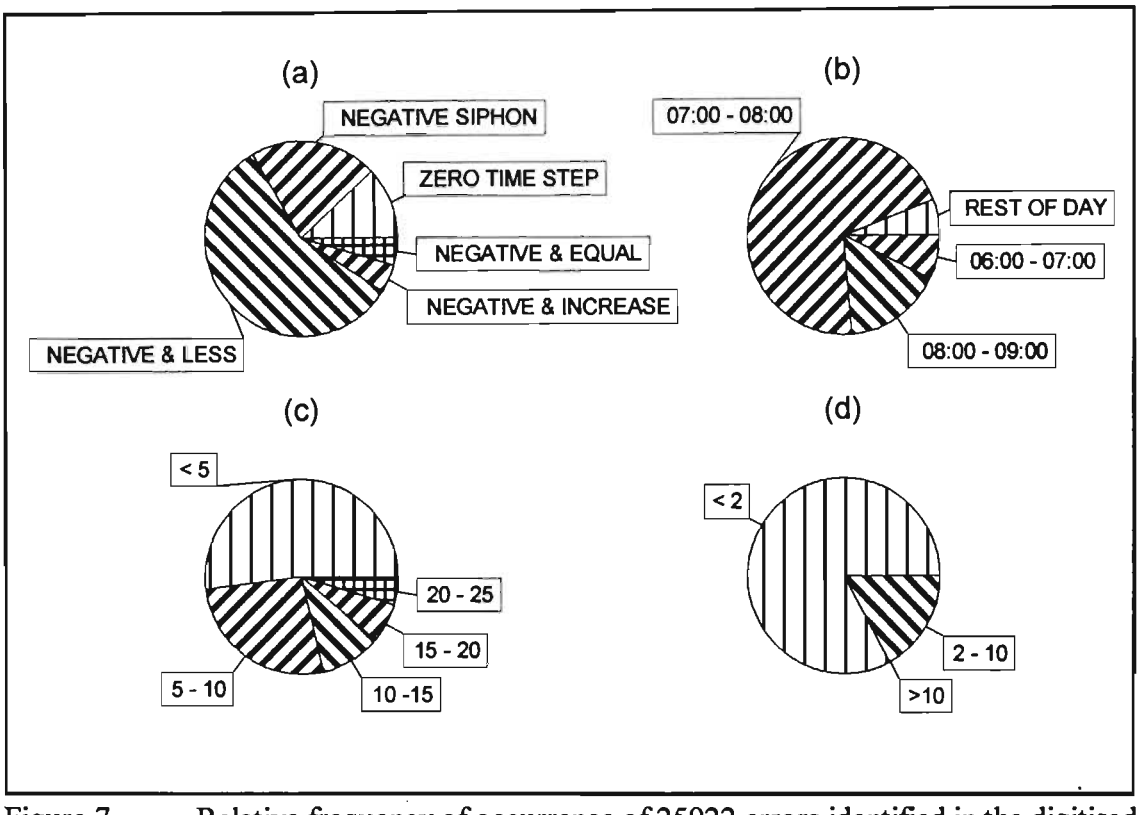


Figure 7 Relative frequency of occurrence of 25922 errors identified in the digitised rainfall database from 29 SAWB stations for the period 1960 to 1990:
(a) Occurrences of negative and zero time steps
(b) Temporal distribution of negative time steps associated with a decrease in digitised rainfall
(c) Magnitude of negative time steps (minutes)
(d) Difference in rainfall depths (mm) of data points associated with zero time steps

4.1.2 Data Correction and Adjustment Procedures

4.1.2.1 Principles applied

The principles used to correct the data were guided by the analysis of errors, such as contained in Figure 7. Each "type" of error was identified, and appropriate remedial actions were performed. The principles applied in these actions are illustrated for a negative time step error associated with an increase in digitised rainfall, as shown schematically by the solid line in Figure 8.

It is assumed that points 1 and 4 are correct and either points 2 or 3 or both are incorrect. One alternative to correcting this error is to delete either points P2 or P3 such that the minimum rainfall intensity is introduced (either I13, the intensity between P1 and P3, or I24, the intensity between P2 and P4). In this case P3 will be deleted and the intensity I24, shown by the dotted line, is introduced into the data. This approach has been termed the Lowest Intensity Adjustment (LIA). An alternative to this technique is to delete either P2 or P3 such that the maximum rainfall intensity is introduced in the database. This approach has been termed Maximum Intensity Adjustment (MIA). A third alternative is to replace P2 and P3 with a single point containing averaged time and rainfall values, and has been termed Average Intensity Adjustment (AIA).

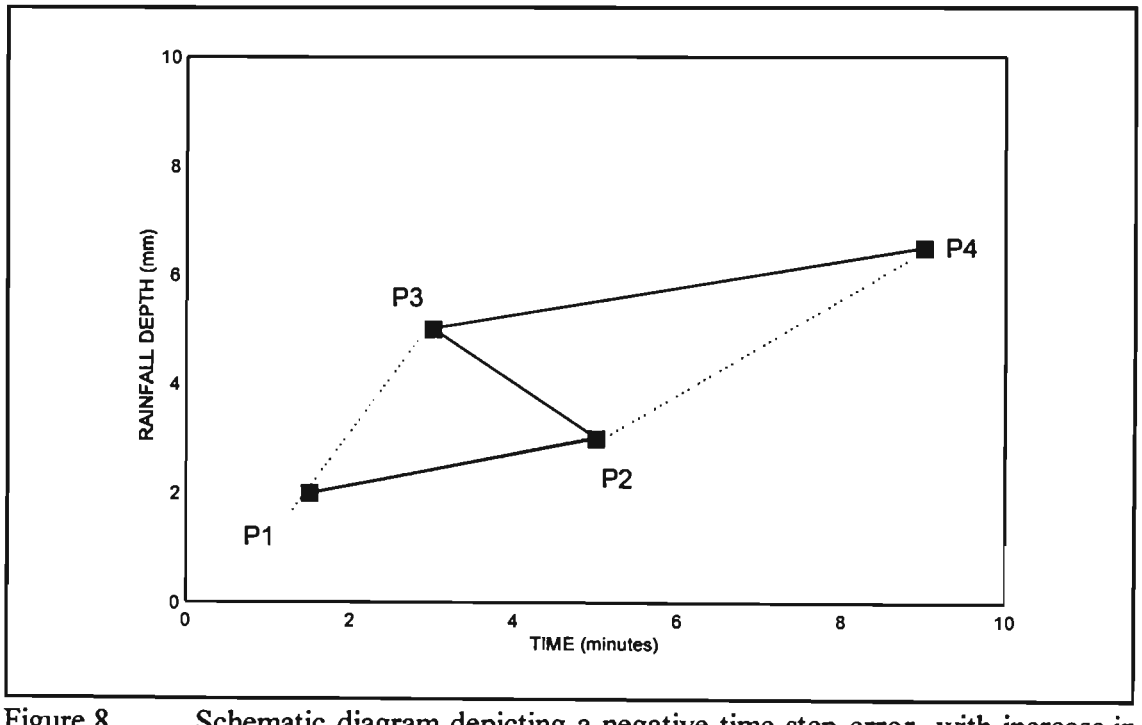


Figure 8 Schematic diagram depicting a negative time step error, with increase in digitised rainfall (P1, P2, P3, P4 are consecutive digitised points in the data)

Three sets of Annual Maximum Series (AMS) were extracted from the database corrected by either the LIA, AIA or MIA procedures. In addition, two AMS were extracted that excluded corrected data points. The first excluded all erroneous data points from the

database prior to the extraction of the AMS and was termed EXPOINT. The data were scanned sequentially and any data point causing an error (e.g. P3 in Figure 8) was discarded, and the AMS was extracted from the remaining data points. The second method excluded from the AMS any event that had any errors contained in the data within the duration of the event and was termed EXEVNT. In order to select which of the LIA, MIA, AIA, EXPOINT or EXEVNT were appropriate procedures, statistical tests are utilised in Section 4.1.5 which test if the 5 different methods of ensuring continuous processing of the data result in significantly different AMS.

4.1.2.2 Chart changes

In some cases the time-off recorded on a chart is often later than the time-on for the following chart. For example, at SAWB Station 0059572 the chart starting on 01/03/42 has a recorded time-off on 02/03/42 at 09:00, while the chart starting on 02/03/42 has a recorded time-on of 08:50. In addition, the last digitised point on a chart is often later than the recorded time-off. For example, at Station 0059572 on 19/12/40 the recorded time-off is 08:30, but the last digitised point on the chart is 08:32.

In addition on some charts, generally for more recent years, the system of recording the correct time-on and time-off, which can then be used to correct the chart time-on and time-off if the clock lost or gained time, seems to have been abandoned. For example, random checks in years 1975, 1980, 1985 and 1990 for Station 0059572 reveal that the time-on and time-off was consistently 8:00 on every day, thus indicating that this is probably not the correct time noted by the observer. As a result the time-on and time-off values cannot be used to correct any time errors on the chart.

For the above reasons it was considered that the recorded time-on and off of charts were too unreliable to use in adjusting negative time steps arising as a result of time clocks running too fast. It was thus assumed that the time when the chart was put on is correct and hence the difference between the first digitised point and the last point digitised on the

previous chart is used to establish the magnitude of the clock time error. This assumes that the clockwork mechanism is not running fast or slow.

4.1.2.3 Automated correction

Owing to the vast number of errors found in the SAWB digitised database and thus the need to automate the correction process, the five correction methods (MIA, AIA, LIA, EXPOINT, EXEVNT) were used to create five different sets of AMS.

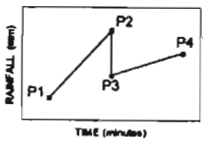
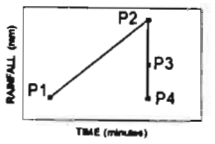
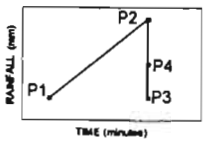
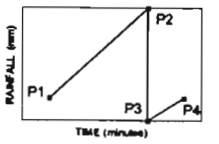
The data points were scanned sequentially and the action undertaken and software routine invoked whenever an error was detected is shown in Table 12. As detailed in the following section, the automated correction procedures were only undertaken after some manual editing had been performed.

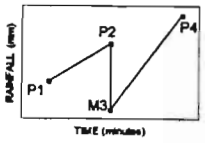
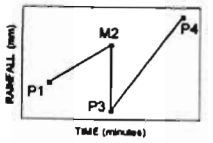
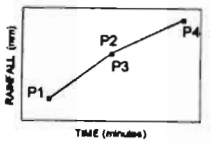
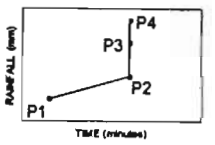
As indicated in Table 12, whenever an adjustment was made, the affected data points were assigned a code. These indicate time adjustments (t), siphon adjustments (s) and corrections (c) to data points where the cause of the error is unknown. A clear distinction is drawn between adjustments, where the probable cause of the error is known, and errors, where the cause of the error is unknown.

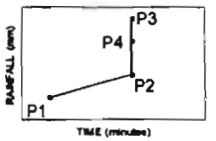
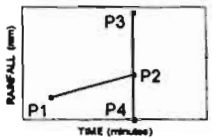
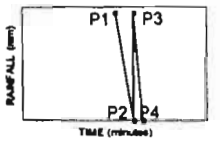
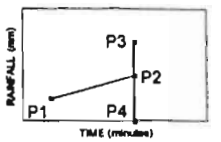
4.1.2.4 Manual correction

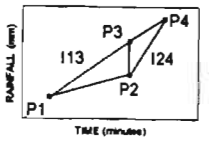
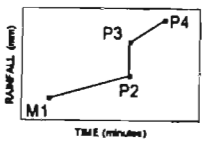
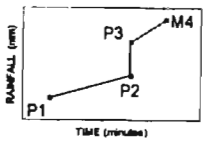
Despite the extensive automatic correction procedures, it was found that using only automatic procedures to correct large negative steps (> 30 minutes) resulted in unrealistic corrections. These large negative time steps were largely a result of what appeared to be either spurious points or the re-digitisation of portions of the same chart. These errors were thus investigated individually and corrected manually.

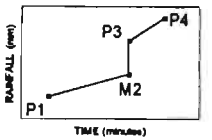
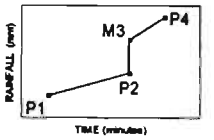
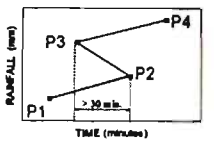
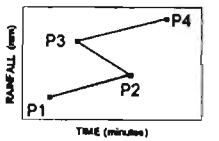
Table 12 Automatic adjustment procedures

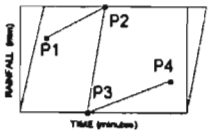
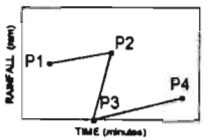
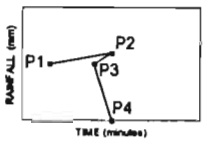
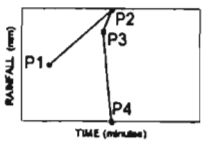
Case	Trace	Error/ Suspected Cause	Assumption(s)	Method	Flag	Action	Routine Invoked
Equal time		Digitising error	Either P2 or P3 incorrect	EXPOINT	c	Delete P3	DISCAD
Equal time and decrease in rainfall trace		Check siphon top and bottom value have been placed in incorrect column		All	c	Delete P3	DISCAD
				All	c	Delete P4	DISCAD
		Siphon		All		None	

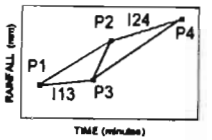
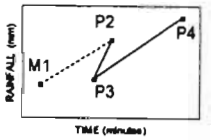
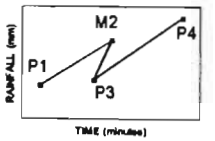
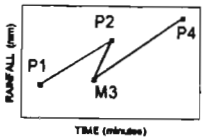
Case	Trace	Error/ Suspected Cause	Assumption(s)	Method	Flag	Action	Routine Invoked
Equal time and decrease in rainfall trace (continued)		Point 3 is a missing code		All	c	Move point 3 such that $T_3=T_2+1$	MINADD
		Point 2 is a missing code		All	c	Move point 2 such that $T_2=T_3-1$	MINSUB
Equal times and equal values		Digitising error Point 2= Point 3	Duplication of same point	All	c	Delete P_3	EQUAL
Equal time and increase in rainfall trace		Digitising error		All	c	Delete P3	DISCAD

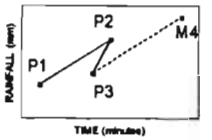
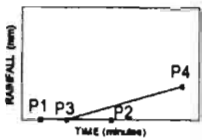
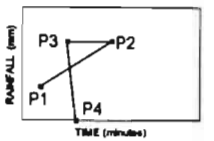
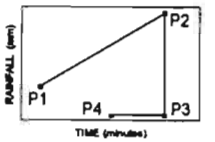
Case	Trace	Error/ Suspected Cause	Assumption(s)	Method	Flag	Action	Routine Invoked
Equal time and increase in rainfall trace (continued)		Digitising error		All	c	Delete P4	DISCAD
		Digitising error at siphon	P2 is not part of siphon	All	c	Delete P2	DISCAD
		Digitising error of 2 consecutive siphons during intense event T2=T3		All	c	Move point 3 such that T3 = T3+1	MINADD
		Digitising error at change of chart		All	c	Delete P2	DISCAD

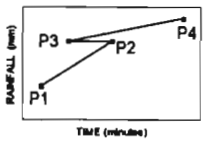
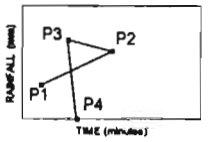
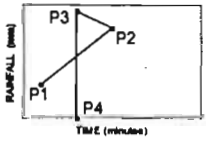
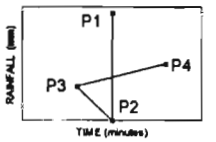
Case	Trace	Error/ Suspected Cause	Assumption(s)	Method	Flag	Action	Routine Invoked
Equal time and increase in rainfall trace (continued)		Digitising error	Either P2 or P3 is an error	MIA	c	If I13 > I24, delete P2	FOUR
						If I24 > I13, delete P3	
				AIA	c	Average P2 and P3	FOUR
			Either P2 or P3 is an error	LIA	c	If I13 > I24, delete P3	FOUR
						If I24 > I13, delete P2	
				EXPOINT	c	Delete P3	FOUR
		Digitising error P1 is either a missing code or P2 is the first point in file	Either P2 or P3 is incorrect	MIA	c	Delete P3	DISCAD
				AIA	c	Average P2 and P3	RAINAV
				LIA	c	Delete P2	DISCAD
		Digitising error P3 is either the last point in the file or P4 is a missing code	Either P2 or P3 is incorrect	MIA	c	Delete P2	DISCAD
				AIA	c	Average P2 and P3	RAINAV
				LIA	c	Delete P3	DISCAD

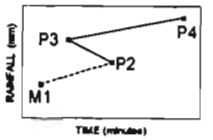
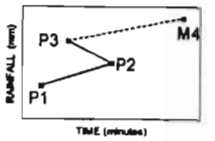
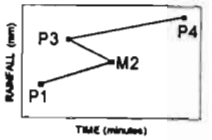
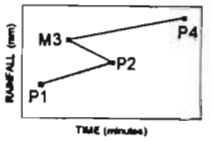
Case	Trace	Error/ Suspected Cause	Assumption(s)	Method	Flag	Action	Routine Invoked
Equal time and increase in rainfall trace (continued)		P2 is a missing code		All	c	Move M2 such that T2=T3-1	MINSUB
		P3 is a missing code		All	c	Move P3 such that T3=T2+1	MINADD
Negative time step		Unknown Manual correction if negative step > 30 minutes		Manual Correction			
		Digitising error	Any point creating an error is deleted	EXPOINT	c	Delete P3	DISCAD

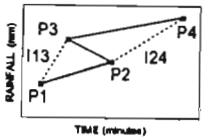
Case	Trace	Error/ Suspected Cause	Assumption(s)	Method	Flag	Action	Routine Invoked
Negative time step with decrease in trace		Parallax error due to chart placement on drum, distorted frame or incorrect digitising of a siphon	All points on chart are affected by the distortion	MIA, LIA, AIA	s	Calculate angle of distortion for each negative siphon on chart, and use maximum angle to correct all points on chart	SIPHON
		Clock running too fast - hence negative time step at change of chart	Time is correct at start of chart. The error in clock time is assumed to be constant over the day (i.e. linear)	MIA, LIA, AIA	t	Move P2 such that T2=T3 Adjust all points on chart which ended on P2 (i.e. 1 day) proportionately backwards	TIMADJ
		Negative step prior to change of chart	P4 is the 1st point of the next chart	MIA, AIA, LIA	c	Delete P3	DISCAD
		Negative step within a siphon	Siphon starts at P2 and ends at P4	MIA, AIA, LIA	c	Delete P3	DISCAD

Case	Trace	Error/ Suspected Cause	Assumption(s)	Method	Flag	Action	Routine Invoked
Negative time step with decrease in trace (continued)		Digitising error	Either P2 or P3 is incorrect	MIA	c	If I13 > I24, delete P2	FOUR-DISCAD
						If I13 < I24, delete P3	
				AIA	c	Average P2 and P3 (rain and time)	FOUR-TPRAVG
				LIA	c	If I13 > I24, delete P3	FOUR-DISCAD
						If I13 > I24, delete P2	
		P2 is the first point in the file or P1 is a missing code	P2 or P3 is incorrect	MIA	c	Move P3 such that T3=T2+1	TIMEP1
				AIA	c	Average times of P2 and P3	TIMAV
				LIA	c	Move P2 such that T2=T3-1	TIMEM1
		P2 is a missing code	Code inserted incorrectly	MIA, AIA, LIA	c	Move P2 such that T2=T3-1	MINSUB
		P3 is a missing code	Code inserted incorrectly	MIA, AIA, LIA	c	Move P3 such that T3=T2+1	MINADD

Case	Trace	Error/ Suspected Cause	Assumption(s)	Method	Flag	Action	Routine Invoked
Negative time step with decrease in trace (continued)		P3 is the last point in the data or P4 is a missing code	P2 or P3 incorrectly digitised	MIA	c	Move P2 such that T2=T3-1	TIMEM1
				AIA	c	Average times of T2 and T3	TIMEAV
				LIA	c	Move P3 such that T3=T2+1	TIMEP1
Negative time step and trace is level		Clock running too fast - hence negative time step at change of chart	Time is correct at start of chart The error in clock time is constant over the day (i.e. linear)	MIA, AIA, LIA	t	Move P2 such that T2=T3 Adjust all points on chart which ended on P2 (i.e. 1 day) proportionately backwards	TIMADJ
		Negative time step and P4 is the start of the next chart		MIA, AIA, LIA	c	Delete P3	DISCAD
		Negative time step after siphon		MIA, AIA, LIA	c	Delete P4	DISCAD

Case	Trace	Error/ Suspected Cause	Assumption(s)	Method	Flag	Action	Routine Invoked
Negative time step and trace is level (continued)		Digitising error	Either P2 or P3 is incorrect	MIA	c	If I13 > I24, delete P2	FOUR-DISCAD
						If I13 < I24, delete P3	
				AIA	c	Average P2 and P3 (rain and time)	FOUR-TPRAVG
				LIA	c	If I13 > I24, delete P3	FOUR-DISCAD
						If I13 > I24, delete P2	
Negative time step and increase in trace		Change of chart	P4 is the first point of the next chart	MIA, AIA, LIA	c	Delete P2	DISCAD
		Siphon	P4 is at the bottom of a siphon	MIA, AIA, LAI	c	Delete P2	DISCAD
		Siphon before negative time step		MIA, AIA, LIA	c	Delete P3	DISCAD

Case	Trace	Error/ Suspected Cause	Assumption(s)	Method	Flag	Action	Routine Invoked
Negative time step and increase in trace (continued)		P2 is the first point in the file or P1 is a missing code		MIA	c	Delete P3	DISCAD
				AIA	c	Average P2 and P3 (rain and time)	TPRAVG
				LIA	c	Delete P2	DISCAD
		P3 is the last point in the file or P4 is a missing code		MIA	c	Delete P2	DISCAD
				AIA	c	Average P2 and P3 (rain and time)	TPRAVG
				LIA	c	Delete P3	DISCAD
		P2 is a code		MIA, AIA, LIA	c	Move P2 such that T2=T3-1	MINSUM
		P3 is a code		MIA, AIA, LIA	c	Move P3 such that T3=T2+1	MINADD

Case	Trace	Error/ Suspected Cause	Assumption(s)	Method	Flag	Action	Routine Invoked
Negative time step and increase in trace (continued)		Digitising error	Either P2 or P3 is incorrect	MIA	c	If I13 > I24, delete P2	FOUR-DISCAD
						If I13 < I24, delete P3	
				AIA	c	Average P2 and P3 (rain and time)	FOUR-TPRAVG
				LIA	c	If I13 < I24, delete P2	FOUR-DISCAD
						If I13 > I24, delete P3	

4.1.3 Flagging of Annual Maximum Events

Two methods of flagging the events contained in the AMS extracted from the five databases were used. The first, termed "Flag_All", flags the AM event with the appropriate flag (c,s or t as defined in Section 4.1.2.3) if any data points within the duration of the AM event are flagged. This is probably too extreme, as the deletion of a single or a number of data points within the duration of an extreme event, with the remainder of the points assumed to be correct and with the siphon type of raingauge accumulating rainfall totals, has no effect on the correct duration of the event or on the total rainfall depth.

A second method was thus adopted, termed "Flag_End", which only flags the AM event if the data points spanning the start or end of the extracted annual maximum event are flagged as being corrected.

The distribution of data points marked as corrected is investigated in the following section. This is in order to ascertain whether, for example, the errors in the data occur predominantly in the larger or smaller events, or if the errors occur randomly through the range of event magnitudes.

Annual maximum events for the each duration considered are extracted from the digitised rainfall data using a moving window which has a duration equal to the duration of the event under evaluation. Each point in the break-point digitised data is considered as the potential starting point of an annual maximum event. The rainfall value at the end point of the event is interpolated linearly from the digitised data points which span the end points of the event.

4.1.4 Frequency Distribution of Corrected Annual Maximum Events

In order to ascertain the effect of the various procedures for correcting the data, an analysis was undertaken to determine whether the corrected points were creating artificially high

rainfall intensities. This was performed for events marked using the "Flag_All" and "Flag_End" methods of flagging events which had corrected data points.

4.1.4.1 "Flag_All" method

For both methods of flagging events which had corrected points, an analysis was initially performed at a single station (SAWB 0059572) and then generalised to 29 SAWB stations that had concurrent data from 1962 - 1991.

4.1.4.1.1 *Station SAWB 0059572 (East London)*

In order to assess the significance of the correction procedures, diagrams showing the frequencies for 10 equally spaced class intervals were constructed for both the entire AMS and for the events which contained a corrected point within the duration of the event. As expected, and illustrated in Figure 9, the number of events in the upper tail of the distribution which have corrected data points contained within the event, increases as the event duration increases. However, relatively few events flagged as corrected are found in the upper tail of the distribution for durations less than 30 minutes. This indicates that artificially high short duration rainfall intensities are generally not created as a result of the correction procedures.

The relative frequency distribution, computed by dividing the number of events which have corrected data points within the event, by the total number of events for each intensity class interval of events, are summarised for all durations and class intervals in Figure 10. As expected, the number of events which have flagged data points contained within the event increases with increasing duration. With some exceptions which are discussed below, relatively few events in the upper tail (intensity class > 7) of the distribution have flagged data points.

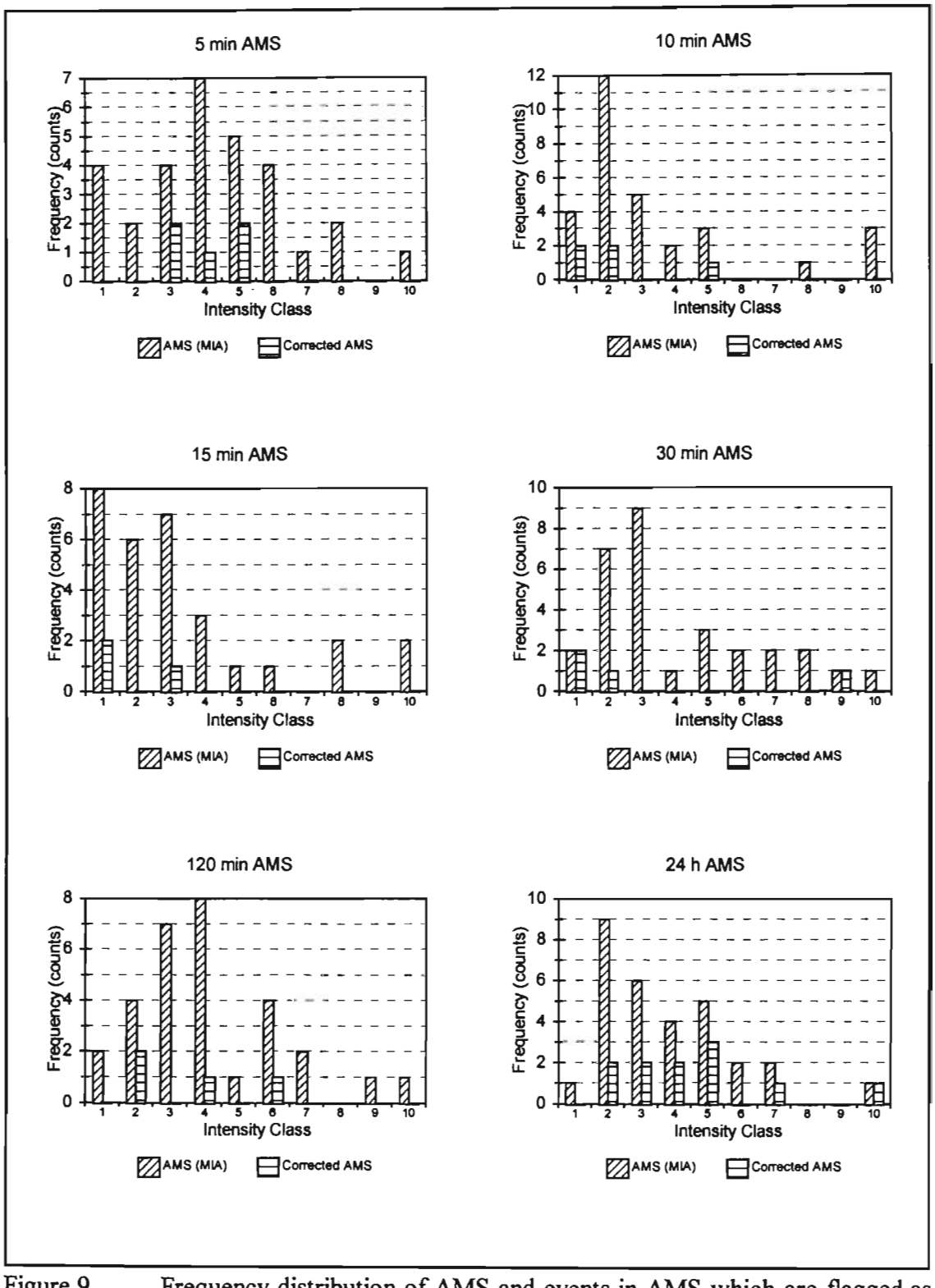


Figure 9 Frequency distribution of AMS and events in AMS which are flagged as corrected using the "Flag_All" method at Station 0059572 (East London)

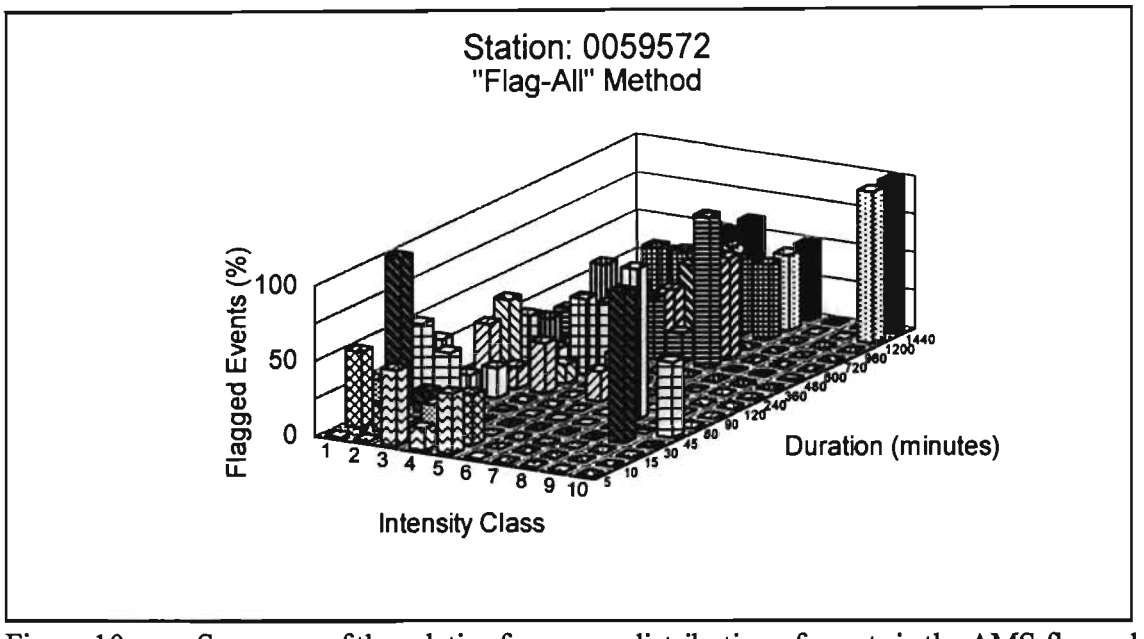


Figure 10 Summary of the relative frequency distribution of events in the AMS flagged using the "Flag_All" method at Station 0059572 (East London)

An apparent anomaly in Figure 10 is the high percentage of corrected points in frequency classes 8, 9 and 10 for durations 30, 45 and 60 minutes. This resulted from the error depicted in Table 13. This shows that an increase in rainfall from P2 to P3 occurs without an increase in time and is corrected, using the MIA method, by deleting P3 and flagging P4 as a corrected point.

Table 13 Zero time step error: SAWB Station 0059572 (East London)

Point	Date	Time	Rainfall Depth (mm*10)
P1	21/07/79	08:29	32
P2	21/07/79	08:35	45
P3	21/07/79	08:35	56
P4	21/07/79	08:37	72

As indicated in the digitised data, the Annual Maximum (AM) event for the 30 min duration started at 08:11 and hence within the 30 min period from 08:11 to 08:41, the corrected (deleted) point was encountered and thus the AM event is marked as a corrected event. The deletion of the point (in this case) has no effect on the intensity of the 30 min duration event. Similarly the AM 45 and 60 min duration events both started at 07:58 and the deleted point had no effect on the AM event, although both were marked as corrected events because the corrected point was contained within their durations.

4.1.4.1.2 *Twenty-nine SAWB stations*

The same analysis as described above was performed on all the SAWB stations that had concurrent data from 1962 - 1991 (29 stations) and the results are summarised in Figure 11.

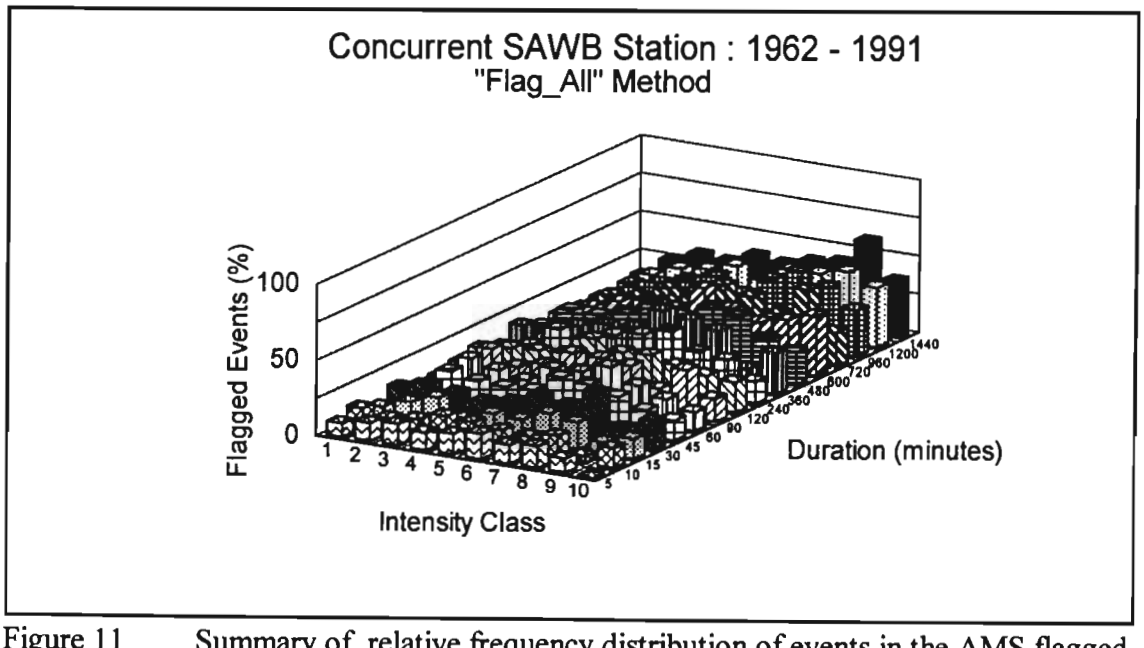


Figure 11 Summary of relative frequency distribution of events in the AMS flagged using the "Flag_All" method at 29 SAWB stations

As shown in Figure 11, when using the "Flag_All" method, relatively few events in the upper tail of the distribution of AMS have flagged points within the events when the MIA

correction method is used. Thus, even when the "Flag_All" method is used, the effect of the automated corrections on the upper tail of the distribution of the AMS is minimal.

4.1.4.2 "Flag_End" method

As discussed previously, the "Flag_All" method may flag events which have corrected data points contained within the event, but which have no effect on the rainfall intensity. Therefore the "Flag_End" method, where an event is flagged only if the corrected points span the start and end of the event, was used in an analysis of the distribution of corrected points at Station 0059572 and at the 29 SAWB stations that had concurrent data for the period 1962-1991.

4.1.4.2.1 *Station 0059572 (East London)*

The frequency distributions for Station 0059572 of both the AMS and the events in AMS flagged using the "Flag_End" method, are contained in Figure 12. The relative frequencies of the events flagged using the "Flag_End" method, expressed as a percentage of total events in each class and for each duration, are summarised in Figure 13 and indicate that the effect of the automated correction procedure on the distribution of the AMS at Station 0059572 is negligible.

4.1.4.2.2 *Twenty-nine SAWB stations*

A relative frequency analysis of the events flagged using the "Flag_End" method was performed for all 29 SAWB stations which contained concurrent data from 1962 - 1991 and the results are summarised in Figure 14.

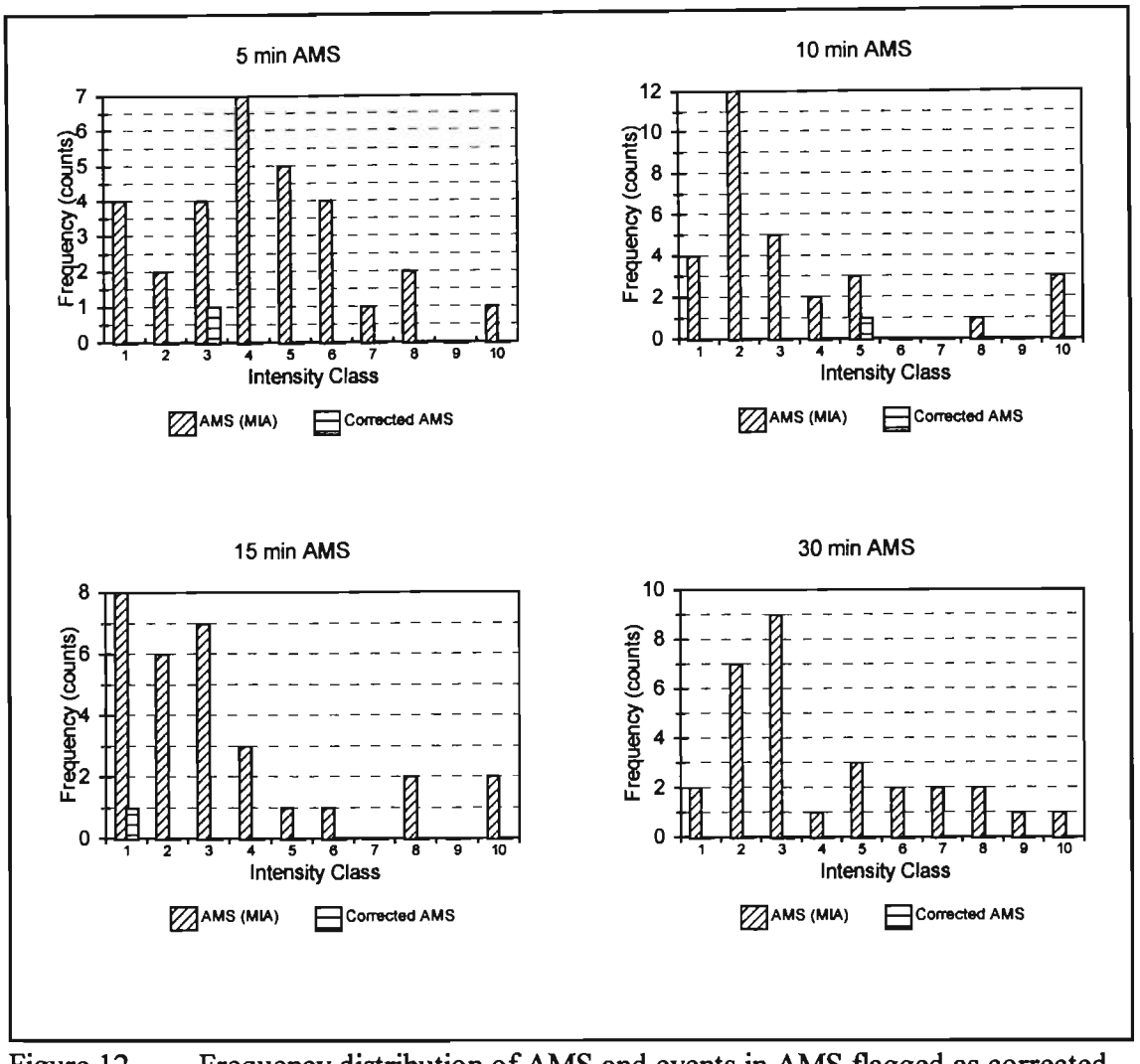


Figure 12 Frequency distribution of AMS and events in AMS flagged as corrected using the "Flag_End" method at Station 0059572 (East London)

As shown previously, the "Flag_All" method flagged events which had flagged data points within the AM event, even though they had no effect on the intensity of the event. Hence the "Flag_All" method was deemed to be inappropriate. As shown in Figure 14, the effect of the automated correction procedure on the distribution of the AMS is relatively small, with the relative frequency less than 5% for most classes and durations.

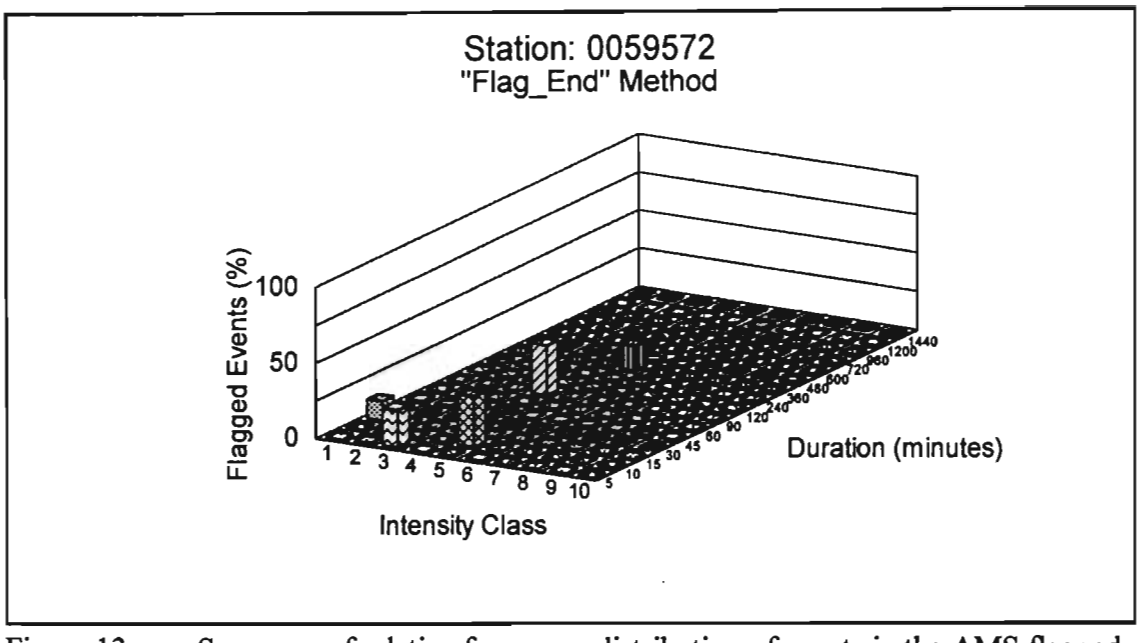


Figure 13 Summary of relative frequency distribution of events in the AMS flagged using the "Flag_End" method at Station 0059572 (East London)

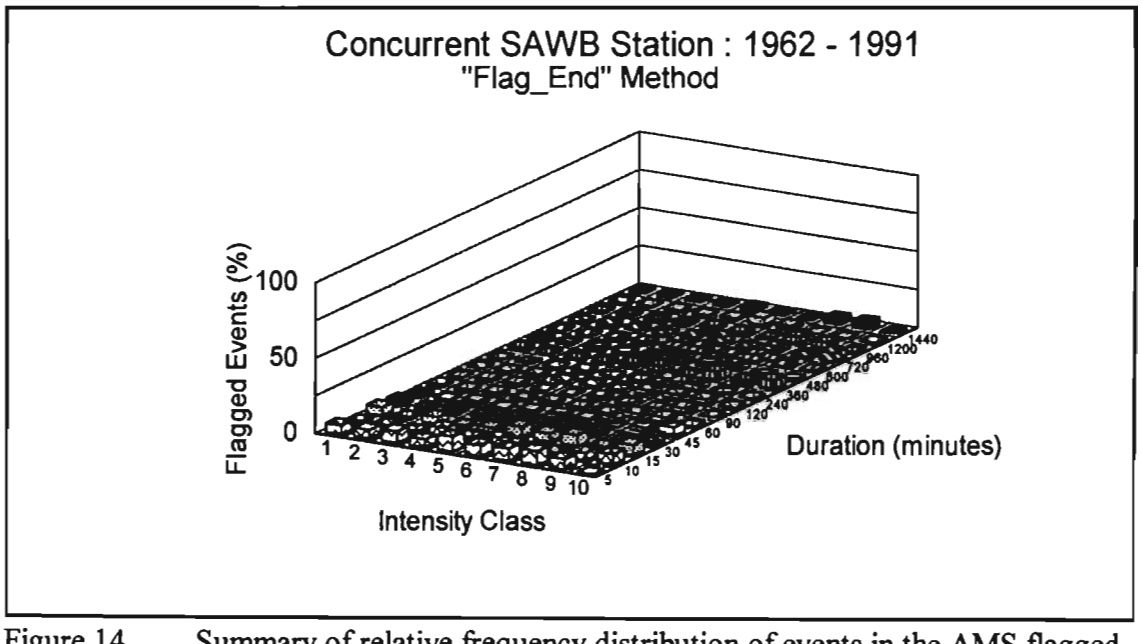


Figure 14 Summary of relative frequency distribution of events in the AMS flagged using the "Flag_End" method at 29 SAWB stations

Having shown that the effect of the correction procedures on the distribution of AMS is not significant, the differences in the various correction procedures were investigated.

4.1.5 Differences in Corrected Databases

Both parametric and non-parametric statistical tests were employed to distinguish differences between the databases corrected using the three different strategies and the database that excluded error points and error events. These were applied to data from Station 0059572 and then to data from the 29 SAWB stations that had concurrent data from 1962 - 1991.

4.1.5.1 Station 0059572 (East London)

The null hypothesis of no significant differences between the means of data groups corrected using the above procedures, was tested by performing an Analysis of Variance (ANOVA) and computing the F-test statistic. Implicit in the ANOVA test are the assumptions of normality of the data and constant variance between groups (Hirsch *et al.*, 1993). A chi-squared test, as described by Kite (1988), which utilises 10 equally spaced probability class intervals was performed on each group of data, either rejecting or accepting the null hypothesis that the data are normally distributed. The homogeneity of variances was tested by Bartlett's method, as described by Steel and Torrie (1980). Results of the normality and homogeneity of variances are contained in Tables 14 and 15. Included in Tables 14 and 15 are the results of the statistical tests performed on 30 years of consecutive data from 1962 - 1991 and on 40 years of data (i.e. all available data from Station 0059572) within the period 1940 - 1992.

As shown in Table 14, with a few exceptions, the AMS are normally distributed for most durations and correction procedure, irrespective of the length of record considered. Similarly, as shown in Table 15 and with the exception of the comparison between the MIA and EXPOINT procedures, the variances of the AMS, after correction by each of the 5 correction procedures, are relatively homogeneous. Thus with the exceptions noted, the assumptions on which the ANOVA are based are generally true and the power of the analysis is not significantly diminished.

Table 14 Acceptance (✓) and rejection (×) at the 95% confidence level of the null hypothesis of normally distributed AMS after various data correction procedures: Station 0059572 (East London)

DATABASE		EVENT DURATION (minutes)															
		5	10	15	30	45	60	90	120	240	360	480	600	720	960	1200	1440
MIA	30 years	✓	✓	×	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
MIA	40 years	✓	×	×	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
AIA	30 years	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
AIA	40 years	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
LIA	30 years	✓	✓	×	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
LIA	40 years	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
EXPOINT	30 years	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
EXPOINT	40 years	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
EXEVNT ("flag_all")	30 years	✓	×	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	×
EXEVNT ("flag_all")	40 years	✓	×	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	×	✓	✓
EXEVNT ("flag_end")	30 years	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
EXEVNT ("flag_end")	40 years	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

An ANOVA was performed for each of the 16 durations at East London between the 3 groups of AMS, which were extracted from 3 databases, each of which had been corrected using either the LIA, AIA or MIA correcting procedure. As indicated in Table 16, the null hypothesis of no significant differences of locations between the 3 data groups, was accepted at the 95% confidence level on all counts for the AMS. Thus the effect at East London of the MIA, LIA or AIA data correcting procedures, which are conceptually very different, on the AMS was negligible.

Results from similar ANOVA tests to those described above and performed on the AMS extracted from the MIA and EXPOINT databases as well as between the MIA and EXEVNT ("Flag_All" and "Flag_End" methods) databases are also contained in Table 16. Both of these tests indicated that there were significant differences, for most durations,

between the AMS extracted after the MIA, EXPOINT and EXEVNT data correcting procedures had been implemented. These results indicate that at East London either the MIA, LIA or AIA procedures are appropriate, but that the EXPOINT and EXEVNT procedures are not appropriate as they result in significantly different AMS compared to when the MIA procedure was used.

Table 15 Acceptance (✓) and rejection (X) at the 95% confidence of the null hypothesis of homogeneity of variance of the AMS after various data correction procedures: Station 0059572 (East London)

DATABASE		EVENT DURATION (minutes)															
		5	10	15	30	45	60	90	120	240	360	480	600	720	960	1200	1440
MIA AIA LIA	30 years	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
MIA AIA LIA	40 years	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
MIA EXPOINT	30 years	✓	X	X	X	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
MIA EXPOINT	40 years	✓	X	X	X	X	X	X	X	X	✓	✓	✓	✓	✓	✓	✓
MIA EXEVNT ("flag_all")	30 years	✓	X	X	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	X	X
MIA EXEVNT ("flag_all")	40 years	✓	X	X	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	X	X
MIA EXEVNT ("flag_end")	30 years	✓	X	X	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
MIA EXEVNT ("flag_end")	40 years	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Table 16 Acceptance (✓) or rejection (×) at the 95% confidence of the null hypothesis of no significant differences between data groups after correction by various procedures: Station 0059572 (East London)

DATABASE		EVENT DURATION (minutes)															
		5	10	15	30	45	60	90	120	240	360	480	600	720	960	1200	1440
MIA AIA LIA	30 years	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
MIA AIA LIA	40 years	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
MIA EXPOINT	30 years	×	×	×	×	×	×	×	×	✓	✓	✓	✓	✓	✓	✓	✓
MIA EXPOINT	40 years	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×
MIA EXEVNT ("flag_all")	30 years	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×
MIA EXEVNT ("flag_all")	40 years	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×
MIA EXEVNT ("flag_end")	30 years	×	×	×	×	×	×	×	×	×	×	✓	✓	×	✓	✓	✓
MIA EXEVNT ("flag_end")	40 years	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×

As shown in Table 16, the MIA, AIA and LIA AMS are not significantly different at the 95% confidence level. However, significant differences at the 95% confidence level between the MIA and both the EXEVNT ("Flag_All"and "Flag_End" method) and EXPOINT AMS are evident for most durations. Similar results were presented by Smithers (1993), who had however excluded both the adjusted and corrected events, and not just the corrected events, as is the case for the results in Table 16.

According to Hirsch *et al.* (1993) and as shown in Tables 14 and 15, the violation of the assumptions of normality or of constant variance, results in loss of power of the ANOVA test. The results from applying the non-parametric Kruskal-Wallis test to the AMS are contained in Table 17. While some differences are noted between the Kruskal-Wallis test and the ANOVA, the trends are similar, thus giving greater confidence to the results of the statistical tests.

Table 17 Acceptance (✓) or rejection (✗) at the 95% confidence of the null hypothesis of identical distributions between data groups after correction by various procedures (Kruskal-Wallis test): Station 0059572 (East London)

DATABASE		EVENT DURATION (minutes)															
		5	10	15	30	45	60	90	120	240	360	480	600	720	960	1200	1440
MIA AIA LIA	30 years	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
MIA AIA LIA	40 years	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
MIA EXPOINT	30 years	✗	✗	✗	✗	✗	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
MIA EXPOINT	40 years	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓	✓
MIA EXEVNT ("flag_all")	30 years	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
MIA EXEVNT ("flag_all")	40 years	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
MIA EXEVNT ("flag_end")	30 years	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓	✓	✗	✓	✓	✓
MIA EXEVNT ("flag_end")	40 years	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗

With some exceptions, the length of record generally has no effect on the significance of the results. No significant differences were found between AMS extracted from databases corrected using the MIA, AIA or LIA methods. However, the AMS extracted from a database corrected using the EXPOINT method, as well when events were excluded (EXEVNT) which have corrected data points, either within the event or at the extremities of the event, were significantly different to other correction procedures.

A similar analysis to the above was performed at 29 SAWB stations which have concurrent data for the period 1962 - 1991 and the results are reported in the following section.

4.1.5.2 Twenty-nine SAWB stations

The results of normality tests for all 29 SAWB stations that have concurrent data from 1962 - 1991 are contained in Table 18.

Table 18 Number of stations where the null hypothesis of normally distributed data was rejected at the 95% confidence level, expressed as a percentage of total number of stations tested (29)

DATABASE		EVENT DURATION (minutes)															
		5	10	15	30	45	60	90	120	240	360	480	600	720	960	1200	1440
MIA	AMS	41	24	24	17	14	17	21	17	10	10	21	10	10	17	14	21
AIA	AMS	38	28	17	10	14	14	17	17	10	10	24	10	7	17	10	14
LIA	AMS	45	24	14	14	17	21	17	21	14	10	21	7	10	21	17	14
EXPOINT	AMS	48	31	24	10	7	14	14	28	3	10	10	14	7	10	10	14
EXEVNT	AMS	38	24	14	10	21	10	24	21	10	14	10	17	21	21	17	10

The results of homogeneity of variance tests for all 29 SAWB stations that have concurrent data from 1962 - 1991 are contained in Table 19.

Table 19 Number of stations where the null hypothesis of homogeneity of variance was rejected at the 95% confidence level, expressed as a percentage of total number of stations tested (29)

DATABASE		EVENT DURATION (minutes)															
		5	10	15	30	45	60	90	120	240	360	480	600	720	960	1200	1440
MIA AIA LIA	AMS	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
MIA EXEVNT	AMS	7	17	14	10	14	14	7	7	10	10	17	10	14	10	14	3
MIA EXPOINT	AMS	17	45	38	38	41	34	48	38	28	24	21	17	24	14	14	14

Results from similar ANOVA tests to those described above and performed on the AMS generated from the MIA and EXPOINT databases as well as between the MIA and EXEVNT databases are also contained in Table 20.

Table 20 Number of stations where the null hypothesis of no significant differences between data groups was rejected at the 95% confidence level, expressed as a percentage of total number of stations tested (29)

DATABASE		EVENT DURATION (minutes)															
		5	10	15	30	45	60	90	120	240	360	480	600	720	960	1200	1440
MIA AIA LIA	AMS	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
MIA EXPOINT	AMS	52	62	59	59	55	52	55	55	41	31	28	24	21	21	17	10
MIA EXEVNT	AMS	72	79	83	79	72	72	66	69	62	62	52	45	48	38	14	17

The results from applying the non-parametric Kruskal-Wallis test to the AMS are contained in Table 21.

Table 21 Number of stations where the null hypothesis of identical distributions between data groups (Kruskal-Wallis test) was rejected at the 95% confidence level, expressed as a percentage of total number of stations tested (29)

DATABASE		EVENT DURATION (minutes)															
		5	10	15	30	45	60	90	120	240	360	480	600	720	960	1200	1440
MIA AIA LIA	AMS	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
MIA EXPOINT	AMS	31	48	55	62	59	55	34	38	31	24	21	24	17	10	7	7
MIA EXEVNT	AMS	59	69	72	72	76	72	66	72	62	45	31	31	21	17	3	0

4.1.5.3 Concluding remarks on differences in corrected databases

In the case of Station 0059572 and for all 29 SAWB stations that had concurrent data from 1962-1991, no significant differences were found between the means and variances of the AMS extracted from the MIA, AIA and LIA databases. Significant differences were found between the AMS extracted from the MIA and both the EXPOINT and EXEVNT databases. The correction approaches used in the MIA, AIA and LIA procedures are different, yet do not produce significantly different AMS, thus indicating that the procedure chosen to correct the database is not critical. The exclusion of all erroneous data points (EXPOINT), or events flagged according to both the "Flag_All" and "Flag_End" methods (EXEVNT), does significantly affect the AMS. Thus it is hypothesised that the MIA correction procedure, or a random selection of the LIA, AIA or MIA procedure, should be

adopted. The effect of randomly selecting either of the MIA, AIA or LIA procedures is investigated in the following section.

4.1.6 Correction by Random Selection of MIA, LIA Or AIA Procedures

When the probable cause of an error in the data is unknown, an option (RANDOM) was developed to randomly invoke the MIA, AIA or LIA procedures, in addition to the options to correct the data using only one of the procedures. It was assumed that the random selection of the correcting procedure would better reflect the nature of the errors.

In order to evaluate the RANDOM procedure, errors were randomly introduced into error-free (clean) data and the RANDOM procedure was used to correct the errors. The correction procedure was then evaluated by comparing the AMS extracted from the error-free data and from the data after the randomly introduced errors had been corrected using the RANDOM procedure.

4.1.6.1 Creating errors in the data for hypothesis testing

Four types of errors were introduced randomly into the data by selecting a line number, in the data file, at random and reading sequentially from that point in the file until the first appropriate point (e.g. siphon) which had not previously been altered. The types of errors introduced are:

- negative time step (not at change of chart or siphon),
- negative time step during siphon,
- negative time step at change of chart, and
- zero time step (infinite intensity).

The four types of errors were introduced randomly, with the seed value for the selection of the random number based on the system clock time. The parameters used by the routine and which are set by the user are:

- number of errors to introduce per year,
- maximum negative time step, and
- maximum number of negative time steps.

Based on records from 29 SAWB autographic rainfall stations which had 30 years of concurrent data, the average number of errors per year was estimated to be 30. Hence the number of errors introduced into the data was set at 30 per year. The maximum negative time step was set to 60 minutes. Thus, when negative time step errors were introduced into the data, a random value between 0 and 60 was used. The maximum number of data points that were moved when adding negative time steps was 2. Hence, either 1 or 2 data points were moved to create the maximum negative time step. A typical sequence of errors introduced into the data is shown in Table 22.

Table 22 Example of errors introduced randomly during a single sequence: Station 0059572 (East London)

Type of Error	Date (dd/mm/yy)	Time	Number of negative time steps	Size of Negative Time Step (minutes)
Zero time step	08/07/40	02:16		
Negative step at chart change	28/10/40	08:27		32
Negative step at chart change	21/12/40	08:27		24
Negative step	12/09/40	01:54	1	37
Negative step at chart change	01/11/40	08:29		11
Negative step	10/08/40	08:22	2	59 51
Negative step	29/02/40	23:58	1	43
Negative step at chart change	11/11/40	08:28		32
Negative step	10/08/40	10:38	2	55 22

SAWB Station 0059572 (East London) was used in a case study to evaluate the RANDOM procedure and results of the evaluation are presented in the following section.

4.1.6.2 Evaluation of RANDOM procedure at Station 0059572 (East London)

The digitised rainfall data from Station 0059572 were corrected and the corrected data used as a control. Errors were randomly inserted into the control (error-free) data and then corrected using the RANDOM procedure, after which the AMS for durations ranging from 5 min to 24 h were extracted. This process was initially repeated 10 times and subsequently 100, times resulting in 11 (control and 10 corrections) and 101 (control and 100 corrections) sets of AMS respectively. The time used on the CCWR's mainframe computer to complete the 100 repetitions of this procedure was approximately 10 days and hence only one case study was performed. The results for only the 100 repetitions are reported.

The null hypothesis of no significant differences existing between the means of the control and 100 repetitions, was tested by performing an Analysis of Variance (ANOVA) and computing the F-test statistic. Implicit in the ANOVA test are the assumptions of normality of the data and constant variance between groups. The homogeneity of variances was tested by Bartlett's method, as described by Steel and Torrie (1980). Results of the normality and homogeneity of variances tests are contained in Tables 23 and 24 and, with the exceptions for durations $\leq$ 10 min, the power of the ANOVA test is not significantly diminished as a result of significant deviations from underlying assumptions.

An ANOVA was performed for each of the 16 durations at Station 0059572 between the control and 100 replications. As indicated in Table 25, the null hypothesis of no significant differences of locations between the 101 sets of AMS, was accepted at the 95% confidence level on all counts for the AMS.

Table 23 Number of times the null hypothesis of normally distributed AMS of the control (no errors) and of 100 corrected series of data using the RANDOM procedure was accepted or rejected at the 95% confidence level : Station 0059572 (East London)

DATABASE		EVENT DURATION (minutes)															
		5	10	15	30	45	60	90	120	240	360	480	600	720	960	1200	1440
Control (No errors)	Accept	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	Reject	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
RANDOM Corrections	Accept	87	15	98	99	1000	95	1000	1000	1000	88	100	99	94	100	96	94
	Reject	13	85	2	1	0	5	0	0	0	12	0	1	6	0	4	6

Table 24 Acceptance (✓) and rejection (X) at the 95% confidence level of the null hypothesis of homogeneity of variance between AMS extracted from 100 corrections using the RANDOM procedure and AMS of control data : Station 0059572 (East London)

DATABASE		EVENT DURATION (minutes)															
		5	10	15	30	45	60	90	120	240	360	480	600	720	960	1200	1440
RANDOM	48 years	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

In addition to the ANOVA test, a Kruskal-Wallis test was also performed on the null hypothesis of no significant differences of locations between the 101 sets of AMS. The results of this analysis are contained in Table 26.

Table 25 Acceptance (✓) or rejection (✗) at the 95% confidence level of the null hypothesis of no significant differences between AMS extracted from the control and from 100 corrections to the data using the RANDOM procedure after errors had been randomly introduced into the control data: Station 0059572 (East London)

DATABASE	EVENT DURATION (minutes)															
	5	10	15	30	45	60	90	120	240	360	480	600	720	960	1200	1440
Control and 100 RANDOM corrections	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

From the above case study at Station 0059572 and for 100 repetitions of errors introduced randomly into the control (error-free) data and corrected using the RANDOM procedure, it appears that the use of the RANDOM correction procedure has no significant effect on the AMS. Similar results were obtained from 10 repetitions. Processing (CPU) time limited the study to only 10 and 100 repetitions at a single site. It is thus postulated that the RANDOM procedure (i.e. a random selection of the MIA, AIA or LIA procedures) to correct the data better reflects the probable random nature of the causes of the errors in the data than do the independent use the MIA, AIA or LIA procedures. Hence the RANDOM procedure was adopted to correct errors in the data.

Table 26 Acceptance (✓) or rejection (X) at the 95% confidence level of the null hypothesis of no significant differences between AMS extracted from the control and from 100 corrections to the data using the RANDOM procedure after errors had been randomly introduced into the control data (Kruskal-Wallis test): Station 0059572 (East London)

DATABASE	EVENT DURATION (minutes)															
	5	10	15	30	45	60	90	120	240	360	480	600	720	960	1200	1440
Control and 100 RANDOM corrections	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

In the following section, the first of the consistency checks on the digitised data is presented where the digitised and manually extracted extreme events are compared.

4.2 COMPARISON OF DIGITISED AND MANUALLY EXTRACTED ANNUAL MAXIMUM SERIES

At selected study sites the values of the AMS extracted from the digitised database, corrected using the MIA procedure, were compared to those reported by Midgley and Pitman (1978), which had been extracted manually from autographic charts. Where differences in the AMS were noted, and where available, comparisons were made between the digitised data, rainfall charts and the manually extracted hourly totals. As noted by *inter alia* Schulze (1984) and Weddepohl (1988) it is expected that the AMS extracted from the digitised data should be greater than the AMS extracted manually from autographic charts, as the manual extraction used fixed 15 min time increments and hence the recorded maxima could have been missed, particularly for shorter durations.

4.2.1 Station 0059572 (East London)

The 15, 30, 45, 60 and 1440 min duration AMS extracted manually from charts and automatically from the digitised data for SAWB Station 0059572 are plotted in Figure 15. Included in Figure 15, and plotted using the right hand side (Y2) scale, is the ratio between the digitised and manually extracted value, expressed as a percentage. As noted above, this percentage is expected to be ≥ 100. However, as shown in Figure 15 the percentage is seldom ≥ 100, particularly for durations less than 1 h. Assuming that the manually extracted data are correct, it is thus evident that a number of extreme events were not adequately digitised. Selected anomalies are discussed below.

As depicted in Figure 15, the manually extracted AMS exceeded the digitised AMS for all selected durations in 1958. The manually extracted hourly totals indicate that, for the all selected durations, the AMS events in 1958 occurred between 08:00 on 21 December and 08:00 on 22 December. The chart for this day appears not to have been digitised as it is not contained in the SAWB digitised database, which does contain data for 20 and 22 December, but not for 21 December 1958.

The AM event during 1967 occurred on 26 May 1967 for all durations. For durations up to 60 min, the manually extracted data exceeds the digitised data, and for the maximum 24 h event, the digitised AM event is larger. The digitised data indicate that data are missing on 26 May from 18:54 to 19:37, which may explain the large differences for durations up to 60 min. A copy of the chart for 26 May 1967 may explain the reason for the missing data and why the manually extracted data exceed the digitised data.

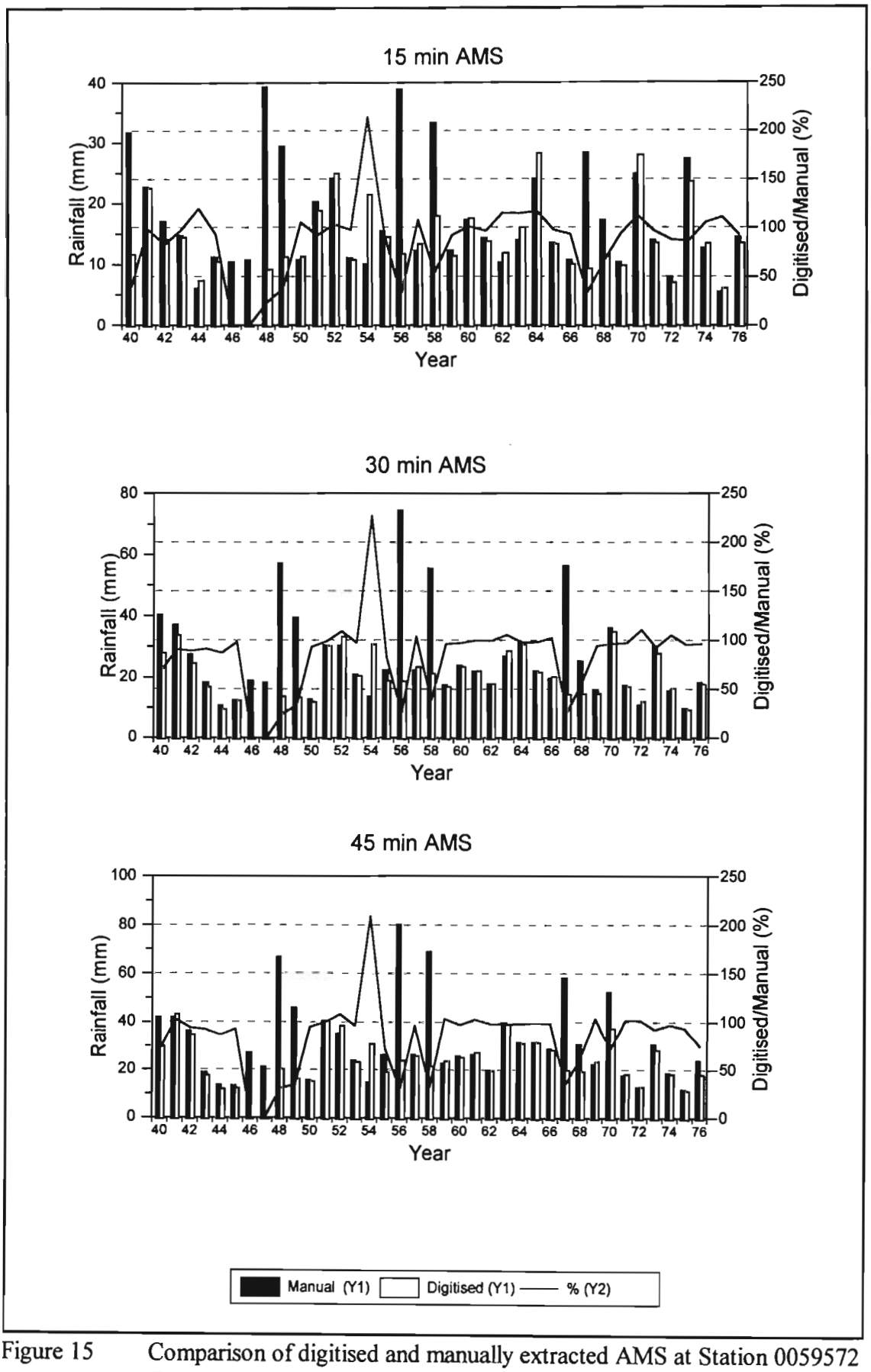


Figure 15 Comparison of digitised and manually extracted AMS at Station 0059572 (East London)

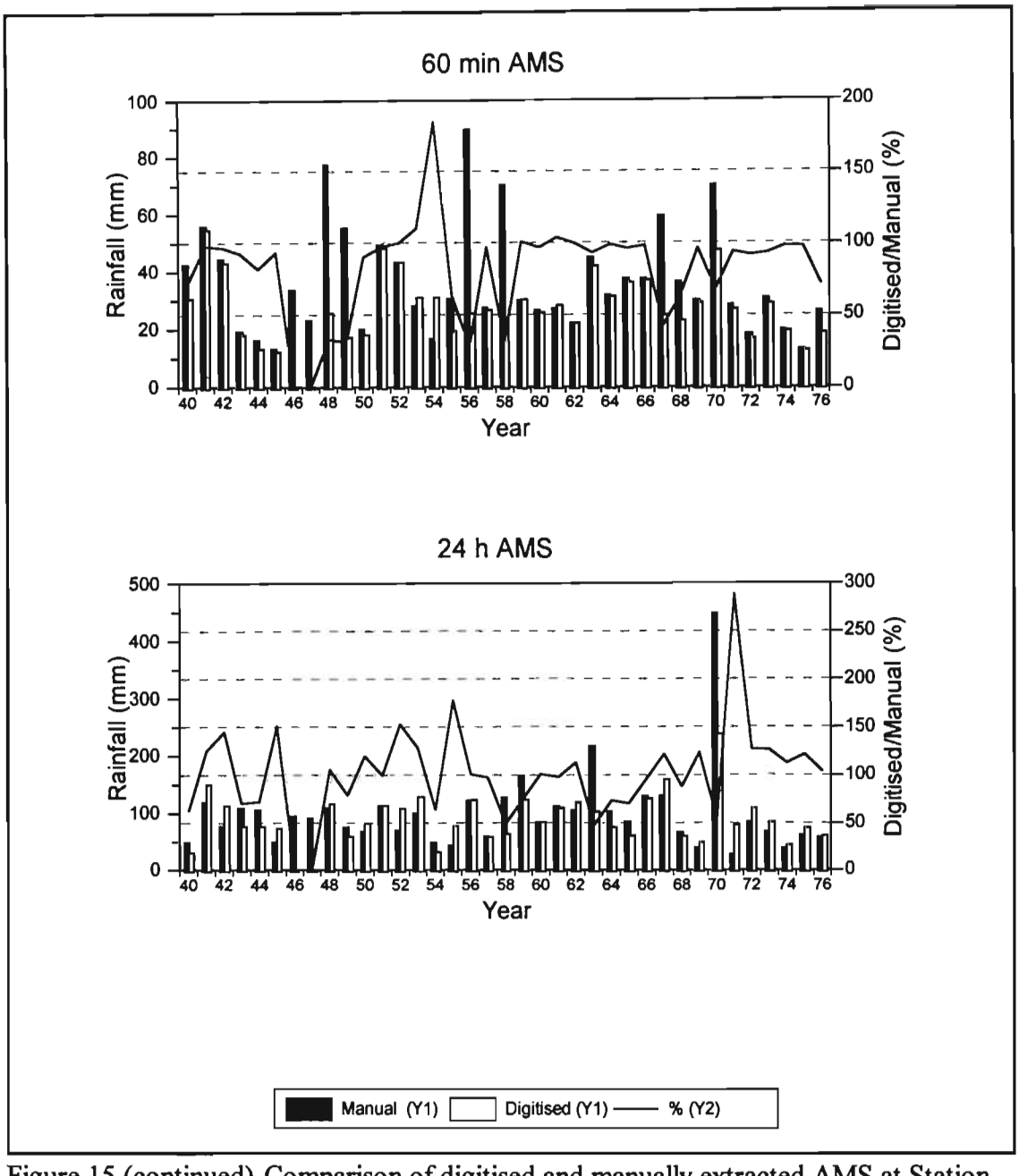


Figure 15 (continued) Comparison of digitised and manually extracted AMS at Station 0059572 (East London)

The maximum 24 h digitised rainfall event during 1970 starts at 02:24 on 27 August 1970 and 237 mm of rainfall is recorded . The manually extracted AM 24 h total for the 24 h period starting at 08:00 on 27 August 1970 is 447 mm. The digitised data are missing for the period 07:34 to 13:34 on the 27 August. The rainfall from the manually extracted hourly data for the period 08:00 to 14:00 is 211 mm. At least 190 mm of rainfall recorded on the

chart during the period of missing data can be reasonably deduced from the indistinct trace. Thus the period of missing digitised data accounts for the difference between the 24 h digitised and manually extracted totals. The probable reason for the entire chart not being digitised is that the ink had run dry and the trace is not that clear.

4.2.2 Station 0317476 (Upington)

Similar to the analyses above, the ratio between the manually extracted and digitised annual maximum event for SAWB Station 0317467 is shown in Figure 16. Generally the digitised AMS exceed the manually extracted AMS, although in some years and for some durations the digitised values may be as little as 60% of the manually extracted value. An anomaly in the manually extracted data is apparent for 1966 where the 15, 30, 45 and 60 minute annual maximum rainfalls are all 7.3 mm and the 24 h rainfall is 9.5 mm, which results in the digitised/manual ratio of 3.94 for this year. Years in which the digitised value is less than the manually extracted value (e.g. 1960) are postulated to be the result of portions of the autographic rainfall charts not being digitised.

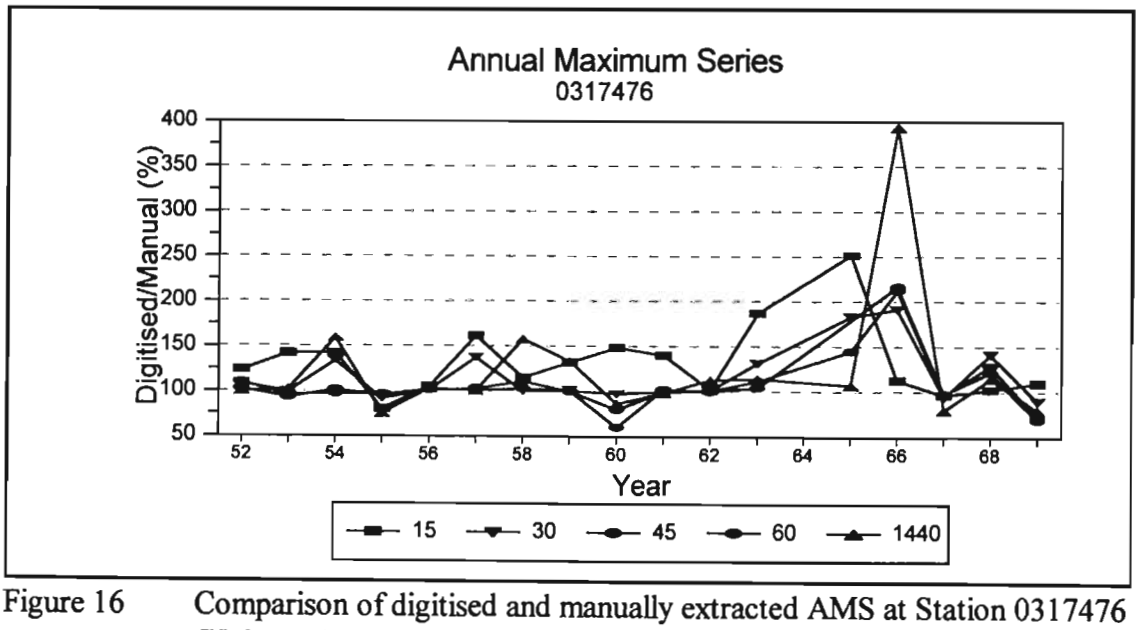


Figure 16 Comparison of digitised and manually extracted AMS at Station 0317476 (Upington)

4.2.3 Station 0677802 (Pietersburg)

The ratio between the manually extracted and digitised AMS for SAWB Station 0677802 is shown in Figure 17. Generally the AMS extracted from the digitised exceeds the manually extracted values, although on occasion the reverse trend occurs. Similar to Station 0317476, the large differences between the two series, particularly for the 24 h duration event, is unexpected, but could be explained by errors occurring during the manual extraction or digitisation of rainfall events.

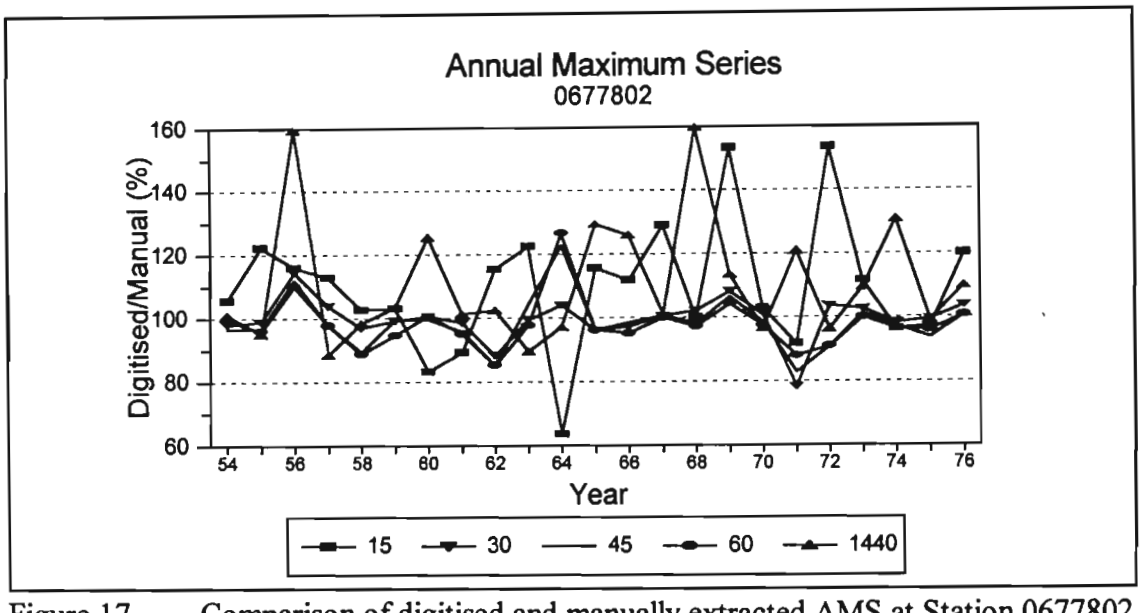


Figure 17 Comparison of digitised and manually extracted AMS at Station 0677802 (Pietersburg)

The three examples presented above illustrate the relatively large differences that do occur between the digitised and manually extracted AMS and that different trends do occur at particular stations. For example, the manually extracted AMS generally exceed the digitised AMS at Station 0059572 while the reverse is generally true at Stations 0317476 and 0677802. Another method of assessing the adequacy of the digitised data is to compare the daily rainfall totals computed from the digitised data to data recorded by the adjacent non-recording daily rainfall raingauge. This is again illustrated by means of selected examples.

4.3 COMPARISON OF DIGITISED AND STANDARD RAINGAUGE DAILY TOTALS

In order assess the reliability of the digitised data and to identify where extreme events were not contained in the digitised rainfall database, a comparison was performed between three sources of data for obtaining totals of daily rainfall:

- Daily rainfall totals derived from the digitised data for fixed 24 h periods ending at 08:00 every day are referred to as *Digitised.*
- Adjacent to each recording raingauge is a standard, non-recording raingauge measure at 24 h intervals at 08:00 every day, and this source of daily rainfall totals is referred to as *SAWB Daily.*
- The daily rainfall total as measured by the adjacent standard, non-recording raingauge is included within the digitised data file obtained from the SAWB, as a control for the days digitised rainfall data, and this daily rainfall total obtained from the digitised rainfall file is referred to as *SAWB Control.* Hence the SAWB Control and SAWB Daily values should be the same as they are recorded by the same raingauge.

The SAWB Daily values were extracted from the SAWB daily rainfall database housed by the Computing Centre for Water Research (CCWR) and, of the three sources of daily rainfall data, were assumed to be the most reliable. This assumption is based on the frequent use of the SAWB daily rainfall database, and hence errors are noted by users. In comparison, this study is the first major user of the digitised database and hence little feedback has been given to the SAWB regarding the quality of the digitised rainfall data. In addition, the processing of the digitised data and the inherent greater potential for problems when recording rainfall continuously and autographically, and the more thorough checking of the daily rainfall data by the SAWB, add credibility to this assumption. The comparisons of daily rainfall totals from these three sources were performed for selected stations.

4.3.1 Station 0034767 (Uitenhage)

Cumulative totals of daily rainfall for the Digitised, SAWB Control and SAWB Daily values, as well as a scatter plot of Digitised vs SAWB Daily values are shown in Figure 18. For Station 0034767, a good comparison is evident between the SAWB Control and SAWB Daily values, but the Digitised total is often less than the SAWB Daily value. A comparison of the daily totals obtained from the three sources for the thirty largest daily rainfall totals during the period January 1954 - December 1975 is listed in Table 27. From Table 27 it is evident, that on numerous days when the Digitised total is substantially less than the SAWB Daily value, no missing data are recorded in the digitised data. Thus, regrettably the missing data flags in the digitised SAWB data are not a reliable indicator of whether data are missing or not.

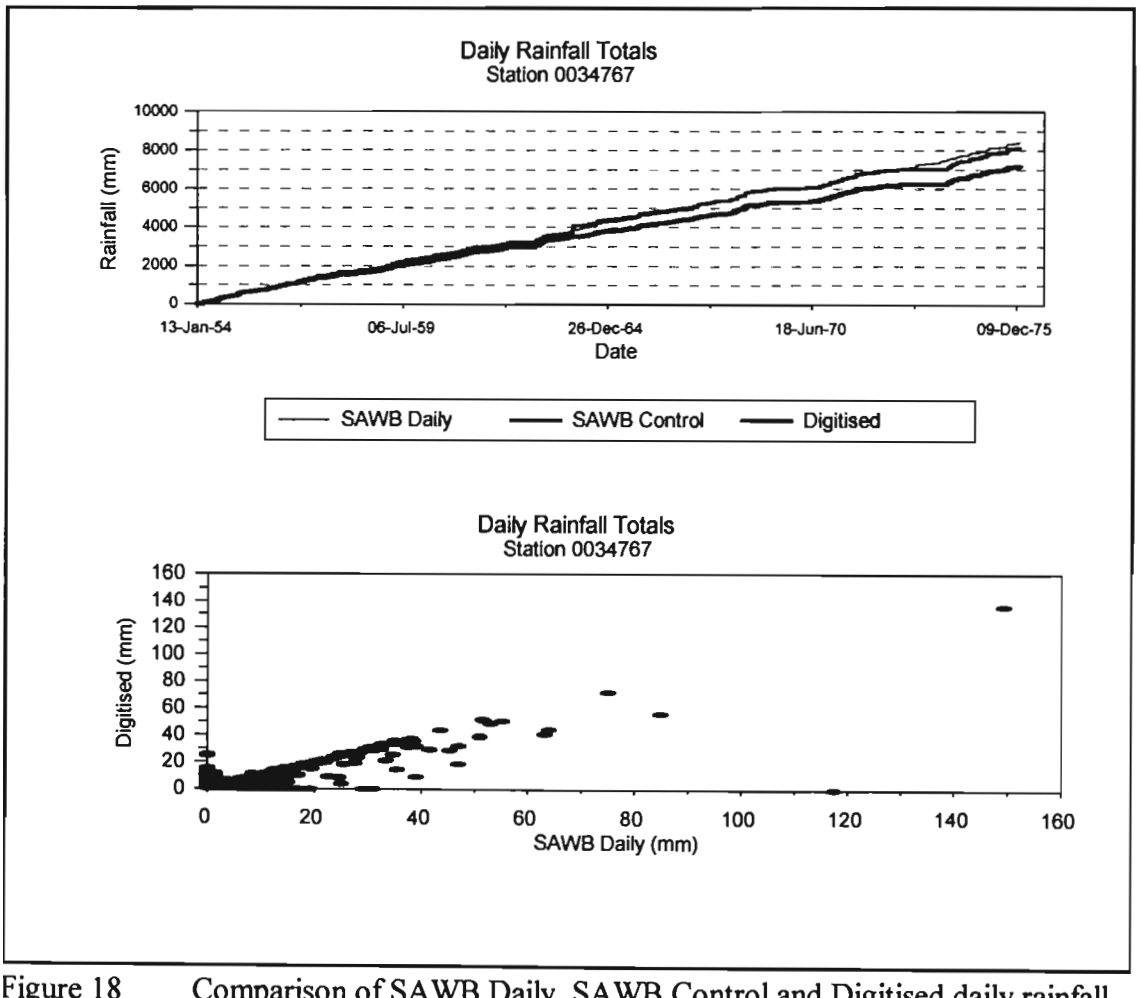


Figure 18 Comparison of SAWB Daily, SAWB Control and Digitised daily rainfall totals at Station 0034767 (Uitenhage)

Table 27 Comparison of daily rainfall totals obtained from three sources for the thirty largest events for period 1954 - 1975 at Station 0034767 (Uitenhage)

Year	Month	Day	Daily Rainfall Total			Digitised Flag (M=Missing)	Digitised / SAWB Daily (ratio)
			SAWB Daily (mm)	SAWB Control (mm)	Digitised (mm)		
68	9	1	149.2	149.2	136.0		0.91
67	4	9	117.4	0.0	0.0		0.00
71	8	21	84.9	84.9	55.8		0.66
54	8	26	75.0	75.0	72.0	M	0.96
67	5	26	64.0	64.0	44.6		0.70
64	9	16	63.2	63.2	40.9		0.65
70	12	6	55.2	55.2	50.7		0.92
55	11	29	53.0	53.0	48.6	M	0.92
75	2	10	52.0	52.0	50.8		0.98
63	3	7	51.5	51.5	51.5		1.00
74	8	22	51.0	51.0	38.9		0.76
68	6	12	47.0	47.0	32.2		0.69
59	8	2	47.0	47.2	18.8	M	0.40
75	8	31	45.4	45.4	28.7		0.63
74	1	26	43.6	43.6	43.4	M	1.00
56	12	20	41.5	41.5	29.3	M	0.71
57	6	30	39.0	39.0	9.3	M	0.24
74	9	2	39.0	39.0	31.3		0.80
66	11	4	38.5	39.7	35.7		0.93
63	1	23	38.0	38.0	37.4		0.98
56	9	18	37.5	37.5	31.2		0.83
62	3	10	37.5	37.5	36.8		0.98
65	11	3	36.5	36.5	34.9		0.96
67	4	8	35.8	35.8	33.6		0.94
65	11	2	35.5	35.5	35.0		0.99
74	5	2	35.4	35.4	14.4	M	0.41
59	7	17	35.0	35.0	35.8		1.02
59	1	25	34.6	34.6	25.2		0.73

4.3.2 Station 0035179 (Port Elizabeth)

The accumulated daily rainfall totals obtained from the standard gauge value (SAWB Control) in the file containing the digitised data and from the SAWB daily rainfall database, as well as the total derived from the digitised rainfall data for Station 0035179 are shown in Figure 19. A comparison of the daily totals obtained from the three sources for the thirty largest daily rainfall totals during the period January 1938 - December 1975 are listed in Table 28. From Table 28 it is evident, that on numerous days when the Digitised total is substantially less than the SAWB Daily value, no missing data are recorded in the digitised data.

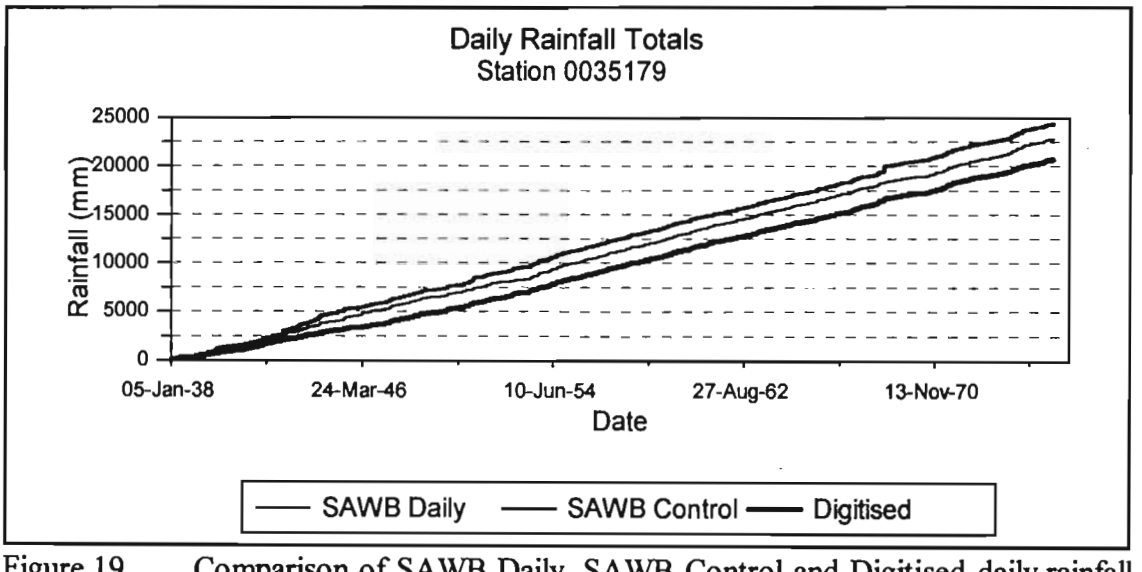


Figure 19 Comparison of SAWB Daily, SAWB Control and Digitised daily rainfall totals at Station 0035179 (Port Elizabeth)

4.3.3 Station 0059572 (East London)

The accumulative daily rainfall totals obtained from the standard gauge value in the file containing the digitised data and from the SAWB daily rainfall database (obtained from CCWR), as well as the total derived from the digitised rainfall data for Station 0059572 are shown in Figure 20. A comparison of the daily totals obtained from the three sources for

the thirty largest daily rainfall totals during the period January 1938 - December 1975 are listed in Table 29. Clearly the SAWB Daily data for Station 0059572 extracted from the database housed on the CCWR are missing from 1973 to 1987.

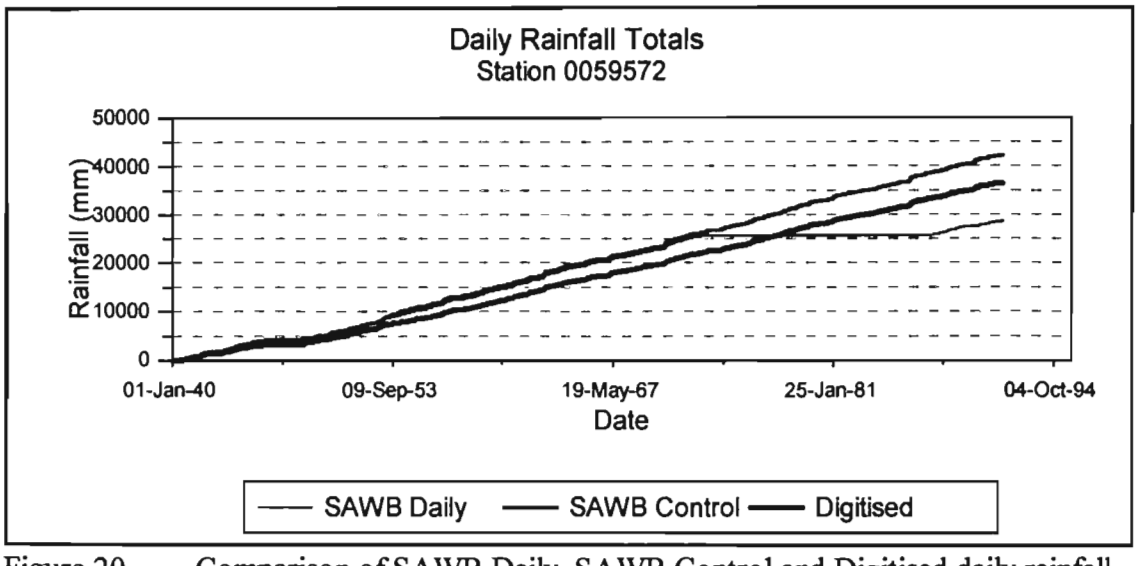


Figure 20 Comparison of SAWB Daily, SAWB Control and Digitised daily rainfall totals at Station 0059572 (East London)

4.3.4 Station 0088293 (Sutherland)

The accumulative daily rainfall totals obtained from the standard gauge value (SAWB Control) in the file containing the digitised data and from the SAWB daily rainfall database (obtained from the CCWR), as well as the total derived from the digitised rainfall data for Station 0088293 are shown in Figure 21. A comparison of the daily totals obtained from the three sources for the thirty largest daily rainfall totals during the period January 1961 - July 1991 are listed in Table 30.

Table 28 Comparison of daily rainfall totals obtained from three sources for the thirty largest events for period 1938 - 1975 at Station 0035179 (Port Elizabeth)

Year	Month	Day	Daily Rainfall Total			Digitised Flag (M=Missing)	Digitised / SAWB Daily (ratio)
			SAWB Daily (mm)	SAWB Control (mm)	Digitised (mm)		
54	8	26	132.5	132.5	111.4	M	0.84
61	2	11	120.6	12.0	118.2		0.98
46	3	22	108.4	0.0	0.0		0.00
62	4	26	105.4	105.4	62.5		0.59
41	12	21	100.0	100.0	88.1		0.88
70	12	6	94.6	94.6	86.1	M	0.91
49	11	16	91.9	91.9	83.9		0.91
53	10	20	91.0	91.0	20.1	M	0.22
51	1	11	88.1	88.1	31.2	M	0.35
67	4	9	88.0	88.0	88.1		1.00
67	5	26	76.0	76.0	77.4		1.02
39	12	3	72.8	72.8	0.0	M	0.00
64	9	16	72.4	72.4	70.5		0.97
49	11	17	72.3	72.3	68.0		0.94
54	5	20	72.0	72.0	63.4		0.88
41	6	27	70.3	70.3	65.3		0.93
43	9	14	68.5	6.8	7.6		0.11
53	6	22	64.8	64.8	65.5		1.01
74	3	3	64.3	64.3	19.1	M	0.30
39	7	6	61.9	0.0	0.0		0.00
74	6	14	59.7	59.7	24.4	M	0.41
53	6	21	59.5	59.5	57.3		0.96
68	6	1	59.2	59.2	59.1		1.00
74	8	22	59.0	59.0	44.1		0.75
45	6	24	58.9	58.9	18.7	M	0.32
60	5	6	56.8	56.8	55.8		0.98
53	7	28	56.2	56.2	20.1	M	0.36
72	5	11	55.4	55.4	44.8		0.81

Table 29 Comparison of daily rainfall totals obtained from three sources for the thirty largest events for period 1940 - 1991 at Station 0059572 (East London)

Year	Month	Day	Daily Rainfall Total			Digitised Flag (M=Missing)	Digitised/ SAWB Daily (ratio)
			SAWB Daily (mm)	SAWB Control (mm)	Digitised (mm)		
63	3	7	199.7	217.2	100.5		0.50
70	8	25	155.3	115.3	129.1		0.83
70	8	28	152.4	0.0	0.0		0.00
70	8	27	147.0	447.0	180.8	M	1.23
67	4	10	130.6	130.6	130.9		1.00
58	12	21	127.5	0.0	0.0		0.00
59	7	18	122.4	122.4	82.0		0.67
56	2	15	122.3	122.3	47.2	M	0.39
56	11	1	122.1	122.1	114.8		0.94
41	4	5	119.3	11.9	83.6	M	0.70
51	3	27	116.3	103.6	102.4		0.88
51	9	4	113.0	113.0	59.6		0.53
44	3	9	112.2	112.2	73.9	M	0.66
61	7	30	112.1	112.1	108.9		0.97
48	4	19	109.2	109.2	91.5		0.84
43	6	21	109.2	102.3	69.5		0.64
53	1	12	107.5	10.7	99.3		0.92
51	9	30	107.4	107.4	70.1		0.65
62	3	10	105.6	105.6	102.5		0.97
41	4	4	105.6	10.6	102.9		0.97
70	10	11	103.1	103.1	20.5	M	0.20
64	2	1	103.0	103.0	65.9	M	0.64
64	6	17	100.8	100.8	59.8		0.59
48	4	18	99.3	9.9	91.6		0.92
71	4	5	97.7	97.7	74.5		0.76
51	1	12	90.6	77.9	80.5		0.89
41	10	30	90.1	9.1	79.5		0.88
59	5	15	89.2	89.2	80.8		0.91

Table 30 Comparison of daily rainfall totals obtained from three sources for the thirty largest events for period 1961 - 1991 at Station 0088293 (Sutherland)

Year	Month	Day	Daily Rainfall Total			Digitised Flag (M=Missing)	Digitised / SAWB Daily (ratio)
			SAWB Daily (mm)	SAWB Control (mm)	Digitised (mm)		
80	3	11	86.0	86.0	83.8		0.97
76	1	20	62.0	62.0	58.4	M	0.94
66	3	20	52.5	0.0	0.0	M	0.00
81	3	25	50.7	50.7	34.5	M	0.68
85	1	16	49.3	0.0	0.0		0.00
67	6	9	49.0	49.0	15.8	M	0.32
91	1	25	42.6	22.0	0.0	M	0.00
65	3	22	41.5	8.6	8.4	M	0.20
83	5	13	41.0	41.0	40.1		0.98
73	3	18	41.0	41.0	29.4	M	0.72
86	6	2	39.8	39.8	24.2	M	0.61
85	1	14	39.3	39.3	34.9	M	0.89
76	2	4	39.0	39.0	34.7	M	0.89
81	1	24	38.0	38.0	33.3		0.88
90	4	21	34.8	34.8	29.9	M	0.86
82	4	6	34.3	34.3	33.9		0.99
73	7	1	34.0	34.0	32.3		0.95
86	4	25	33.6	33.6	9.4	M	0.28
75	12	22	31.0	31.0	28.9		0.93
80	11	28	30.7	30.7	29.1		0.95
76	11	4	29.5	0.0	0.0	M	0.00
76	11	23	28.3	28.3	27.6		0.98
90	2	3	28.0	28.0	24.5	M	0.88
74	6	25	27.7	27.7	25.5		0.92
76	2	5	27.2	27.2	25.7		0.94
62	4	22	27.0	2.5	2.3		0.09
85	12	19	27.0	27.0	26.3		0.97
81	3	24	26.7	0.0	0.0		0.00

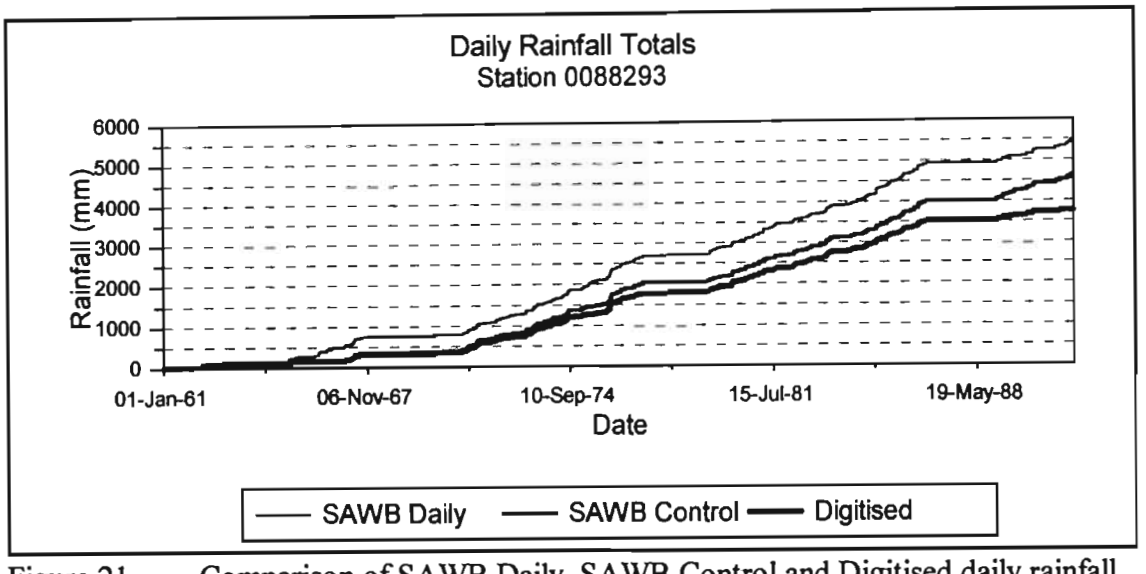


Figure 21 Comparison of SAWB Daily, SAWB Control and Digitised daily rainfall totals at Station 0088293 (Sutherland)

4.3.5 Concluding Remarks on Comparison of Digitised and Standard Raingauge Daily Totals

In the four stations examined, there are differences between the three sources of data which, when accumulated over a number of years of record, amount to a large amount of rainfall. The reason for the differences between the SAWB Daily and SAWB Control values can only be attributed to typographical errors when inputting the data, as the source of the data is the same. Some of the daily rainfall data for SAWB Station 0059572, obtained from the CCWR, appear to be missing. In all the cases investigated, the daily rainfall totals derived from the digitised data are less than the standard gauge values, and in some cases when the digitised daily rainfall total is less than the standard gauge values, no missing data flags have been inserted in the data. It is conceded that on occasion the daily rainfall total derived from the digitised data may correctly be less than the standard gauge value. However, the reasons for the consistent under-estimation of daily rainfall totals in the absence of missing data flags needs to be investigated by the SAWB.

4.4 MAGNITUDE AND FREQUENCY OF ERRORS IN DAILY RAINFALL TOTALS

In order to further quantify how reliable the digitised data are for a particular site, the differences between the standard raingauge (SAWB, obtained from CCWR) and digitised daily totals were computed and categorised. The categories used were differences of 0-5 mm, 5 -10 mm, 10-15 mm, 15-20 mm and > 20 mm, with negative categories indicating that the digitised daily total is greater than the standard gauge totals. For example, the results of the above analysis for SAWB station 0239482 (Cedara) are contained in Figure 22. For this station the majority of raindays have differences between the standard gauge and digitised rainfall totals of less than 5 mm. However, it is disturbing to note that on 58 days the standard gauge values exceeded the digitised values by more than 20 mm, and on 158 days the standard gauge value exceeded the digitised rainfall by more than 10 mm.

As a result of the occasional malfunctioning of the autographic raingauges, it is expected that the standard raingauge totals would exceed those of the digitised values. Hence the days when the digitised values exceed the standard raingauge values in Figure 22 require special investigation. Missing data flags in the digitised data were ignored in the compilation of Figure 22.

A summary of the above analysis for 330 SAWB stations, but with the number of days when the differences fall into different classes expressed as a percentage of the total number of raindays, is shown in Figure 23. Nearly 3% of the recorded raindays from the 330 SAWB stations have differences between the standard raingauge and digitised daily rainfall totals of greater than 20 mm. These differences clearly need further investigation. In Figure 23 missing data flags in the digitised data are ignored.

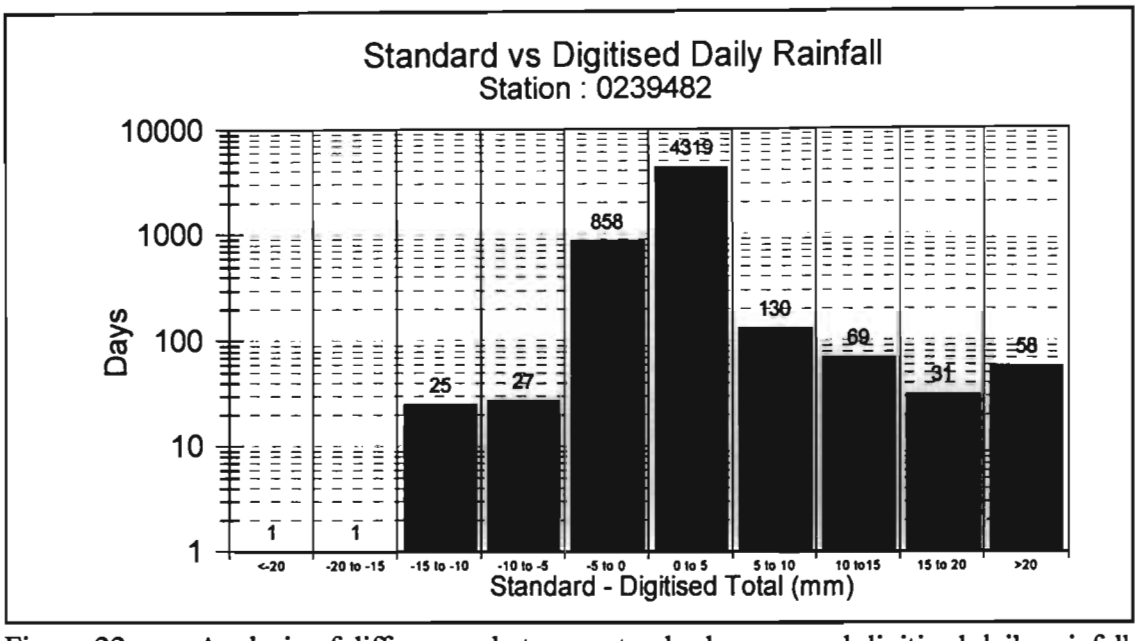


Figure 22 Analysis of differences between standard gauge and digitised daily rainfall totals at Station 0239482, Cedara (days with some missing digitised data included)

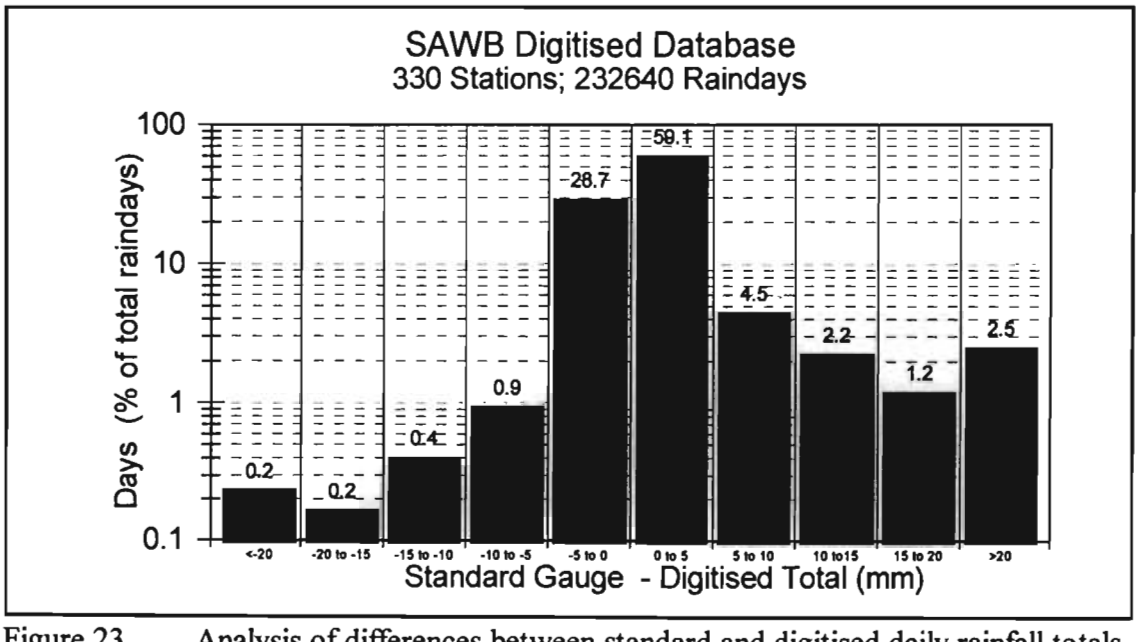


Figure 23 Analysis of differences between standard and digitised daily rainfall totals at 330 SAWB stations (days with some missing digitised data included)

As shown in Figure 24, even when days which have missing digitised data are excluded, there remains an excessive number of days which have large differences between the

standard gauge and digitised daily rainfall totals. When days which have missing digitised data are excluded, nearly 3% of the standard gauge daily totals exceed the digitised data by a magnitude of more than 15 mm.

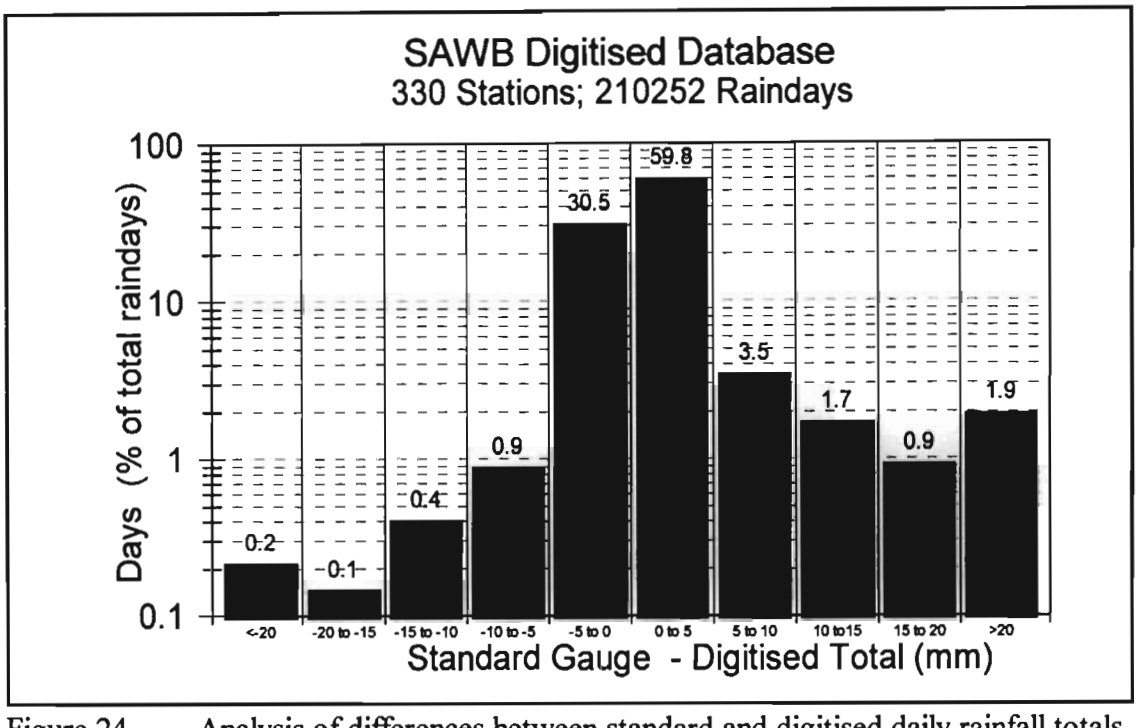


Figure 24 Analysis of differences between standard and digitised daily rainfall totals at 330 SAWB stations (days with some missing digitised data excluded)

4.5 ERRORS IN DAILY RAINFALL TOTALS VS EVENT MAGNITUDE

Based on the assumption that the standard gauge daily rainfall total is the "correct" value, it has been shown that some large errors are contained in the digitised data. However, it is necessary to determine whether the large differences in the digitised and standard gauge daily rainfall totals occur only during large events or whether they occur over a range of rainfall events. For example, in order to investigate the occurrence of the errors as a function of the daily rainfall total, the error (standard - digitised daily rainfall total) for SAWB Station 0239482 (Cedara) was plotted against the standard gauge total, as shown in Figure 25.

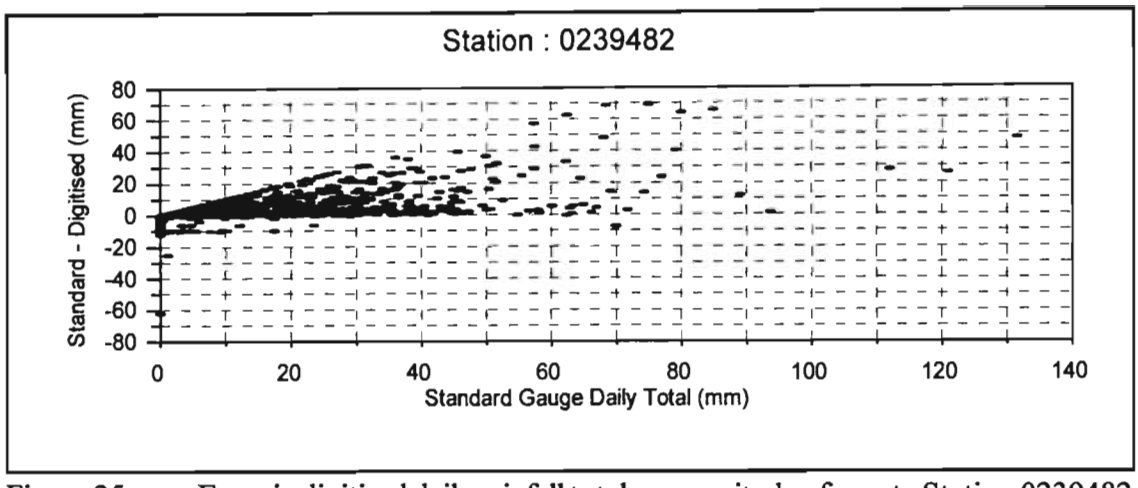


Figure 25 Error in digitised daily rainfall total vs magnitude of event : Station 0239482, Cedara (days with missing data flags in digitised data included)

From Figure 25 it is apparent that errors in the measurement of daily rainfall totals from digitised data occur throughout the range of daily rainfalls. However, it is significant that the largest events could have more than half the rainfall unrecorded in the digitised data. The digitised daily totals in Figure 25 were calculated by ignoring the missing data flags. In Figure 26, days which contained missing data flags were excluded. Assuming that the missing data flags were inserted in the data correctly according to the recorded trace on the chart, then the similarity between Figures 25 and 26 and the errors still evident in Figure 26 indicate that many occasions when the gauge malfunctioned are not reflected in the digitised data.

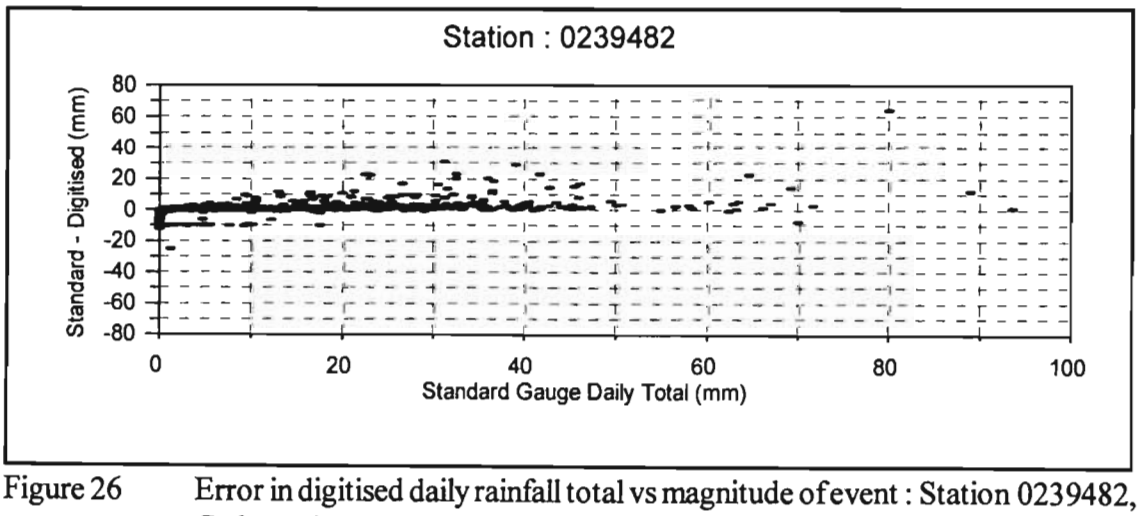


Figure 26 Error in digitised daily rainfall total vs magnitude of event : Station 0239482, Cedara (days with missing data flags in digitised data excluded)

The plot in Figure 25, which includes days with missing digitised data, is summarised in Figure 27. This figure depicts the number of days in which the standard gauge and errors fell into defined classes. A similar analysis utilising data from 330 SAWB stations is shown in Figure 28.

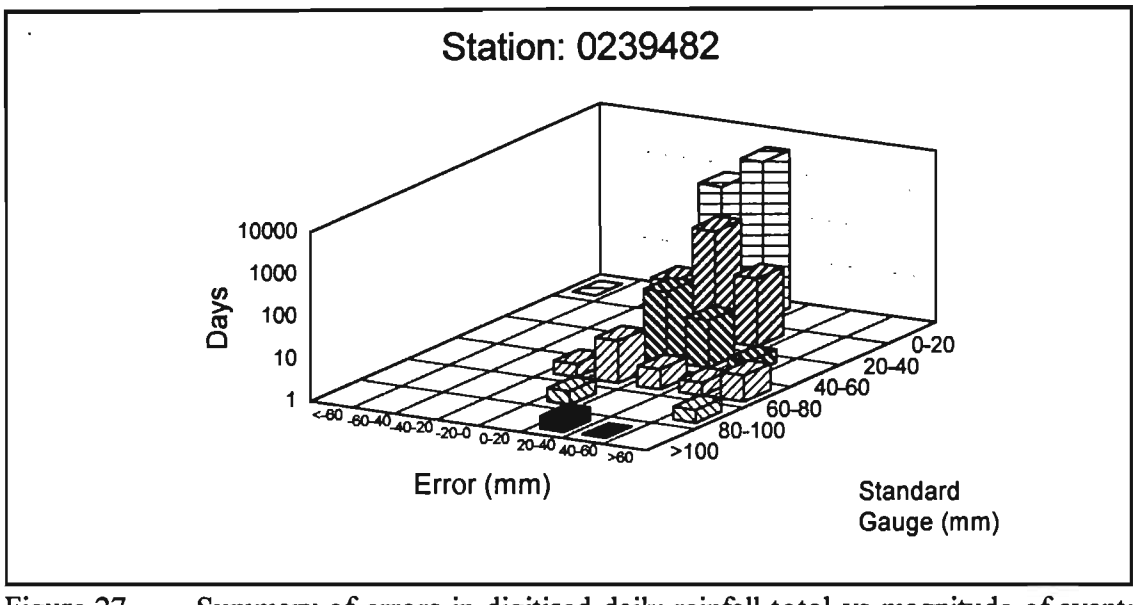


Figure 27 Summary of errors in digitised daily rainfall total vs magnitude of event: Station 0239482, Cedara (days with missing data flags in digitised data excluded)

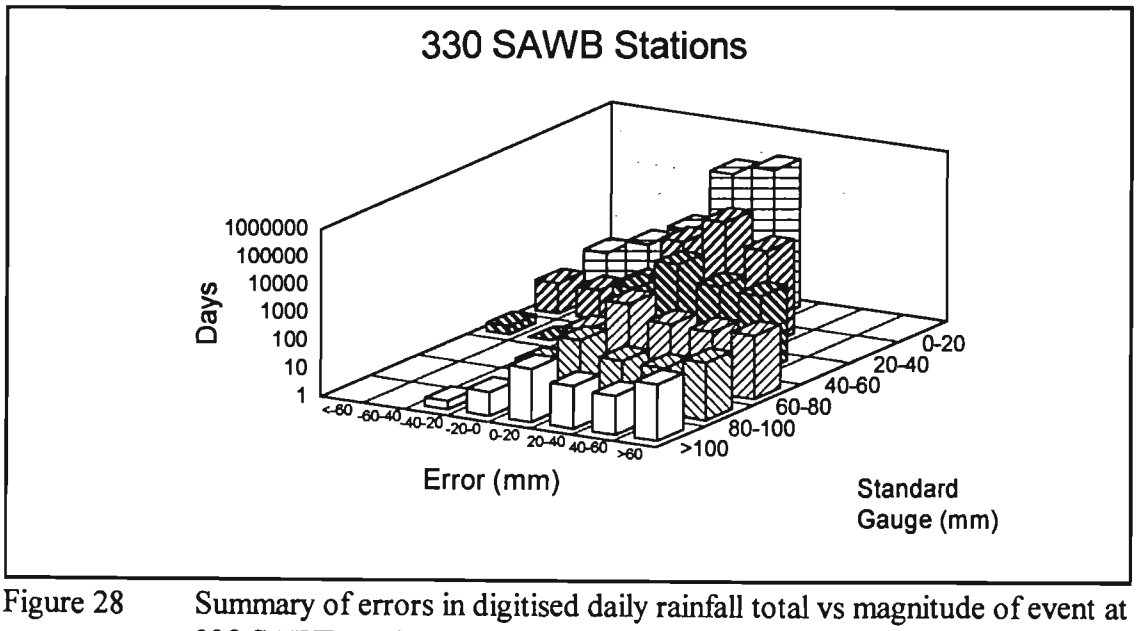


Figure 28 Summary of errors in digitised daily rainfall total vs magnitude of event at 330 SAWB stations

From Figure 28 it is evident that errors in the daily totals computed from digitised data occur across the range of daily rainfall totals, and hence the digitised data need to be adjusted to compensate for the apparent errors.

A Reliability Index (RI) for each SAWB station was developed. This was expressed as the percentage of total raindays where the difference between the digitised and standard gauge daily rainfall totals exceeded 5 mm. A frequency analysis of the RI values for all SAWB stations is shown in Figure 29. Only 1.3% of the SAWB stations have a RI of ≤ 2% and 75.4% of the gauges have a difference larger than 5 mm between the standard and digitised raingauge daily rainfall totals on more than 10% of the raindays.

The processing errors in the SAWB data were corrected and the RANDOM procedure was adopted. However, it was established that considerable amounts of rainfall were either not recorded by the autographic gauges or were not digitised and hence are not contained in the digitised data. In addition many of these missing data are not reflected in the digitised data file as missing data. Hence it is necessary to establish the impact the missing data has on the estimation of design storms.

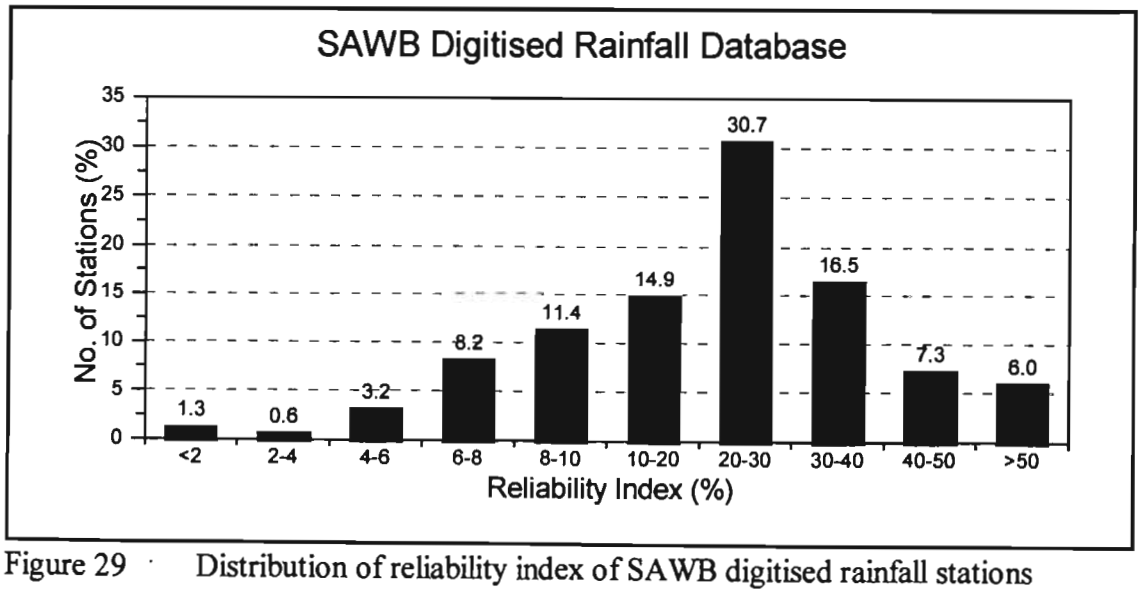


Figure 29 Distribution of reliability index of SAWB digitised rainfall stations

4.6 IMPACT OF INCOMPLETE DATA ON DESIGN RAINFALL ESTIMATES

As shown in Figure 29, 24.7% of the raingauges have a difference larger than 5 mm between the standard and digitised raingauge daily rainfall totals on less than 10% of the raindays. The analysis of the impact of incomplete data on design rainfall estimates was performed at a single station which has a relatively long record length and which has an RI value less than 10%. SAWB Station 0059572 (East London), which has a record length of 51 years and RI=5.8%, was selected as a suitable gauge on which to perform the analysis. The data from SAWB Station 0059572 are viewed as relatively reliable as approximately 95% of the SAWB stations have a reliability index greater than the value for SAWB 0059572.

4.6.1 Methodology

The Partial Duration Series (PDS) and Annual Maximum Series (AMS) for SAWB Station 0059572, used as a case study, were extracted and design rainfall estimates were computed from the AMS for 16 durations ranging from 5 min to 24 h. These values were used to represent design values based on a data set with no missing values.

Thereafter, the AMS was extracted from the same PDS to create an AMS with some of the "true" extreme events missing. This was achieved by not selecting the maximum value in all years, but for a preselected number of years which were randomly chosen, a user specified rank was extracted from the ranked PDS (e.g. second largest, third largest, etc.). Thus an AMS was constructed having "missing" data (the largest values) and design values were computed from the modified AMS. This process was repeated 100 times and for varying numbers of years having "missing" data and for the second and third largest values used in the modified AMS for the randomly selected years.

Statistical tests were then performed based on the null hypothesis that there were no significant differences between the design rainfall estimates computed from the AMS series extracted from the PDS having no missing data, and extracted from the PDS with some "missing" data. The t-test statistic was used to test the significance of the null hypothesis that the mean of the 100 repetitions of design rainfall values was within 5% of the control value.

4.6.2 Results

The t-test statistic was evaluated for design values at 2 to 100 year return periods and for durations ranging from 5 minutes to 24 h. Results produced when randomly excluding the largest value from 10% to 50% of the years, and thus extracting the second or third largest value in those years as the annual maximum, are contained in Table 31.

A case study was performed at Station 0059572 to estimate the number of years when the "true" AMS values were not contained in the digitised data. It was assumed that the manually extracted data used by Midgley and Pitman (1978) contained all the maximum events and that, where the manually extracted annual maxima exceeded the digitised annual maxima, the digitised event was not the same as the manually extracted event. Based on these assumptions, Table 32 contains estimates of the percentage of years in which the digitised data do not contain the "true" maximum event.

Based on the above analysis on data from East London, it is concluded that if only the largest event is not contained in the digitised data for 10% of the years, the design rainfall estimates for all durations are not significantly different for all durations and return periods. This generally also holds true for the case when the annual maximum event is excluded for 20% of the years, particularly for longer durations. However, when the annual maximum events are excluded from 30% or more of the years, significantly different design values are obtained for most durations and return periods. In the case when the two largest events are excluded in the randomly selected years, similar trends are evident.

Table 31 Acceptance (✓) and rejection (✗) at the 95% confidence level of the null hypothesis that the mean of 100 design rainfall values, estimated by randomly excluding the largest event(s) from varying percentages of the years, falls within the 5% of the control value: Station 0059572 (East London)

Return Period (Years)	Percentage of Years Selected for which Largest Events Excluded	Rank of Event Excluded in Randomly Selected Years	Event Duration (minutes)															
			5	10	15	30	45	60	90	120	240	360	480	600	720	960	1200	1440
2	10	1	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	20	1	✓	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓	✓	✓
	30	1	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
	40	1	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
	50	1	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
2	10	2	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	20	2	✓	✓	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓	✓	✓
	30	2	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
	40	2	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
	50	2	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
5	10	1	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	20	1	✓	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	30	1	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
	40	1	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
	50	1	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
5	10	2	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	20	2	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	30	2	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
	40	2	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
	50	2	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗

Return Period (Years)	Percentage of Years Selected for which Largest Events Excluded	Rank of Event Excluded in Randomly Selected Years	Event Duration (minutes)															
			5	10	15	30	45	60	90	120	240	360	480	600	720	960	1200	1440
10	10	1	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	20	1	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	30	1	×	×	×	×	×	×	×	×	×	×	×	×	×	✓	✓	✓
	40	1	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×
	50	1	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×
10	10	2	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	20	2	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	30	2	×	×	×	×	×	×	×	×	×	×	×	×	×	×	✓	✓
	40	2	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×
	50	2	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×
20	10	1	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	20	1	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	30	1	✓	✓	×	×	✓	✓	×	×	✓	✓	✓	✓	✓	✓	✓	✓
	40	1	×	×	×	×	×	×	×	×	×	×	✓	×	×	×	✓	✓
	50	1	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	✓
20	10	2	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	20	2	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	30	2	✓	×	×	×	×	×	×	✓	×	✓	✓	✓	✓	✓	✓	✓
	40	2	×	×	×	×	×	×	×	×	×	×	×	×	×	×	✓	✓
	50	2	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×

Return Period (Years)	Percentage of Years Selected for which Largest Events Excluded	Rank of Event Excluded in Randomly Selected Years	Event Duration (minutes)																
			5	10	15	30	45	60	90	120	240	360	480	600	720	960	1200	1440
50	10	1	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	20	1	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	30	1	✓	✓	×	×	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	40	1	✓	×	×	×	×	×	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	50	1	✓	×	×	×	×	×	×	✓	×	✓	✓	✓	×	×	✓	✓
50	10	2	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	20	2	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	30	2	✓	×	×	✓	✓	×	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	40	2	✓	×	×	✓	✓	×	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	50	2	×	×	×	×	×	×	×	✓	×	✓	✓	✓	×	×	✓	✓
100	10	1	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	20	1	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	30	1	✓	✓	×	×	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	40	1	✓	×	×	✓	✓	×	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	50	1	✓	×	×	×	×	×	×	✓	✓	✓	✓	✓	✓	✓	✓	✓
100	10	2	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	20	2	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	30	2	✓	×	×	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	40	2	✓	×	×	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	50	2	✓	×	×	×	×	×	×	✓	✓	✓	✓	✓	✓	✓	✓	✓

Table 32 Estimated percentage of years with "true" annual maxima missing in the digitised data: Station 0059572 (East London)

	Duration (minutes)				
	15	30	45	60	1440
Years (%)	32	32	35	35	38

4.6.3 Concluding Remarks on the Impact of Incomplete Data on Design Rainfall Estimation

Based on the deductions made above, it is estimated that at Station 0059572 the annual maximum events are not contained in the digitised data in at least 30% of the years (cf. Table 32). Hence it is concluded that the digitised data at this station, when used to perform design rainfall estimation, will underestimate the true design values. As shown in Figure 29, the reliability index of 5.8% for Station 0059572 indicates that the data for this station are relatively reliable, and that approximately 95% of the SAWB stations have data which are less reliable. It is thus hypothesised that at the majority of SAWB stations the impacts of missing data on design rainfall values would be similar to or greater than the impacts obtained at Station 0059572.

4.7 CHAPTER CONCLUSIONS

A short duration rainfall database consisting of 412 stations was compiled. The major portion (81%) of the data were contributed to the database by the SAWB. Numerous errors such as negative and zero time step errors were found in the SAWB digitised data which prompted the development of automated correction procedures. A clear distinction was drawn between adjustments, where the probable cause of the error is known, and errors, where the cause of the error was unknown. Five procedures were developed to correct

these errors with unknown causes. The effect on the AMS of the different procedures was investigated and it was concluded that the exclusion of erroneous data points or events, which had an error at the beginning or end of the event, was not an acceptable procedure. The recommended method to correct the errors in the data was a random selection of either the MIA, LIA or AIA procedures, and the RANDOM procedure was shown to have no significant effect on the extracted AMS.

A comparison of the digitised and manually extracted AMS at a number of sites indicated that many extreme events were not contained in the digitised data. This was attributed to inadequate digitisation procedures as the same autographic charts were used in both methods of data extraction. The adequacy of the digitised data was further assessed by a comparison of daily rainfall totals computed from the digitised data with daily rainfall values recorded by standard raingauges at the same location. At all the sites investigated, the majority of the daily rainfall totals derived from the digitised data were less than the standard raingauge values, thus indicating significant periods of missing data in the digitised record. It was found that these periods of missing data were frequently not flagged as missing in the digitised data and hence the missing codes in the digitised data were viewed as unreliable.

The reliability of the digitised data was established by the frequency of the differences between the digitised and standard daily rainfall totals. More than 75% of the SAWB stations have greater than 10% of raindays which have differences larger than 5 mm between the digitised and standard gauge daily rainfall totals. It was found that nearly 3% of the recorded raindays from 330 SAWB stations have differences between the digitised and standard raingauge daily totals of greater than 20 mm. These errors were found to occur over the whole range of daily rainfall totals, and were not only associated with smaller events and thus could not be ignored for the purposes of design rainfall estimation.

The impact of incomplete or missing data on design rainfall values at East London was assessed by randomly removing maxima and it was found that, for most return periods and particularly for longer durations, there was no significant effect on the design values if up

to 20% of the years in the AMS do not contain their true maximum value. However, if 30% or more of the years have their annual maximum event missing, then significant differences in the design values were noted. At Station 0059572, which is considered to have relatively reliable data, it was estimated that at least 30% of the annual maxima which were manually extracted from the autographic charts were not contained in the digitised data. It is postulated that the effect of missing data on design rainfall estimates at the majority of SAWB stations are likely to be similar to, or larger, than those demonstrated at Station 0059572, because approximately 95% of the SAWB stations have digitised rainfall data which are less reliable than the data for Station 0059572.

A considerable amount of evidence in this chapter indicates that the majority of the SAWB digitised rainfall data were not reliable enough to use in the estimation of design rainfalls. Further evidence of this assertion is further illustrated in Chapters 5, 6 and 7 where comparisons between the 24 h and 1 day design rainfall values are made.

The re-digitisation of the SAWB charts, or even the re-digitisation of charts which should contain large events as recorded by the standard raingauge was, from a labour and cost point of view, not a viable option for this study. A list of days when large events occurred was provided to the SAWB for possible re-digitisation of the charts for these days, but no new data was forthcoming. What is thus required is to develop techniques to estimate design storms from the digitised database and to make some compensation for the inadequate digitised data and/or to develop techniques to estimate short duration design storms from the more reliable and spatially more dense standard daily raingauge network. The results from one such technique, the use a regional approach to design rainfall estimation, is presented in Chapter 5 following.

CHAPTER 5

DESIGN RAINFALL ESTIMATION USING A REGIONALISED APPROACH

As shown in Chapter 4, only 49 stations in South Africa have short duration rainfall data with a record length ≥30 years. In addition, the data contributed by the SAWB, who contributed the majority of the data to the short duration rainfall database compiled for South Africa in this study, are regarded as generally unreliable. Hence the problem of estimating short duration design rainfalls for South Africa using a database with relatively few stations which have short record lengths, is exacerbated by the majority of the data not being reliable. One technique which has been successfully applied in other studies for improving the reliability of design rainfall estimates from limited data, as discussed in Chapter 2, is to adopt a regional approach.

As discussed in Section 2.2.1, the advantages of using a regionalised approach to design storm estimation is that the information from the limited and relatively short record lengths available is supplemented with spatial information, thereby enabling more reliable design estimates to be obtained. Various methods of regionalisation are summarised in Table 5 and desirable concepts and principles to be incorporated in a regional approach to design storm estimation are outlined in Section 2.2.2. The regional, index storm approach based on L-moments, reported by Hosking and Wallis (1997) and termed the Regional L-Moment Algorithm (RLMA), incorporates these concepts and principles. In addition, a number of studies reviewed in Chapter 2 have successfully used the RLMA and it was concluded that this approach was appropriate for this study. The use of a cluster analysis of site characteristics to group stations, and not any of the other methods listed in Section 2.2.3.2, enables independent testing of clusters of stations for homogeneity using statistics computed from at-site data.

After initial screening of the data to identify gross errors and inconsistencies, as addressed in Section 5.1 for selected sites, relatively homogeneous regions are identified by a cluster

analysis of site characteristics (e.g. latitude, longitude, altitude, MAP etc) and the heterogeneity of the regions, or clusters, is evaluated using at-site data. The regions are assumed to be homogeneous and thus the frequency distribution at all the sites in the region are assumed to be identical apart from a site-specific scaling factor, the index rainfall. The regional average L-moment ratios are computed by weighting according to an individual site's record length. These regional average L-moment ratios are equated to the population L-moment ratios and used to fit the distribution. Hence it is necessary to determine the most appropriate distribution to use for each cluster. This distribution, after appropriate re-scaling by the at-site index value, is used at each site to estimate quantiles. The results of the implementation of the RLMA in South Africa are reported in Section 5.2. At ungauged sites or at sites where the data are unreliable, it is necessary to estimate the index value in order to use the regional growth curve to estimate design rainfalls at that site. The regional growth curve, as described in Section 2.2.3, is the relationship between the ratio of the design storm and an index storm and return period. The accuracy of design storms estimated using regional growth curves is assessed in Section 5.3. The results of estimating the 24 h index storm at ungauged sites in South Africa are presented in Section 5.4 and the selection of an appropriate probability distribution is addressed in Section 5.5.

5.1 EVALUATION OF DISCORDANCY MEASURE

When performing a regional rainfall frequency analysis it is necessary to ensure that the data are a true representation of the rainfall and must be homogeneous i.e. all the data are drawn from the same frequency distribution. Statistical tests for outliers and trends in the data are well established in the literature. In a regional context and using L-moments, Hosking and Wallis (1993) developed a discordancy index (D), as described in Section 2.2.3.1 and formalised in Equation 11, based on the L-skewness vs L-CV plot to test for incorrect data values, outliers, trends and shifts in the mean of samples. Any points on the L-skewness vs L-CV plot which are far from the centre of the cloud are flagged as being discordant. For samples sizes > 14, a station with $D > 3$ is considered to be discordant with the rest of the group (Hosking and Wallis, 1997). This index was used to screen and identify discordant

data. Examples of using the discordancy measure on data from the Cedara and Ntabamhlope research catchments are presented in the following two sections.

5.1.1 Cedara Catchments

Rainfall data from 12 sites, listed in Table 33, from the Cedara (C) catchments are available with record lengths varying from 12 to 21 years. The discordancy index, as described in Section 2.2.3.1, was computed for 16 rainfall durations ranging from 5 min to 24 h using Fortran routines provided by Hosking (1996). Based on the Hosking and Wallis (1997) criterion ($D>3$), the data for the 10 min duration from site C163 was discordant from the rest of the data. The L-moment statistics are plotted in Figure 30 and it is clear that the statistics from one site (C163), which is circled in Figure 30, are different to those from the other sites.

Table 33 Cedara rainfall stations used in the evaluation of discordancy

Station Number	Latitude (S)			Longitude (E)		
	°	'	"	°	'	"
C161	29	35	13	30	13	38
C162	29	34	40	30	13	53
C163	29	33	50	30	15	10
C164	29	34	0	30	14	22
C165	29	33	0	30	14	45
C172	29	34	10	30	15	50
C173	29	33	50	30	15	0
C182	29	35	18	30	14	50
C191	29	32	37	30	16	34
C201	29	32	40	30	16	57
C202	29	32	0	30	17	0
C181	29	35	43	30	15	43

Plots of the 10 min AMS from sites C163, C164 and C182 are shown in Figure 31. From Figure 31 it is evident that the extreme event recorded at C163 for the 1989 wet season is much larger than that at the neighbouring sites.

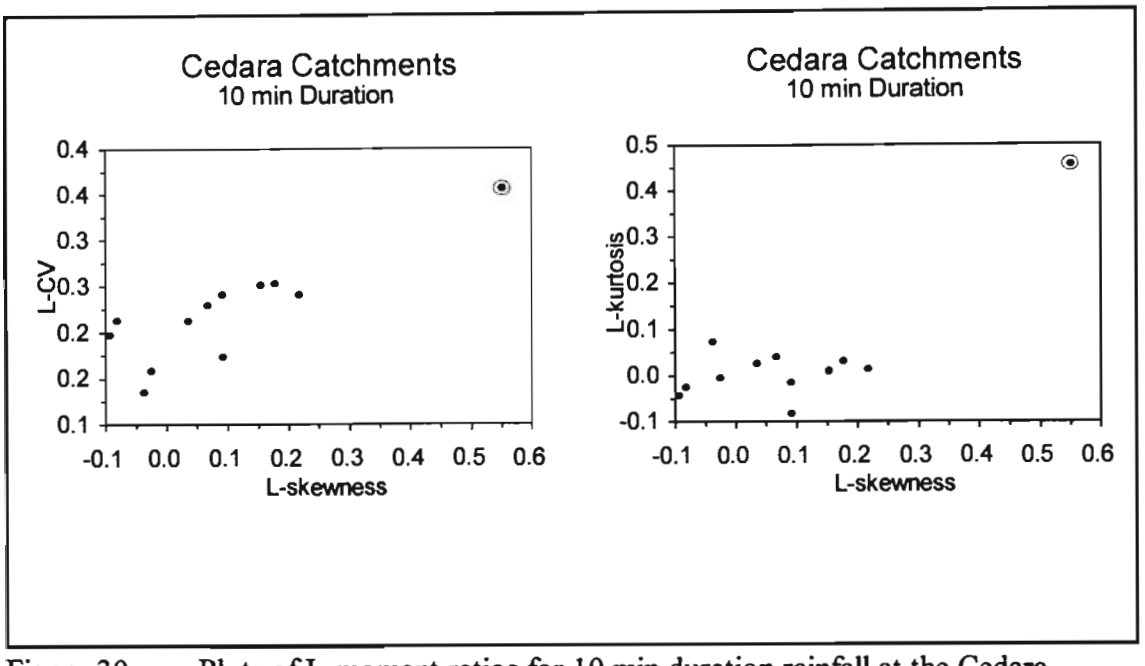


Figure 30 Plots of L-moment ratios for 10 min duration rainfall at the Cedara catchments

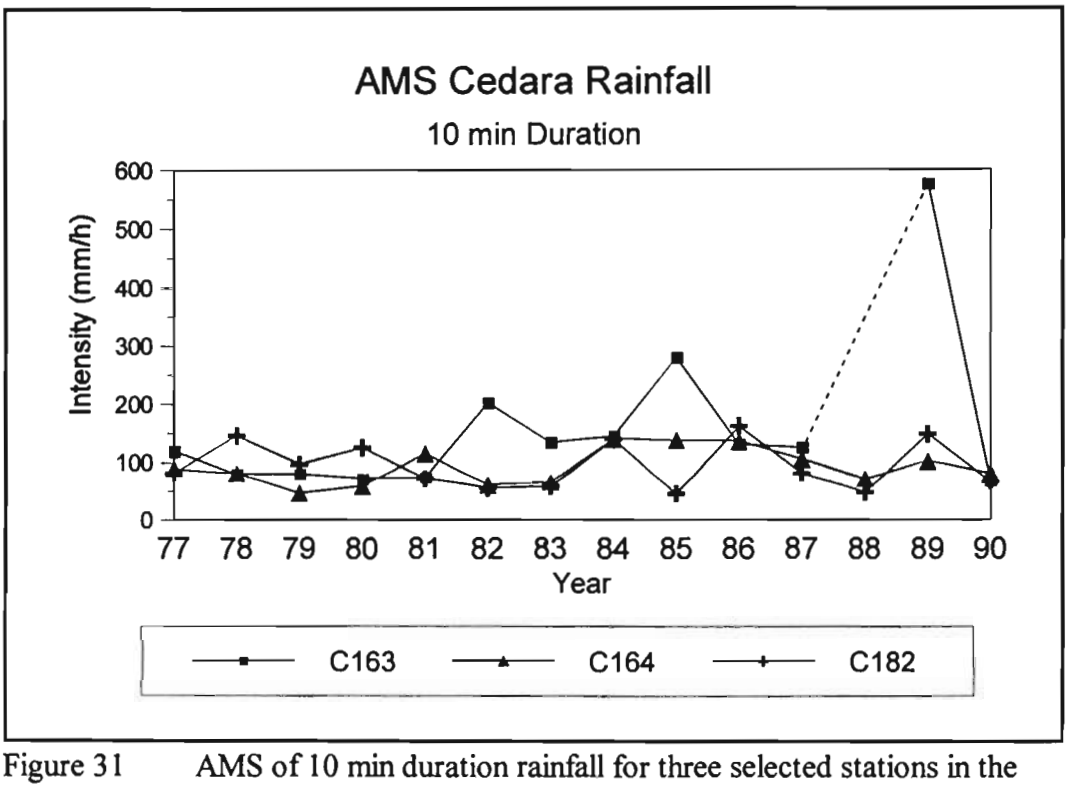


Figure 31 AMS of 10 min duration rainfall for three selected stations in the Cedara catchments (dashed line indicates missing data)

It is therefore hypothesised that the data for 1989 from C163 are suspect. This could have been due to the change-over from autographic recorders to data loggers which occurred in 1989. A comparative plot for the period October 1988 to September 1989 of accumulated daily rainfall at C163 and at neighbouring stations is shown in Figure 32 and confirms that the data from C163 are suspect. Thus, at the Cedara catchments, the discordancy measure (*D*) successfully identified inconsistencies in the data. Discordant data, such as from C163, were not included in further analyses.

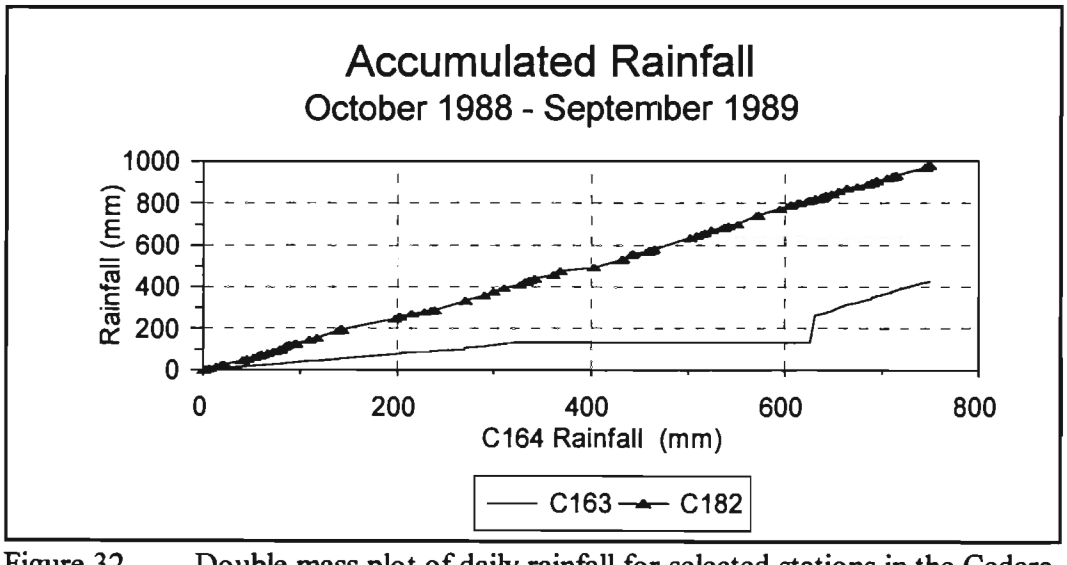


Figure 32 Double mass plot of daily rainfall for selected stations in the Cedara catchments for the period October 1988 - September 1989

5.1.2 Ntabamhlope Catchments

Similar to the Cedara catchments, the Ntabamhlope catchments are research catchments maintained by the DAEUN. Thus the quality of data from both catchments is expected to be better than other data which are recorded as part of a national operation. No discordant data were detected from the 10 De Hoek (D) and Ntabamhlope (N) catchment raingauge sites, listed in Table 34. By way of example, the L-moment ratio plots for the 24 h annual maximum events are shown in Figure 33.

Table 34 Ntabamhlope rainfall stations used in evaluation of discordancy

Station Number	Latitude (S)			Longitude (E)		
	°	'	"	°	'	"
D1	29	00	07	29	39	55
D4	29	00	40	29	39	10
N11	29	00	44	29	37	38
N14	29	02	04	29	39	57
N18	29	02	26	29	39	43
N20	29	01	10	29	40	21
N21	29	02	39	29	38	47
N23	29	03	29	29	39	23
N40	29	02	08	29	35	54
N41	29	04	06	29	37	44

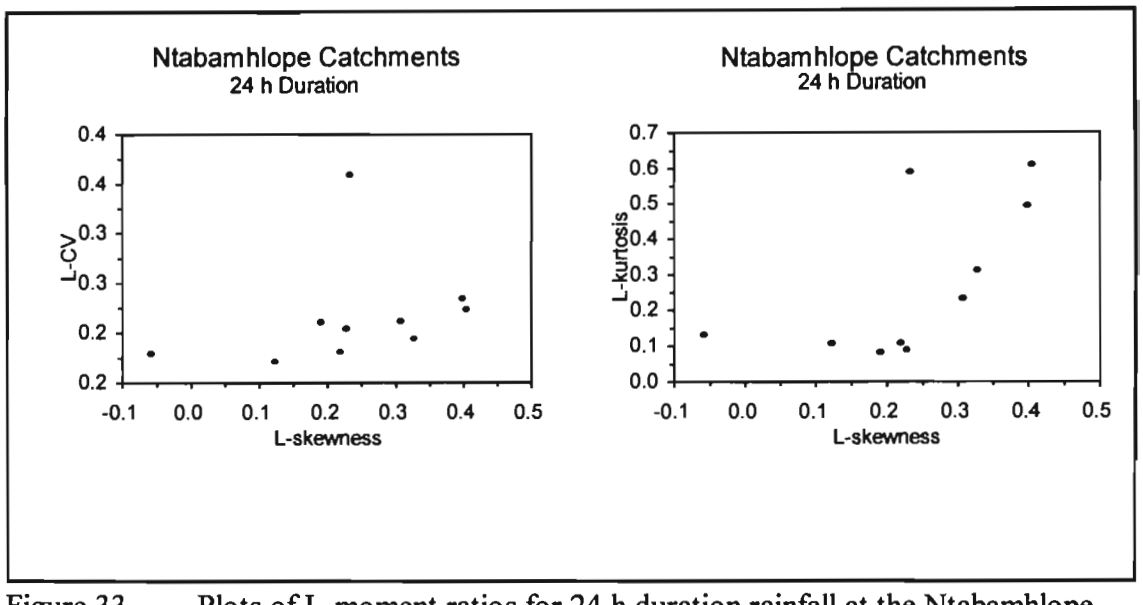


Figure 33 Plots of L-moment ratios for 24 h duration rainfall at the Ntabamhlope catchments

5.1.3 Concluding Remarks on Discordancy Measure

Based on the above analyses, it appears that the discordancy measure developed by Hosking and Wallis (1993; 1997) is an effective tool for initial screening of the data and thus to detect probable errors in the data. The index is easy to use and is compatible with the

regional L-moment approach to frequency analysis. Thus it was adopted for use on all short duration rainfall data prior to initial regionalisation of the data.

5.2 REGIONALISATION USING L-MOMENTS

The results of a homogeneity test of the frequency distributions of the 24 h AMS extracted from all available short duration rainfall data in SA which had 10 or more years of data, indicated that sub-division or regionalisation was necessary. Initial regionalisation of the frequency distribution of short duration rainfall was performed using criteria used previously in SA (Midgley and Pitman, 1978) for short duration rainfall frequency analysis, which were based on identifiable criteria such as Mean Annual Precipitation (MAP) and distance from sea (inland/coastal). Attempts to create geographically contiguous and relatively homogeneous regions based on these criteria proved to be fruitless. Hence the regional L-moment algorithm (RLMA) advocated by Hosking and Wallis (1997) was adapted and applied.

The *rationale* behind the RLMA, as described in Section 2.2.3, is that homogeneous regions are identified based only on site characteristics. The homogeneity of the regions can then be checked independently based on site statistics computed from the at-site data.

5.2.1 Stations Used

Rainfall stations which had 10 or more years of record and which contained the necessary information to perform a regional frequency analysis were extracted and 172 (DAEUN=15; CTCE=2, CSIR=2; SASEX=4, SAWB=137; UZ=12) stations in South Africa met these requirements. The location of the stations are shown in Figure 34. The site characteristics and cluster locations of all these stations used in the cluster analysis are listed in Appendix A. Regionalisation of sites using only site characteristics was performed by cluster analysis using routines from the SAS statistical software (SAS, 1989). The cluster analysis is the

most subjective aspect of the RLMA and it may be necessary to relocate sites/create new clusters subjectively, but based on geographical and physical considerations (Hosking and Wallis, 1997). In the cluster analysis, a vector of site characteristics is associated with each site and standard multivariate statistical analysis is performed to group sites according the similarity of the vectors (Hosking and Wallis, 1997).

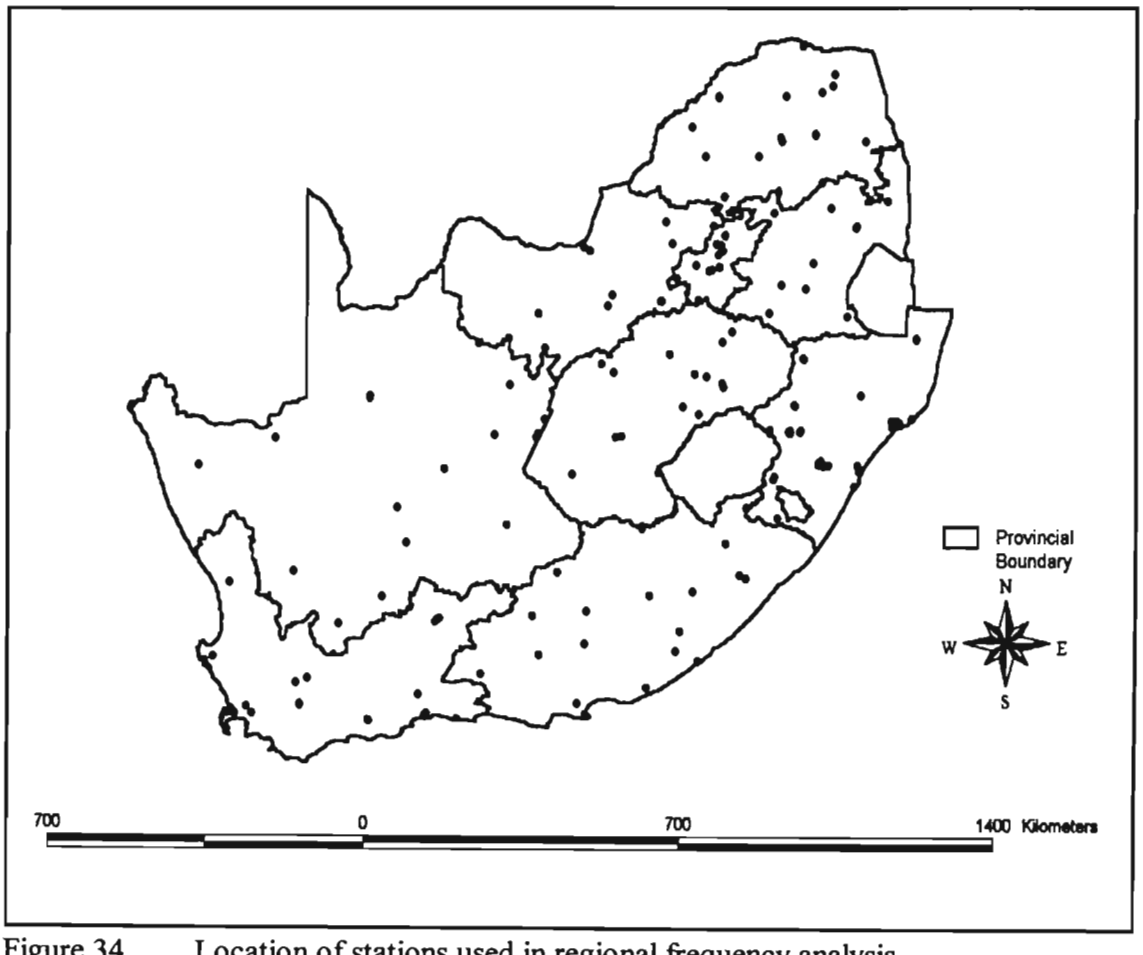


Figure 34 Location of stations used in regional frequency analysis

5.2.2 Site Characteristics Used

The following site characteristics were used in the cluster analysis:

- latitude (°),

- longitude (°),
- altitude (m),
- concentration of precipitation (%),
- mean annual precipitation (mm),
- seasonality (category), and
- distance from sea (m).

The rainfall seasonality information was extracted from Schulze (1997) and is computed as:

$$P_{\%,i} = 0.25 \times \frac{(P_{m,i-1} + 2P_{m,i} + P_{m,i+1})}{MAP} \times 100 \qquad ...58$$

where

$P_{\%,i}$ = smoothed concentration of precipitation for i-th month,

$P_{m,i}$ = median monthly rainfall for i-th month (mm), and

MAP = mean annual precipitation (mm).

Using $P_{\%,i}$ a site is categorised as all year ($P_{\%,1\text{-}12} > 20\%$), winter ($P_{\%,6\text{-}8} > 8\%$), early summer ($P_{\%,12} > 8\%$), mid summer ($P_{\%,1} > 8\%$), late summer ($P_{\%,2} > 8\%$) or very late summer ($P_{\%,3\text{-}5} > 8\%$).

Gridded values of the concentration of precipitation were generated by Schulze (1997), which are based on Markham's technique (Markham, 1970). This is a monthly rainfall index and an index of 100% would imply that the rainfall all fell within one month of the year and an index of 0% would indicate that each month of the year received the same amount of rainfall.

5.2.3 Initial Transformation of Site Characteristics

Cluster analysis was used in the regionalisation in order to identify groupings of sites which were relatively homogeneous. Cluster analysis is very sensitive to the Euclidian distance or

scale (Hosking and Wallis, 1997). A number of different transformations were evaluated and the final transformations which gave the best results and which were implemented are summarised in Table 35. The site characteristics from the 172 stations were used in a cluster analysis using Ward's minimum variance hierarchical algorithm (SAS, 1989), which tends to form clusters of roughly equal size (Hosking and Wallis, 1997).

Table 35 Initial transformations of site characteristics

Site Characteristic (X)	Cluster Variable (Y)	Site Characteristic (X)	Cluster Variable (Y)
Latitude (° decimal)	$\frac{X}{90} \times 100$	Concentration of Precipitation (%)	X (Untransformed)
Longitude (° decimal)	$\frac{X}{90} \times 100$	Seasonality (category)	$\frac{X}{10} \times 100$
Altitude (m)	$\frac{X}{X_{max}} \times 100$	Distance to Sea (m)	$\frac{X}{X_{max}} \times 100$
MAP (mm)	$\frac{X}{X_{max}} \times 100$		

Fifteen regions were identified in the cluster analysis of site characteristics. These were tested for homogeneity based on a heterogeneity measure (H), which utilises L-moment ratios as described in Section 2.2.3.2 and in Equation16, and was implemented using routines provided by Hosking (1996). As discussed in Section 2.2.3.2, the objective is to estimate the degree of heterogeneity within a group of sites and to test whether the region may reasonably be treated as a homogeneous region. According to Hosking and Wallis (1997) a region with a value of $H < 1$ is considered to be "acceptably homogeneous", when $1 < H < 2$ it is "possibly heterogeneous" and when $H > 2$ it is "definitely heterogeneous". Table 36 contains the results of the heterogeneity measure.

The 24 h AMS was used to assess the homogeneity of the clusters. A different set of relatively homogeneous clusters could be obtained for different storm durations. However, as the cluster analysis is based on site characteristics, the allocation of stations to clusters should not change, except for the subjective relocation of clusters. In addition, having a different set of clusters for each duration is not practical (Wallis, 1997). This approach of using the same clusters for different durations was also used by Werick *et al.* (1993) in the creation of a National Drought Atlas for the USA.

From Table 36 and the spatial distribution of the clusters it was evident that for Cluster15, which is definitely heterogeneous, very large spatial distances between the sites in the region were noted. Therefore, it was suspected that the transformation used for the latitude and longitude results in a smaller range for these characteristics which therefore have less weight in the cluster analysis. The reasons for the heterogeneity in the other regions (6 and 7) are not clear. However, as pointed out by Hosking and Wallis (1997), the cluster analysis is the most subjective aspect of the RLMA and it may be necessary to relocate sites/create new clusters subjectively, but based on geographic and physiographic considerations.

Table 36 Results of heterogeneity tests for clusters identified using site characteristic transformations listed in Table 35

Cluster	Number of sites	Heterogeneity Measure (H)	Cluster	Number of sites	Heterogeneity Measure (H)
1	13	1.0	9	24	1.1
2	6	1.1	10	9	0.4
3	9	0.3	11	10	1.2
4	23	0.4	12	4	0.8
5	16	1.3	13	7	0.6
6	7	2.2	14	5	0.5
7	10	5.6	15	6	3.6
8	7	0.8			

5.2.4 Modified Transformations of Site Characteristics

In an effort to decrease the heterogeneity within clusters, as shown in Table 36, and to decrease the spatial distances between sites within a cluster, the transformations listed in Table 37 were implemented. These modified transformations attempted to ensure equitable scales between the different site characteristics.

Table 37 Final transformations of site characteristics

Site Characteristic (X)	Cluster Variable (Y)	Site Characteristic (X)	Cluster Variable (Y)
Latitude (° decimal)	$\frac{X - X_{min}}{X_{max} - X_{min}} \times 100$	Concentration of Precipitation (%)	$\frac{X - X_{min}}{X_{max} - X_{min}} \times 100$
Longitude (° decimal)	$\frac{X - X_{min}}{X_{max} - X_{min}} \times 100$	Seasonality (category)	$\frac{X - X_{min}}{X_{max} - X_{min}} \times 100$
Altitude (m)	$\frac{X}{X_{max}} \times 100$	Distance to Sea (m)	$\frac{X}{X_{max}} \times 100$
MAP (mm)	$\frac{X}{X_{max}} \times 100$		

The characteristics of the 172 sites were transformed as shown in Table 37 and the results of using Ward's minimum variance hierarchical algorithm on the transformed variables, are presented in Figure 35. In this analysis 17 clusters were created, based on the results of simulation experiments performed by Hosking and Wallis (1997). These indicated that, although the accuracy of the design values estimated using the RLMA increases with an increasing number of stations in a homogeneous region, there is relatively little benefit in having more than 20 stations per cluster when estimating quantiles with return periods $\leq$1000 years.

The regions identified in the cluster analysis of site characteristics shown in Figure 35 were tested for homogeneity using the Hosking and Wallis (1997) heterogeneity test. Table 38 contains the results of the heterogeneity measure (H) for the clusters depicted in Figure 35.

Table 38 Results of heterogeneity tests for clusters depicted in Figure 35

Cluster	Number of Sites	Heterogeneity Measure (H)	Cluster	Number of Sites	Heterogeneity Measure (H)
1	19	0.95	10	8	0.59
2	10	1.04	11	20	0.57
3	32	0.64	12	10	1.20
4	6	-0.76	13	5	-0.79
5	8	1.59	14	7	-0.45
6	9	-1.07	15	5	-0.46
7	14	-0.35	16	3	2.93
8	6	3.06	17	2	1.02
9	8	-0.10			

The negative measures of heterogeneity contained in Table 38 indicate that there is less dispersion in the at-site sample L-CV values than would be expected. However, Hosking and Wallis (1997) indicate that if many large negative values (< -2) are obtained, then the probable cause is positive correlation between the data. Since no values of H < -2 were obtained, the negative values of H were considered not to be the result of positive correlation between the data.

Using the data transformations listed in Table 37, the results in Table 38 indicate that only two regions (8 and 16) were definitely heterogeneous and require further attention.

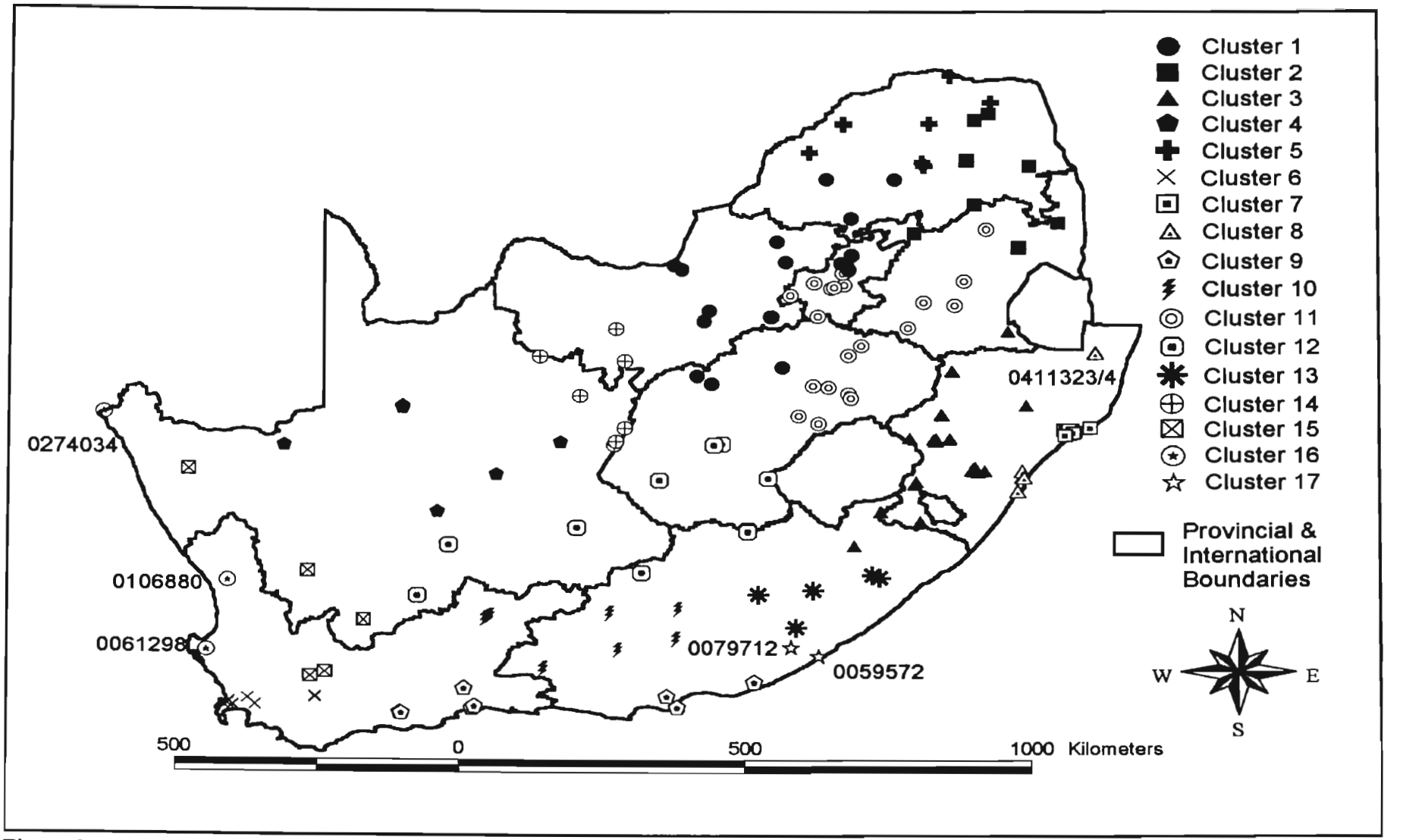


Figure 35 Results from a cluster analysis using final transformations of site characteristics listed in Table 37

5.2.5 Modifications to Regions

According to Hosking and Wallis (1997), subjective intervention, within plausible physical limits, may be required in the final determination of homogeneous clusters. Stations from Clusters 8 and 16 were moved to adjacent regions as indicated in Table 39. In addition, the two stations from cluster 17 were also relocated as it was deemed that a cluster consisting of only two stations was not satisfactory. The location of stations moved between clusters are indicated in Figure 35 by their SAWB station numbers. The relocation of the stations resulted in 15 clusters, with Clusters 16 and 17 having been eliminated. The distribution of the 15 clusters is presented in Figure 36.

Table 39 Relocation of stations between clusters

Station Number	Moved from Cluster	Moved to Cluster
0411323	8	7
0411324		7
0061298	16	6
0106880		15
0274034		15
0079712	17	13
0059572		13

The modified clusters were tested for homogeneity using the Hosking and Wallis' (1997) test. Table 40 contains the results of the heterogeneity measure for the clusters depicted in Figure 36. From the results contained in Table 40 it is concluded that the regions are sufficiently homogeneous for the RLMA to be applied. Thus growth curves, which depict the relationship between return period and the ratio of the design storm and an index storm, can be derived for each cluster.

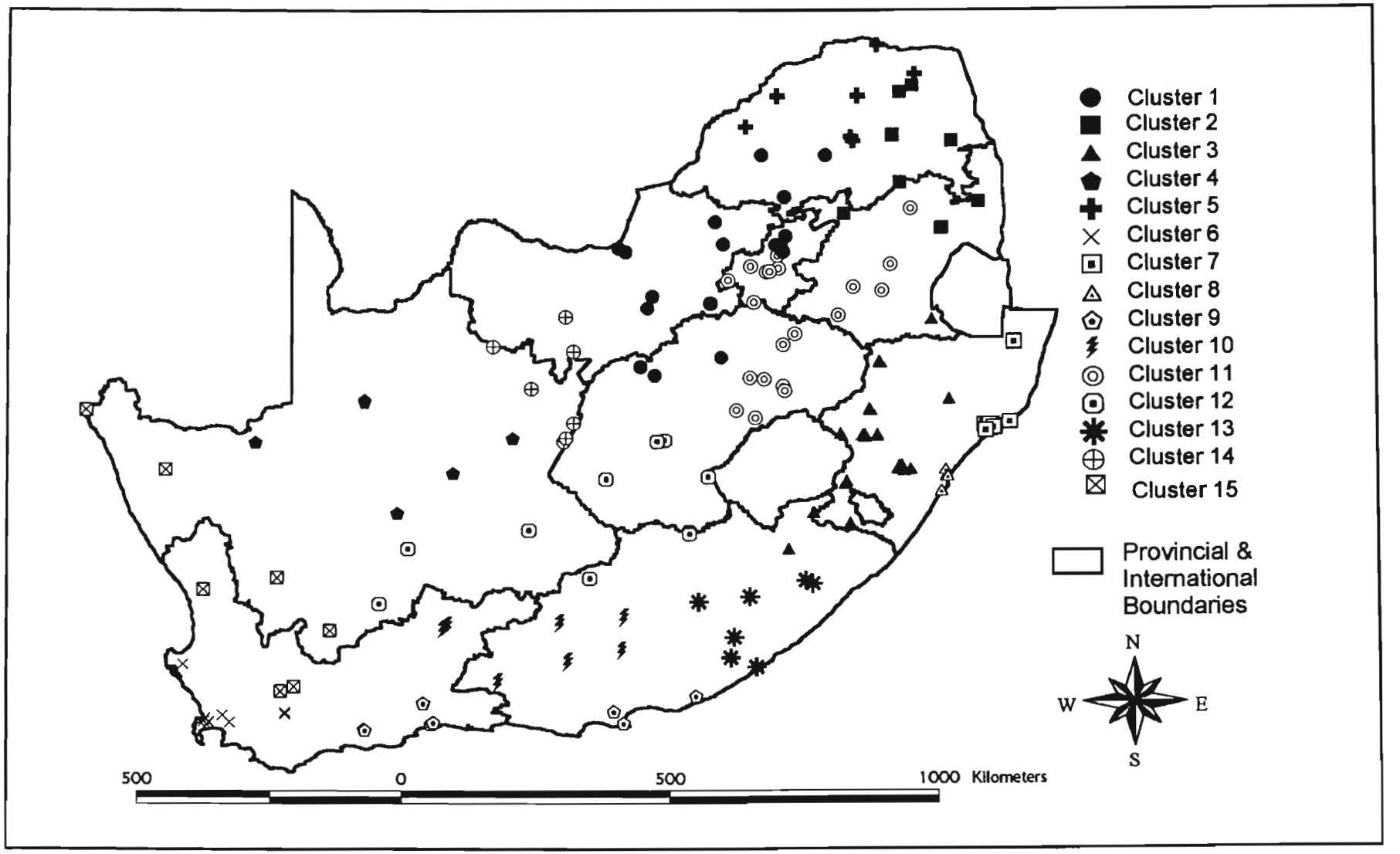


Figure 36 Results from a cluster analysis after relocation of stations as listed in Table 39

Table 40 Results of heterogeneity tests

Cluster	Number of Sites	Heterogeneity Measure (H)	Cluster	Number of Sites	Heterogeneity Measure (H)
1	19	0.95	9	8	-0.10
2	10	1.04	10	8	0.59
3	32	0.64	11	20	0.57
4	6	-0.76	12	10	1.20
5	8	1.59	13	7	0.69
6	10	-1.13	14	7	-0.45
7	16	1.02	15	7	1.67
8	4	0.26			

5.3 REGIONAL GROWTH CURVES

Regional growth curves, developed for each cluster and various durations, relate the ratio between the design rainfall and an index value to return period. Examples of growth curves for selected clusters and various durations are shown in this section. The GEV distribution, which is shown in Section 5.5 to be an appropriate distribution for South Africa, was used to estimate design storms.

5.3.1 Examples

The variation of the regional growth curve of quantiles in Clusters 1 to 6 for two durations are depicted in Figure 37. These examples indicate that the variation between the growth curves for different regions and durations increases with return period. The relatively similar growth curves for some regions may indicate that some regions may be combined. However, Hosking and Wallis (1997) caution against this, arguing that the absence of statistical difference may merely reflect an insufficiency of data.

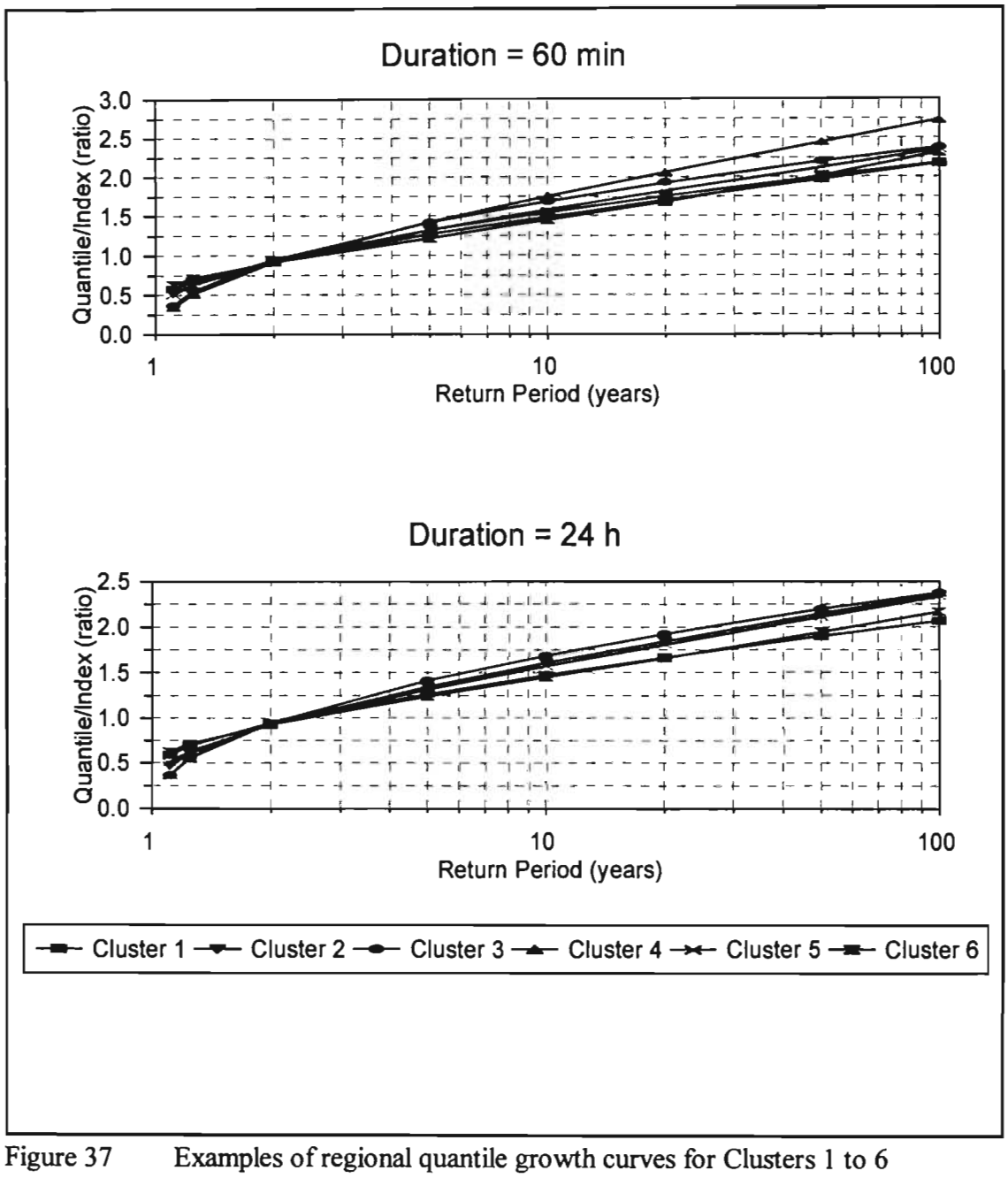


Figure 37 Examples of regional quantile growth curves for Clusters 1 to 6

Another example of the variation in the growth curve with duration is shown for Cluster 3 in Figure 38. In Cluster 3 the growth curve for various durations are very similar for return periods < 10 years, but diverge for longer return periods.

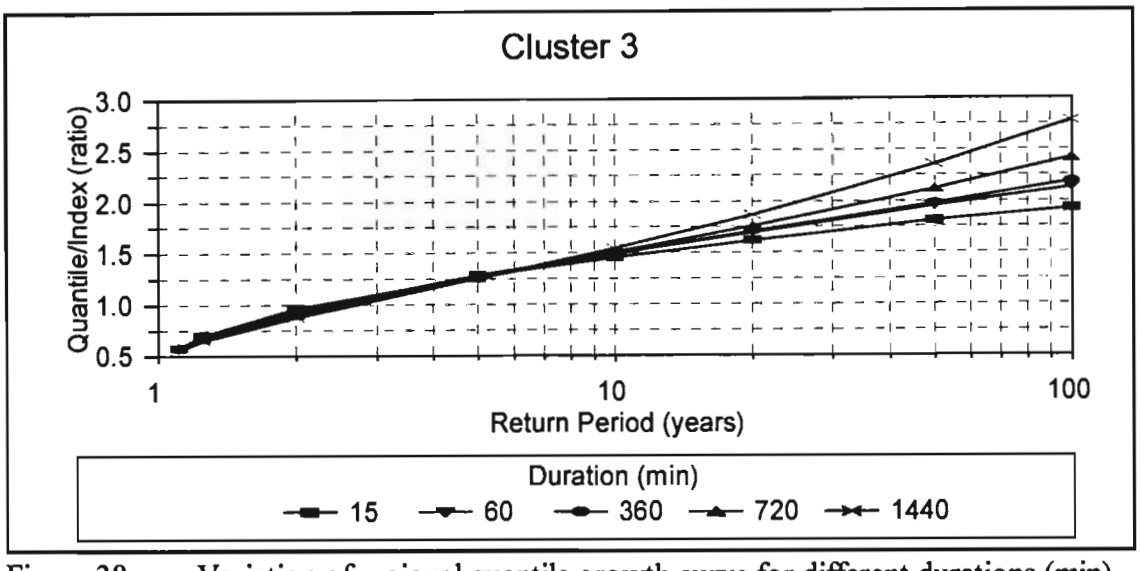


Figure 38 Variation of regional quantile growth curve for different durations (min) in Cluster 3

5.3.2 At-site vs Regional Quantiles

The advantage of using a regionalised approach to design storm estimation is that at-site information is supplemented with information from the entire homogeneous region. Thus the regional estimates of design rainfall are deemed to be more reliable than estimates based only on at-site information. An example of the differences between quantiles estimated using at-site data and the RLMA are shown for 1 h duration events in Figure 39 for five selected stations in Cluster 3. The variation between the quantiles estimated from the at-site data and regional approaches shown in Figure 39, which are less than 15% for all durations, are typical for Cluster 3 and for most other clusters.

Station N23, which has a record length of 32 years, is located in the Ntabamhlope Research Catchments monitored by the DAEUN and was not used in the cluster analyses or in the estimation of the regional growth curves. As shown in Figure 40 there is good agreement between quantiles estimated from the at-site data and from regional analysis for all durations and return periods. Hence it would appear that the RLMA is capable of estimating design

storms reliably. However, a formal assessment of the accuracy and confidence limits of quantiles estimated using the RLMA is necessary.

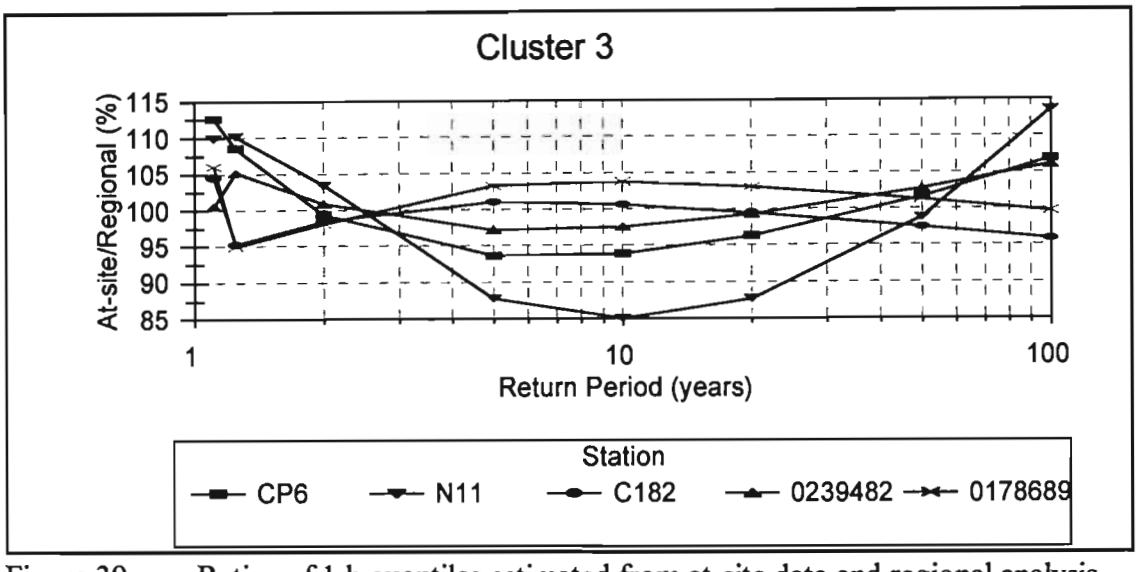


Figure 39 Ratios of 1 h quantiles estimated from at-site data and regional analysis for selected stations in Cluster 3

5.3.3 Assessment of Accuracy of Design Rainfalls Estimated Using the RLMA

Uncertainty is inherent in statistical analysis and hence it is necessary to assess the magnitude of the uncertainty. Traditionally the uncertainty is quantified by constructing confidence intervals for the estimated model parameters and quantiles, assuming that all the statistical model's assumptions are satisfied. The assumptions are rarely, if ever, all true when performing a frequency analysis. Thus a realistic assessment of the accuracy of a regional frequency analysis should account for the possibility of heterogeneity in the regions, inappropriate frequency distribution and dependence between observed data at different sites. Hosking and Wallis (1997) thus advocate the use of Monte Carlo simulation procedures to estimate the accuracy of the quantiles in a regional frequency analysis.

The procedure outlined by Hosking and Wallis (1997) and described in Section 2.2.3.5 was adopted. For each site in each cluster a random sample is generated, which has the same record length as the observed data, using the selected frequency distribution at each site with population equal to the observed data. Thus, for each cluster a region was simulated having the same number of stations, record lengths and regional average L-moment ratios as the observed data. This procedure was repeated 100 times, to give 100 simulated regions. The simulations assumed the regions to be homogeneous with a GEV frequency distribution and routines provided by Hosking (1991b) were used to implement the procedure. For each of the 100 repetitions, the errors in the simulated quantiles were calculated and then accumulated and averaged to estimate the bias and RMSE of the quantiles estimated from the actual data. Thus, the 90 % confidence interval can be constructed by selecting the 5th and 95th percentiles from the 100 ranked errors between the simulated region and actual data. For example, the 90% confidence interval for the regional growth curve for Cluster 3 is given in Table 41 and shown in Figure 41.

Table 41 Accuracy measures for estimated growth curve for Cluster 3

Duration (h)	Return Period (Years)	Growth Curve	RMSE	90 % Confidence Interval	
				Upper	Lower
1	2	0.949	0.045	0.923	0.975
	5	1.288	0.044	1.233	1.320
	10	1.502	0.064	1.402	1.565
	20	1.699	0.094	1.538	1.818
	50	1.943	0.140	1.697	2.112
	100	2.118	0.179	1.803	2.347
24	2	0.889	0.096	0.832	0.921
	5	1.260	0.069	1.178	1.250
	10	1.549	0.088	1.402	1.589
	20	1.862	0.146	1.619	1.989
	50	2.329	0.251	1.856	2.647
	100	2.731	0.341	2.053	3.222

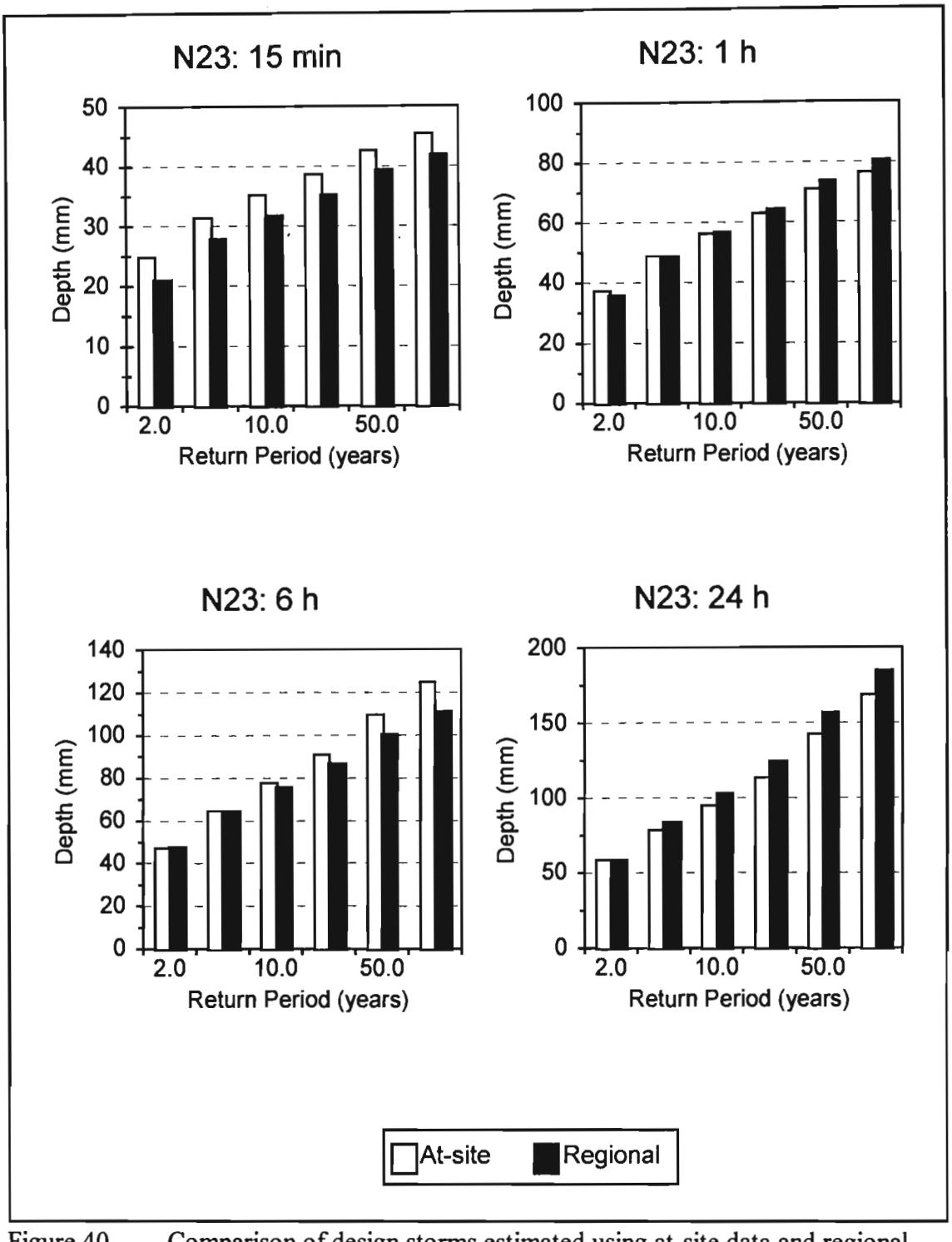


Figure 40 Comparison of design storms estimated using at-site data and regional analysis: N23

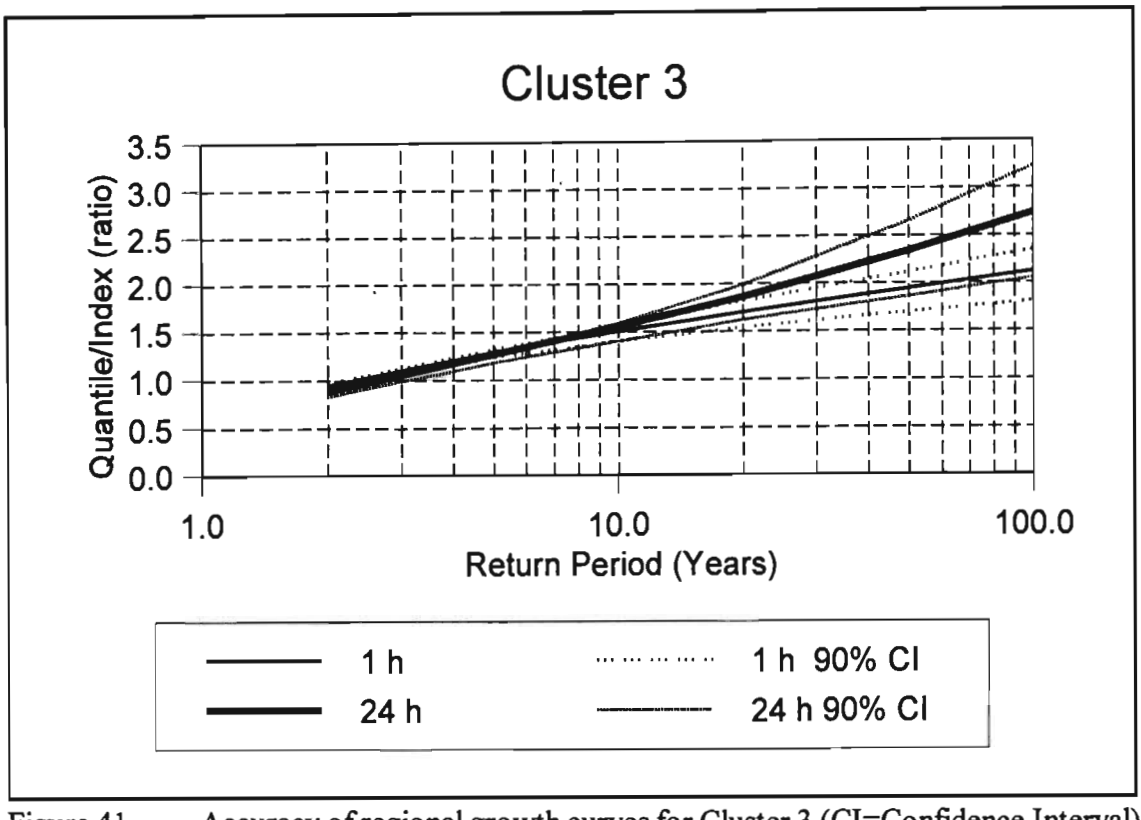


Figure 41 Accuracy of regional growth curves for Cluster 3 (CI=Confidence Interval)

For each of the fifteen relatively homogeneous clusters in South Africa, and for 16 durations ranging from 5 min to 24 h, growth curves were developed which relate the ratio, of the design rainfall and an index value, to return period. The index value used for each duration was the mean of the AMS (L_1) for that duration. Hence quantiles for a particular site can be estimated from the regional growth curve and the index (L_1) value for that site. The accuracy of the quantiles for a particular site can be evaluated using the confidence intervals for the regional growth curve. For example, the 90% confidence interval for the estimated design storms at Ntabamhlope (N11), which is located in Cluster 13, are shown in Figure 42.

In order to estimate the quantiles at a particular site using the regional growth curve, it is necessary to estimate the L_1 value at that site, either from the observed data if that is available, or by some other means if the observed data are not available or are not reliable. In the following section, the results from estimating the 24 h L_1 values using multiple

linear regressions of site characteristics are presented. The methodology used could equally be applied to other durations. However, when investigating scaling relationships in Chapter 6, it is necessary to estimate the 24 h L_1 value and hence only the results for this duration are presented.

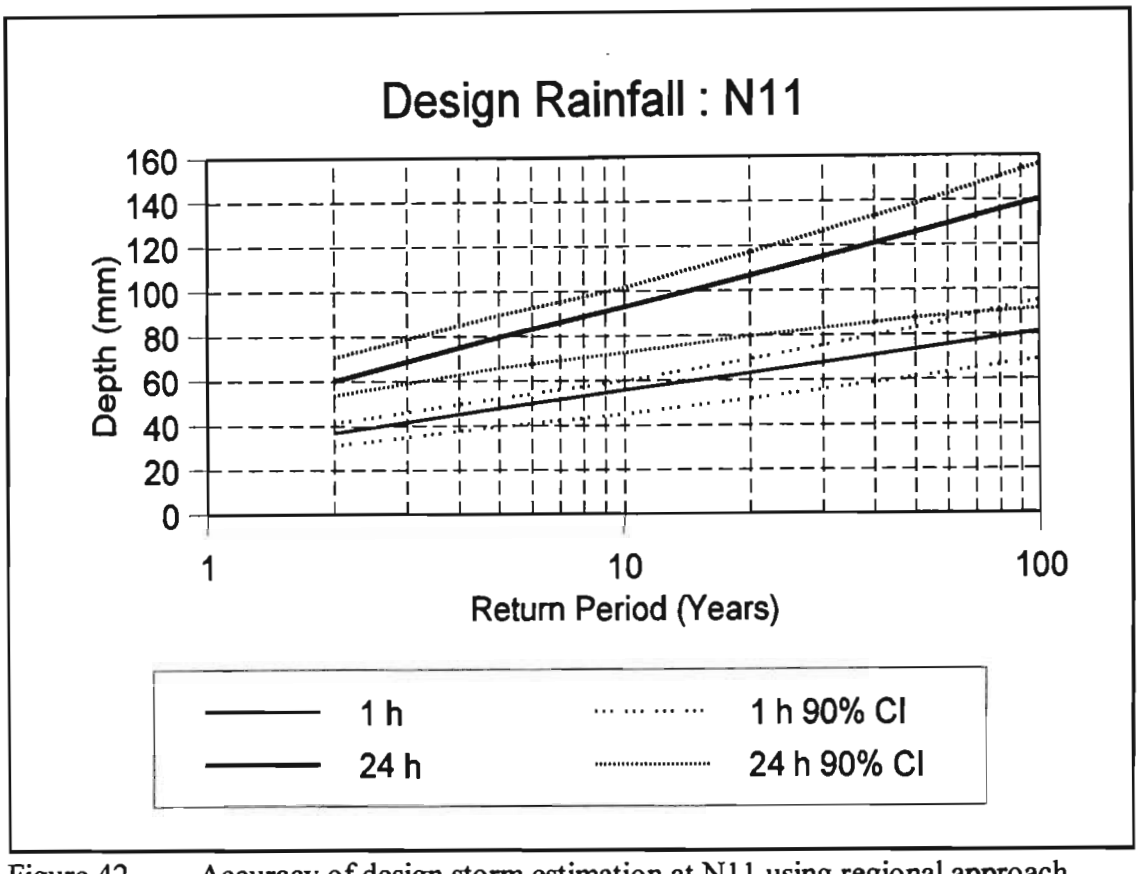


Figure 42 Accuracy of design storm estimation at N11 using regional approach

5.4 ESTIMATION OF THE 24 HOUR INDEX STORM

In order to estimate design storms at ungauged sites, or at sites where the data are unreliable, it is necessary to estimate the index storm used to develop the regional growth curve and thus dimensionalise the curve. For the 24 h duration storm, the index storm used was the mean of the 24 h annual maxima (L_1). Multiple linear relationships were sought, using SAS statistical software, between L_1 and the site characteristics used to establish the

homogeneous regions. It was found that the stepwise method of model selection generally gave lower Predicted Residual Sum of Square (PRESS) values than the methods which optimised the R^2 value. Since the best estimate of the L_1 value was required, the stepwise method of model selection which resulted in the lowest PRESS values was adopted. The significance level for entry of variables into the model was set very low (0.9) and similarly the significance level for keeping a variable in the model was also relaxed to 0.4, thus keeping variables in the model to reduce the PRESS values and improve the estimates of L_1. The coefficients in the linear regression model shown in Equation 59, correlation coefficient and scatter plot around a line of perfect agreement of the data are presented in Table 42.

$$\hat{L\_1} = \left(\sum_{i=1}^{m} Var_i \times Cof_i \right) + Cons \qquad ...59$$

where

$\hat{L\_1}$ = first L-moment,

m = number of variables (7),1=Latitude, 2=Longitude, 3=MAP, 4=Altitude, 5=Seasonality, 6=Precipitation Concentration (Ppt. Conc.), 7=Distance to Sea (Dist. Sea),

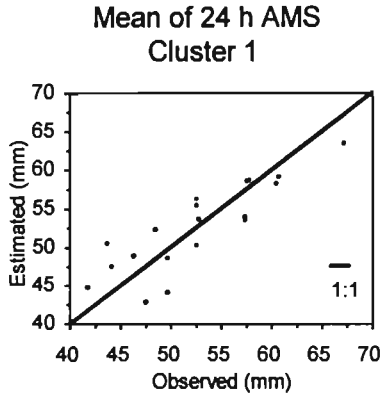
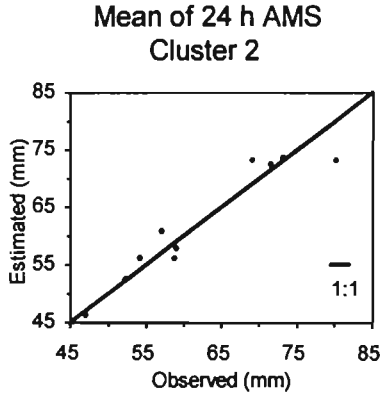
Var_i = i-th variable,

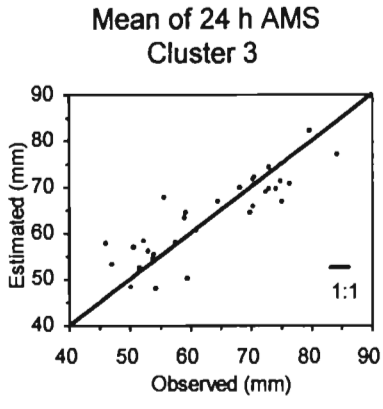
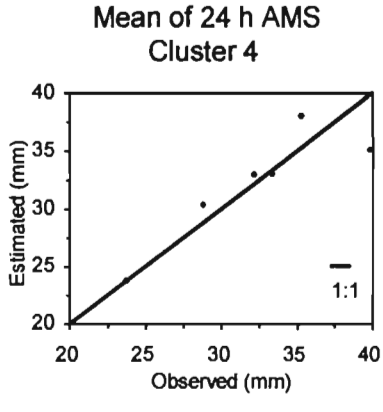
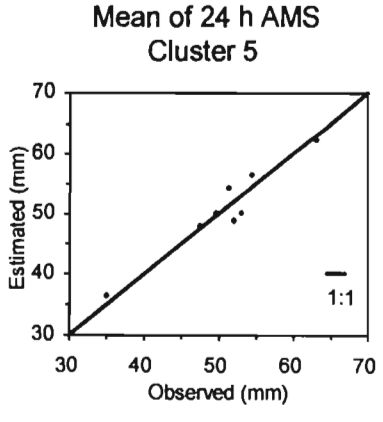
Cof_i = coefficient for i-th variable, and

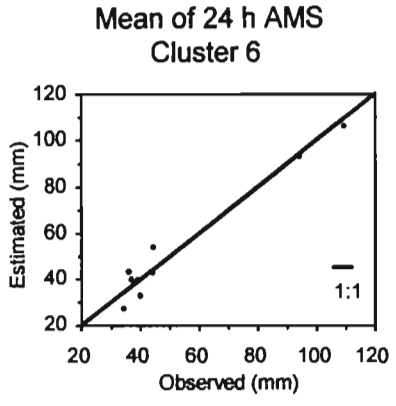
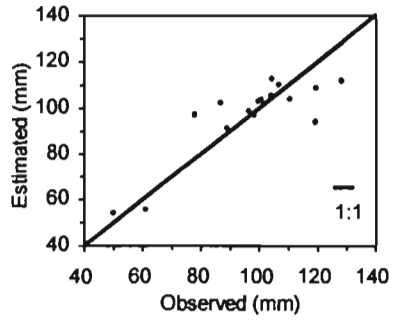
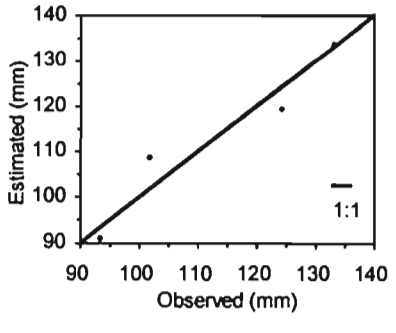
$Cons$ = constant.

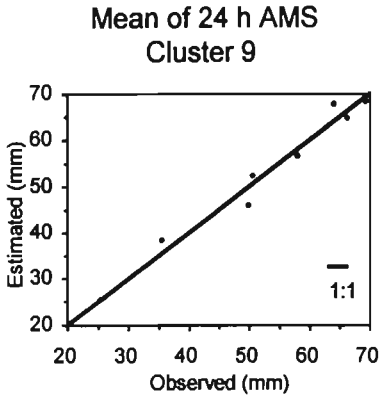
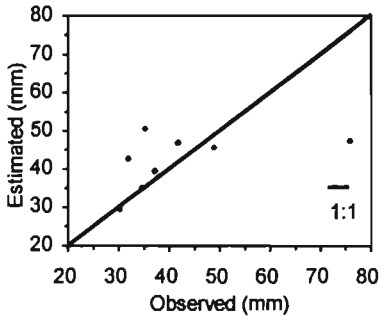
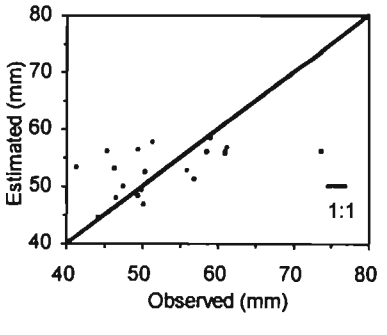
It is conceded that the validity of the regression equations may be affected by dependencies between the selected "independent" variables. However, the choice of independent variables was based on the variables that were successfully used in the cluster analysis of site characteristics. The limited number of short duration rainfall stations resulted in fewer degrees of freedom than the number of independent variables in some clusters. Hence, the results from these clusters (4, 8 and 9) should be used only with extreme caution and the success of the methodology should be judged from the results at the remaining sites.

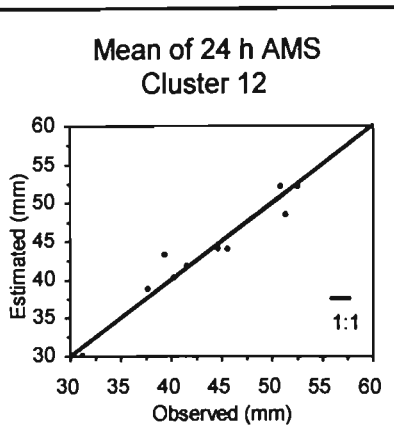
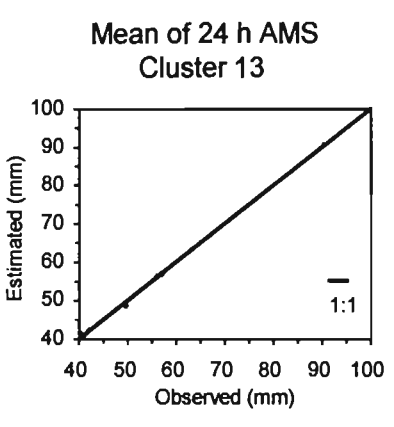
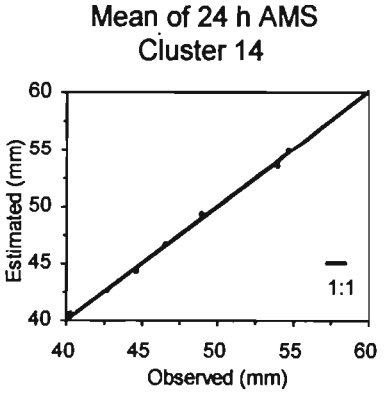
Table 42 Regression analysis of 24 h annual maxima (L_1) as a function of site characteristics and region

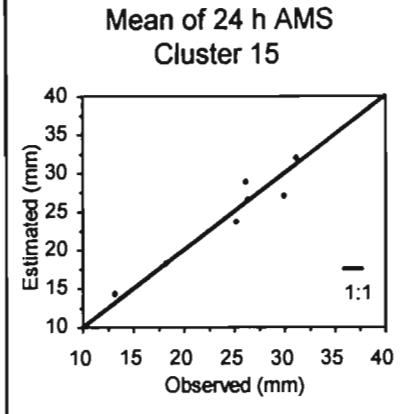
Cluster (No. of stations)	Regression Coefficients		R^2	Scatter Plot
	Variable	Value		
1 (19)	Constant	-121.33139978	0.73	
	Latitude (°)	-4.32194141		
	Longitude (°)			
	Altitude (m)	-0.01709296		
	MAP (mm)	0.09016661		
	Seasonality (-)	-2.71852305		
	Ppt. Concentration (%)	0.62619464		
	Dist. from Sea (m)			
2 (10)	Constant	177.00086849	0.91	
	Latitude (°)	5.38817351		
	Longitude (°)			
	Altitude (m)	-0.01992853		
	MAP (mm)	0.03934102		
	Seasonality (-)			
	Ppt. Concentration (%)			
	Dist. from Sea (m)			

Cluster (No. of stations)	Regression Coefficients		R^2	Scatter Plot
	Variable	Value		
3 (32)	Constant	-1092.41031260	0.77	
	Latitude (°)			
	Longitude (°)	32.64016742		
	Altitude (m)	0.04122272		
	MAP (mm)			
	Seasonality (-)	39.80853547		
	Ppt. Concentration (%)	-0.73429309		
	Dist. from Sea (m)	0.00005997		
4 (6)	Constant	12.88186896	0.78	
	Latitude (°)			
	Longitude (°)			
	Altitude (m)			
	MAP (mm)			
	Seasonality (-)			
	Ppt. Concentration (%)			
	Dist. from Sea (m)	0.00004616		
5 (9)	Constant	801.61697120	0.96	
	Latitude (°)	10.86019307		
	Longitude (°)			
	Altitude (m)	-0.01864936		
	MAP (mm)	-0.03278059		
	Seasonality (-)			
	Ppt. Concentration (%)	-7.01056913		
	Dist. from Sea (m)			

Cluster (No. of stations)	Regression Coefficients		R^2	Scatter Plot
	Variable	Value		
6 (9)	Constant Latitude (°) Longitude (°) Altitude (m) MAP (mm) Seasonality (-) Ppt. Concentration (%) Dist. from Sea (m)	-792.30726324 -79.25404727 -84.90316270 0.07792625 -6.44610538	0.96	
7 (16)	Constant Latitude (°) Longitude (°) Altitude (m) MAP (mm) Seasonality (-) Ppt. Concentration (%) Dist. from Sea (m)	18.34826103 0.06675284 0.05697078	0.71	
8 (4)	Constant Latitude (°) Longitude (°) Altitude (m) MAP (mm) Seasonality (-) Ppt. Concentration (%) Dist. from Sea (m)	-26798.59576036 843.67627112 0.78885764	0.93	

Cluster (No. of stations)	Regression Coefficients		R^2	Scatter Plot
	Variable	Value		
9 (8)	Constant Latitude (°) Longitude (°) Altitude (m) MAP (mm) Seasonality (-) Ppt. Concentration (%) Dist. from Sea (m)	629.14362760 15.55315626 -0.22293808 10.35689960 -0.00118391	0.97	
10 (8)	Constant Latitude (°) Longitude (°) Altitude (m) MAP (mm) Seasonality (-) Ppt. Concentration (%) Dist. from Sea (m)	225.70304539 -8.38218559 0.05897167	0.23	
11 (19)	Constant Latitude (°) Longitude (°) Altitude (m) MAP (mm) Seasonality (-) Ppt. Concentration (%) Dist. from Sea (m)	150.41255017 3.14207324 -2.61165855 0.03569725 11.12382858	0.27	

Cluster (No. of stations)	Regression Coefficients		R^2	Scatter Plot
	Variable	Value		
12 (10)	Constant	-461.88956151	0.93	
	Latitude (°)	-12.49114058		
	Longitude (°)	-4.06683470		
	Altitude (m)			
	MAP (mm)	0.13835642		
	Seasonality (-)	-3.59786280		
	Ppt. Concentration (%)	3.85641434		
	Dist. from Sea (m)			
13 (7)	Constant	496.26529036	1.00	
	Latitude (°)	10.50701317		
	Longitude (°)			
	Altitude (m)			
	MAP (mm)			
	Seasonality (-)	13.73595647		
	Ppt. Concentration (%)	-4.16359992		
	Dist. from Sea (m)	0.00005234		
14 (7)	Constant	-19.11471592	1.00	
	Latitude (°)	5.78368758		
	Longitude (°)	6.95978145		
	Altitude (m)			
	MAP (mm)	0.06912302		
	Seasonality (-)	5.66589950		
	Ppt. Concentration (%)			
	Dist. from Sea (m)			

Cluster (No. of stations)	Regression Coefficients		R^2	Scatter Plot
	Variable	Value		
15 (7)	Constant Latitude (°) Longitude (°) Altitude (m) MAP (mm) Seasonality (-) Ppt. Concentration (%) Dist. from Sea (m)	-21.15989495 -1.23354254 0.00901883	0.92	

With the exception of Clusters 10 and 11, the mean 24 h annual maximum rainfall event were predicted adequately and hence the regressions can be used at ungauged sites, or at sites which have unreliable data, to dimensionalise the regional growth curve and thus to estimate the design values at these sites. Further subdivision or relocation of stations in Clusters 10 and 11 did not improve the regressions. Hence it is recommended that caution should be exercised when applying the RLMA at ungauged sites in Clusters 10 and 11.

The RLMA has been successfully applied and hence it is reasonable, with the exceptions of Cluster 10 and 11, to estimate design rainfalls for 24 h durations at ungauged sites. Similar multiple linear regression analysis could be performed to estimate the *L_1* for each duration < 24 h as a function of site characteristics, and thus enable the estimation of design values using the regional growth curve for that particular duration. Alternatively, the index value used in the estimation of the regional growth curve for durations < 24 h could be replaced by the 24 h *L_1* value, which could be estimated at ungauged sites using the results presented in Table 42. Thus, instead of developing regressions to estimate the *L_1* value for each individual duration, the regional growth curves could be estimated using only the 24 h *L_1* as index values, which could be estimated using the results presented in Table 42.

Using the RLMA, no fixed boundaries exist between adjacent clusters. Therefore at an ungauged location, it is necessary to estimate the Euclidean distance between the site characteristics of the ungauged location and the mean of the site characteristics of all sites within each cluster. The ungauged site is then assigned to the cluster which has the closest Euclidean distance to the ungauged site. This gives an estimation of the regional growth curve at that site. Hence, in order to estimate design values at the ungauged site, it is only necessary to estimate the index value at that site, as has been performed for the 24 h duration.

The assumption in the application of the RLMA is that within each relatively homogeneous cluster, a single probability distribution is applicable to all sites after scaling using an at-site index value. Hence it is necessary to investigate which probability distribution to adopt for the estimation of design rainfalls in each of the clusters.

5.5 CHOICE OF FREQUENCY DISTRIBUTION

One option was to determine the most appropriate probability distribution for each duration in each of the 15 relatively homogeneous clusters. However, from a practical point of view it was decided to determine, for a selected duration, an appropriate distribution which is applicable to all clusters and which is then assumed to apply to all durations. This approach of a single appropriate distribution for all clusters is supported by Wallis (1997). The assumption that an appropriate distribution for a selected duration is applicable to other durations at the same site agrees with the property of scale invariance noted by, *inter alia,* Gupta and Waymire (1990) and Burlando and Rosso (1996), which implies that the probability distributions of rainfall depth is the same at different time scales. The selection of the most appropriate distribution was conducted on the 24 h digitised data. However, it is conceded that possibly more reliable results at many more sites would be obtained from the use of daily rainfall totals recorded by standard non-recording raingauges. Thus, these results may need to be revised after the same analysis has been performed on the daily data.

Hosking and Wallis (1997) developed a Goodness-of-Fit (GOF) criterion, described in Section 2.2.2.3, which is based on L-moment ratios to determine suitable probability distributions to use in a regional frequency analysis. In addition to non-parametric tests, Smithers (1996) also used L-moment statistics as well parametric tests such as the Chi-squared test and deviations from a plotting position. All of these techniques are used in the following section to determine suitable probability distributions for use in South Africa. All tests are performed using the 24 h duration events from the digitised rainfall data.

5.5.1 At-site Parametric Statistics

In order to determine the most appropriate probability distribution to use at all the clusters, Chi-squared and standardised deviations parametric tests were performed.

5.5.1.1 Chi-squared test

A chi-squared test was employed which utilises 10 equally spaced probability class intervals and either rejects or accepts the null hypothesis that the sample of data could have been drawn from the distribution being evaluated (Kite, 1988). In this study the LN2, 3 parameter log-normal (LN3), LP3, Pearson type 3 (PE3), Gumbel (EV1), log-EV1 (L-EV1), General Extreme Value (GEV), generalised Pareto (GPA), generalised logistic (GLO) and Wakeby (WAK) probability distributions were employed. The probability density functions and, where possible, the cumulative density functions for these distributions are defined in Appendix B. The results from the Chi-squared tests performed for the 24 h duration event and for the 15 relatively homogeneous clusters are contained in Table 43.

The results in Table 43 indicate that the GEV, GLO, EV1 and LN3 probability distributions were accepted most frequently as suitable distributions in all clusters.

Table 43 Number of rejections of the null hypothesis that the 24 h AMS could have been drawn from a parent distribution, at the 95% confidence level, with results expressed as a percentage of total number of sites in each cluster

Cluster Number	Probability Distribution									
	LN2	LN3	LP3	L-EV1	EV1	GEV	PE3	GLO	GPA	WAK
1	32	10	26	21	5	15	10	16	32	15
2	10	0	20	20	0	0	0	0	0	10
3	9	22	13	3	21	6	25	3	25	19
4	33	17	50	83	17	17	17	0	17	33
5	22	11	22	11	0	11	11	11	22	22
6	0	0	11	0	0	0	0	0	11	11
7	19	13	25	13	13	6	19	13	13	13
8	25	0	25	50	0	0	0	0	0	0
9	25	13	13	0	13	13	13	25	38	38
10	0	13	0	13	0	13	0	0	25	13
11	5	0	0	10	10	0	0	0	5	15
12	10	0	30	10	0	0	10	10	30	0
13	0	0	14	42	0	0	0	14	14	14
14	14	0	14	28	14	0	14	0	14	14
15	0	28	0	0	28	28	14	28	14	28
Sum	206	130	266	305	122	109	136	120	260	245

5.5.1.2 Standardised deviations

The Standardised Deviation (*SD*) GOF method adopted is similar to techniques used by Benson (1968), Bobee and Robitaille (1977) and Kite (1988). The *SD* is computed as shown in Equation 60. Return periods of 2, 5, 10, 20, 50 and 100 years, which correspond to non-exceedance probabilities of 0.50, 0.80, 0.90, 0.95, 0.98 and 0.99 respectively, were used in the calculation of the *SD*. The choice of plotting position equation was shown by the NERC (1975) and Smithers (1994) to affect the computed *SD*, although Kite (1988) expressed the opinion that the relative rankings of distributions would not be influenced by the choice of plotting position.

$$SD_j = \frac{1}{df}\sum_{i=1}^{k}\frac{\left(y_i - x_i\right)^2}{y_i} \qquad ...60$$

where

SD_j = standardised deviation of j-th candidate distribution,

y_i = recorded data, interpolated (if necessary) but not extrapolated to correspond to the i-th return period, with probabilities assigned to observed data using a plotting position equation,

x_i = event magnitude computed from the j-th probability distribution for the i-th return period,

k = maximum number of recurrence intervals (5) used in the computation, and

df = degrees of freedom used to fit the trial distribution.

The Weibull plotting position, as shown in Equation 61, has been shown by means of a survey conducted in different countries by Cunnane (1989), to be the most frequent plotting position used, despite its bias in graphical quantile estimates.

$$P_e = \frac{r}{N+1} \qquad ...61$$

where

P_e = exceedance probability of r-th ranked data,

r = rank of data, and

N = number of points in the data series.

The results from ranking the distributions according to the *SD* statistic are presented in Table 44. The results in Table 44 indicate that suitable probability distributions to use are the PE3, GEV, LP3 and LN3.

Table 44 Relative ranking of 10 probability distributions for 24 h events according to computed *SD* at all 15 clusters (1 = best, 10 = worst), using the Weibull plotting position to assign probabilities to observed data

Cluster Number	Probability Distribution									
	LN2	LN3	LP3	L-EV1	EV1	GEV	PE3	GLO	GPA	WAK
1	8	2	6	10	5	3	1	9	4	7
2	8	6	3	10	5	2	1	9	4	7
3	5	7	1	8	10	3	6	9	2	4
4	9	3	8	10	5	1	2	6	7	4
5	1	8	3	9	1	6	4	10	5	7
6	7	6	2	9	8	5	3	10	1	4
7	4	6	1	3	7	8	1	10	5	9
8	9	2	7	10	4	3	1	8	5	6
9	2	5	1	9	6	7	3	10	4	8
10	4	3	1	7	5	1	6	9	8	10
11	6	3	2	10	5	4	1	9	8	7
12	7	2	5	10	9	3	1	8	4	6
13	1	4	8	10	2	3	5	6	9	7
14	6	4	3	7	9	5	2	10	1	8
15	4	5	6	10	1	2	3	9	7	9
Sum	81	66	57	132	82	56	40	132	74	103

5.5.2 At-site Non-parametric Tests

A non-parametric test was performed to evaluate the ability of the different probability distributions to provide estimates of the 100 year return period event. Similar tests have been performed on flood flow data in the USA by Vogel *et al.* (1993b) and in Australia by Vogel *et al.* (1993a). The test uses a "station year" approach and assumes that the AMS from the sites within a cluster are independent and the extreme events occur independently from year to year. Thus it may be assumed that the number of exceedances follow a binomial distribution (Vogel *et al.*, 1993a). The test comprises of counting, for each distribution and at each cluster, the number of times (X) an observed value exceeds the

estimated T year return period event. Assuming X follows a binomial distribution, the mean of m site-years within each cluster is $E[X] = mP_e$ and variance $Var[X] = mP_e(1-P_e)$, where $P_e = 1/T$. Confidence intervals at the 95% levels may be computed as

$$0.95 = \sum_{x=x_{0.025}}^{x_{0.975}} \binom{m}{x} P_e^{\,x}(1-P_e)^{m-x} \qquad \ldots 62$$

A 95% confidence interval was computed as shown in Equation 63 for the expected number of exceedances using the normal approximation of the binomial distribution, as described by Steel and Torrie (1980).

$$X_{0.025}^{0.975} = \left(P_e \pm Z_{0.05} \times \sqrt{\frac{P_e \times (1-P_e)}{m}} \right) \times m \qquad \ldots 63$$

where

$Z_{0.05}$ = 5% exceedance value of Normal distribution with μ=0 and σ=1.

Results from the tests based on the above assumptions are contained in Table 45 and indicate that the LN2, EV1 and PE3 distributions were the only distributions which did not exceed the 95% confidence interval in all the clusters. No expected probability adjustment was used in generating the results in Table 45. This non-parametric test's assumptions (independence) may be compromised by the relatively close locality of the sites to each other within each cluster.

Table 45 Number of data values in the AMS that exceed the 100 year return period event, as estimated by different probability distributions, fitted to the data using L-moments (* indicates results falling outside the 95 % confidence interval)

Cluster	Station Years (95% Confidence Level)	Probability Distribution									
		LN2	LN3	LP3	L-EV1	EV1	GEV	GPA	PE3	GLO	WAK
1	444 (3-15)	5	4	15	0*	4	4	12	6	2*	1*
2	195 (0-8)	1	0	9*	0	2	0	5	0	0	0
3	574 (5-18)	13	4*	10	4*	12	4*	7	9	6	9
4	150 (0-6)	0	0	10*	0	0	0	7*	0	0	1
5	201 (0-8)	0	0	5	0	0	0	5	0	0	0
6	228 (0-9)	1	0	1	0	1	0	5	0	0	0
7	190 (0-8)	1	0	3	0	1	0	0	0	0	0
8	93 (-1-5)	0	0	3	0	0	0	1	0	0	0
9	201 (0-8)	0	0	1	0	1	0	2	0	0	0
10	178 (0-7)	1	1	4	0	2	1	3	2	0	1
11	402 (3-14)	4	3	13	0*	4	3	11	5	1	2
12	227 (1-10)	4	2	6	0*	4	2	7	2	2	2
13	178 (0-7)	1	2	3	0	2	1	5	2	1	2
14	195 (0-8)	1	1	4	0	1	1	1	1	0	0
15	198 (0-8)	1	1	2	0	2	0	6	1	0	0

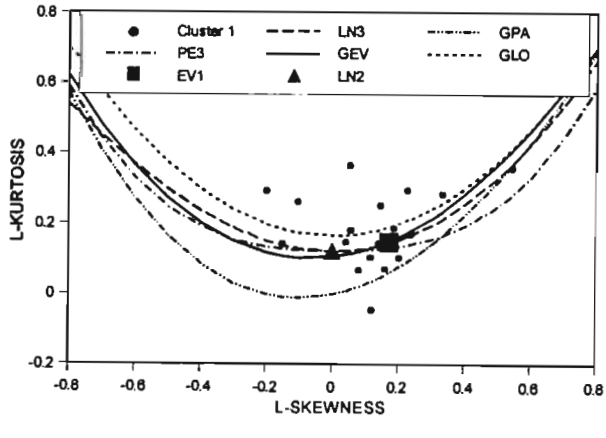
5.5.3 Statistics Based on Regional Average L-moment Ratios

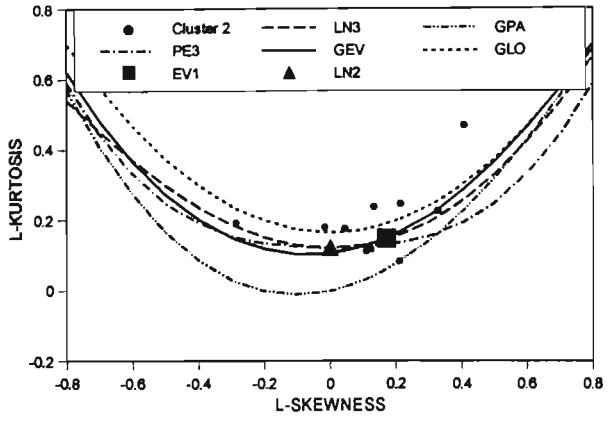
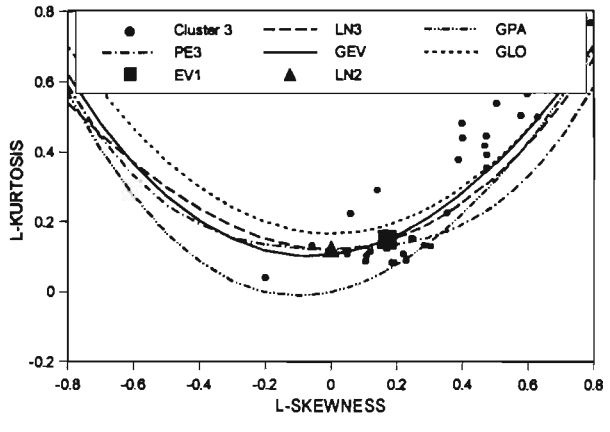
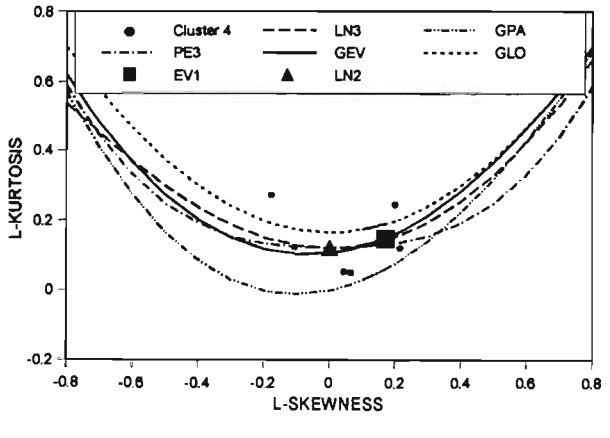
The choice of a regional distribution using L-moment ratios is based on fitting an assumed distribution to the regional record length weighted L-moment ratios (Hosking and Wallis, 1997). Thus the fitted distribution will have the same L-CV as the regional average values and the quality of fit is judged by the difference between the L-kurtosis of the fitted distribution (t_4^{PD}) and the regional average (t_4^R). The sampling variability (σ_4) is obtained by repeated simulations of a homogeneous region, having the fitted distribution, with the same number of sites and record lengths as the observed data. In practice, Hosking and Wallis

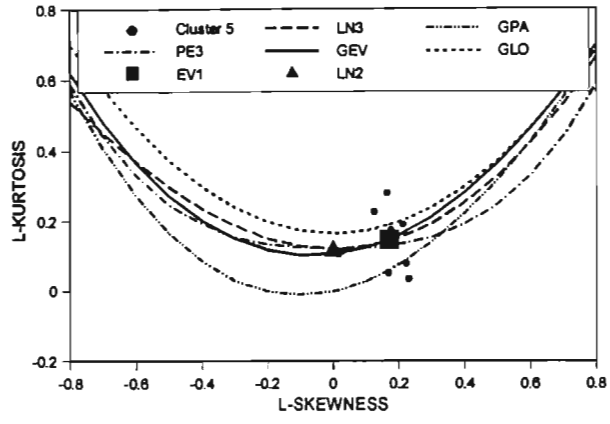
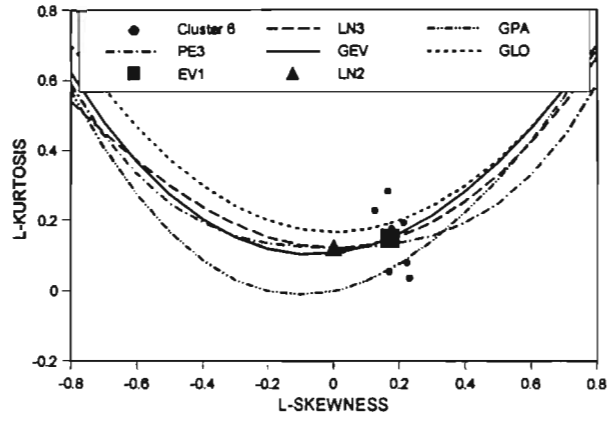
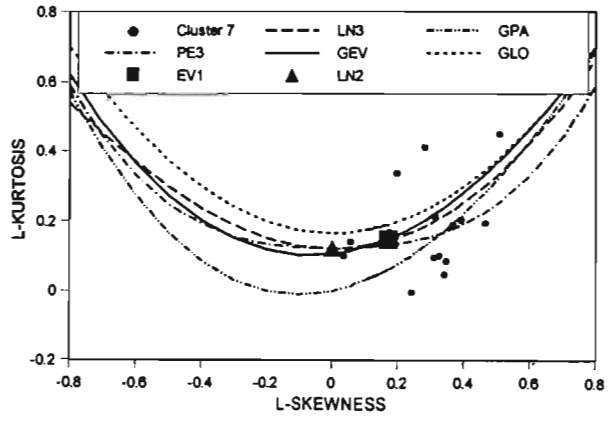
(1997) assume that reasonable estimates of the sampling distribution can be obtained by using the flexible 4-parameter Kappa distribution, instead of repeated simulations with different candidate distributions. The statistic Z is computed as shown in Equation 64. Values of $|Z| \leq 1.64$ are deemed to indicate that the fit of the assumed distribution is adequate. A formal definition of the statistic is presented in Section 2.2.3.3. The results of the analysis and associated L-moment diagrams are contained in Table 46.

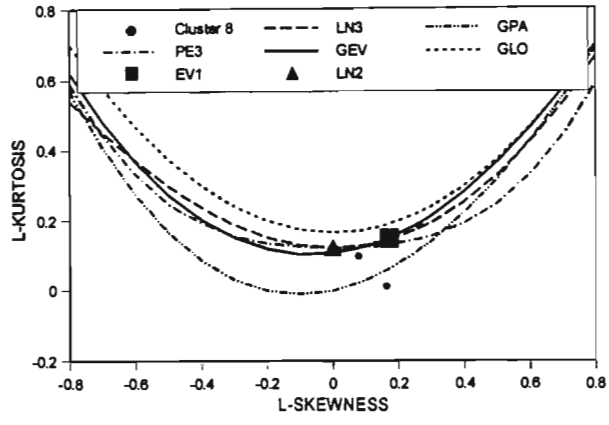
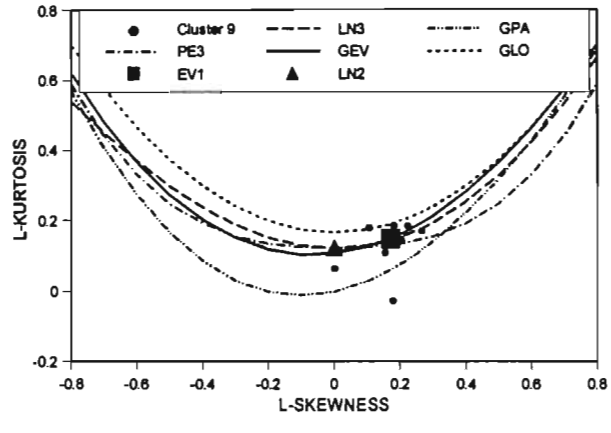
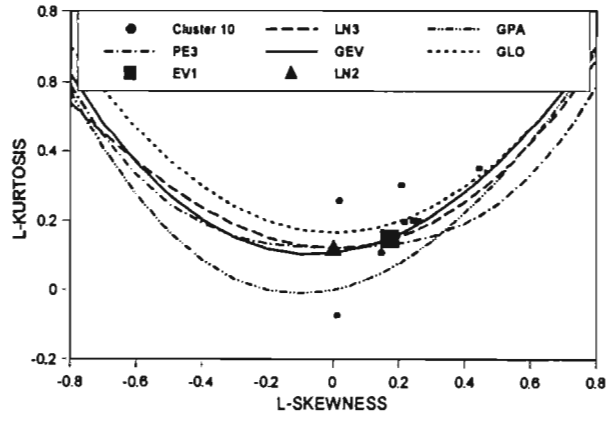
$$Z = \frac{\left(t_4^R - t_4^{PD}\right)}{\sigma_4} \qquad \ldots 64$$

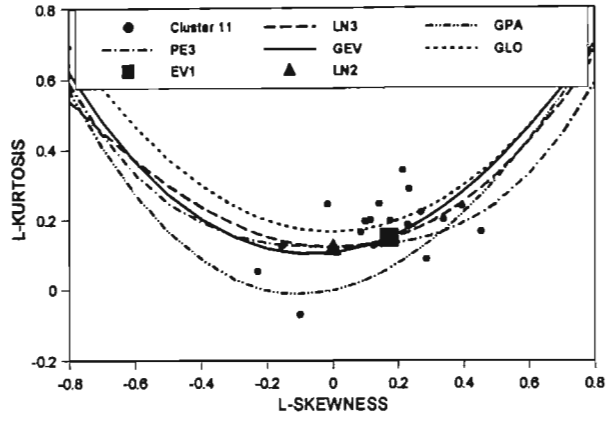
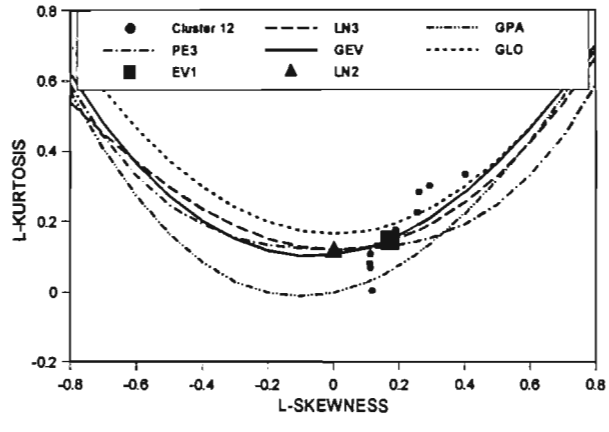
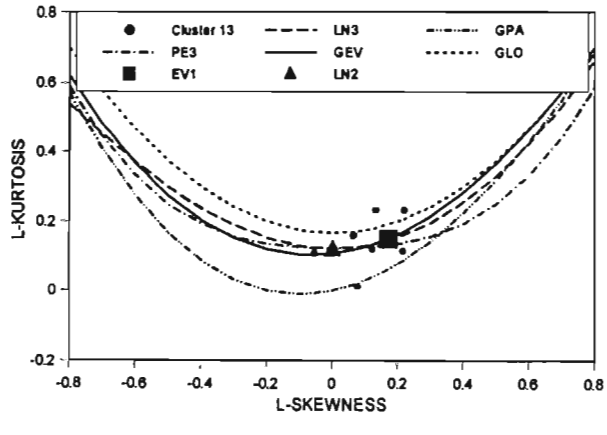
Table 46 Acceptable probability distributions, Z-test statistic and L-moment ratio diagrams for 15 relatively homogeneous clusters in South Africa

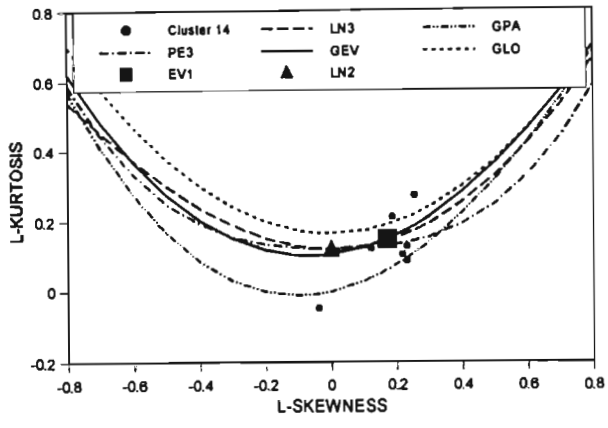
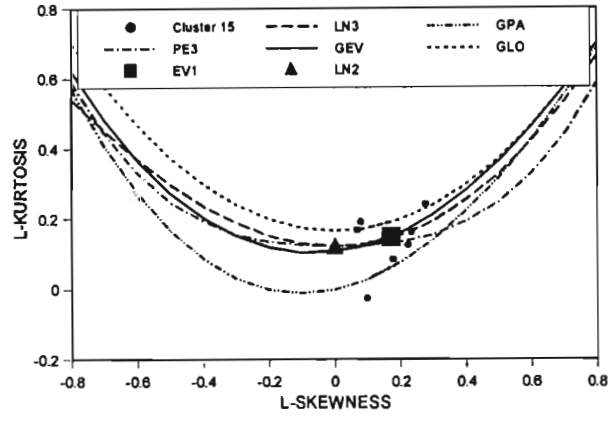
Cluster Number	Acceptable Distributions	Z	L-Moment Diagram
1	GLO	-0.65	

Cluster Number	Acceptable Distributions	Z	L-Moment Diagram
2	GLO GEV LN3	-0.22 -1.58 -1.60	
3	-		
4	GLO GEV LN3 PE3	1.14 -0.55 -0.33 -0.45	

Cluster Number	Acceptable Distributions	Z	L-Moment Diagram
5	GLO GEV LN3 PE3	1.03 -0.35 -0.43 -0.79	
6	GLO GEV LN3 PE3	1.64 0.12 -0.13 -0.74	
7	GLO GEV LN3 PE3 GPA	0.88 0.21 -0.31 -1.19 -1.61	

Cluster Number	Acceptable Distributions	Z	L-Moment Diagram
8	GEV LN3 PE3 GPA	1.03 1.00 0.74 -1.49	
9	GEV LN3 PE3	0.32 0.07 -0.50	
10	GLO GEV	0.04 -0.98	

Cluster Number	Acceptable Distributions	Z	L-Moment Diagram
11	GLO	-1.37	
12	GLO GEV LN3	0.32 -0.99 -1.31	
13	GLO GEV LN3	1.09 -0.48 -0.43	

Cluster Number	Acceptable Distributions	Z	L-Moment Diagram
14	GEV LN3 PE3	0.37 0.10 -0.49	
15	GLO GEV LN3 PE3	1.18 -0.17 -0.36 -0.84	

Hosking and Wallis (1997) recommend that in regions where no distribution is suitable (e.g. Cluster number 3), the Kappa or Wakeby distribution should be used, as they are "robust to the mis-specification of the form of the frequency distribution in a regional frequency analysis". The number of homogeneous regions in which the candidate distributions gave an acceptable fit to the 24 h AMS are listed in Table 47.

Table 47 Number of homogeneous regions in which candidate distributions gave an acceptable fit to the 24 h AMS

Number of homogeneous regions in which the distribution gave an acceptable fit to the 24 h AMS				
GLO	GEV	LN3	PE3	GPA
11	12	11	7	2

The results contained in Table 47 indicate that, if a single probability distribution was to be adopted for the all regions according to the regional L-moment ratios test, the GEV would be the most appropriate distribution.

5.5.4 Concluding Remarks on Choice of Frequency Distribution

There is generally good agreement for most clusters between the probability distributions deemed to be most suitable by the Chi-squared test (GEV, GLO. EV1, LN3) and the regional L-moment ratios test (GEV, GLO, LN3). However, the *SD* test indicated that the most appropriate distributions were the PE3, GEV and LP3 distributions while the non-parametric exceedance test selected the PE3, EV1 and LN2 distributions. It is thus recommended that, if a single distribution were to be adopted for all regions, the GEV distribution would be the most appropriate probability distribution to use. A similar conclusion for South Africa was made by Smithers (1996) using data from individual sites and employing both parametric and non-parametric tests, but not regional tests based on L-moment ratios.

5.6 CHAPTER CONCLUSIONS

In this chapter the RLMA, which is described in Chapter 2 and is based on the methodology developed by Hosking and Wallis (1993; 1997), has been applied using data from 172 short

duration rainfall stations in South Africa. The discordancy index developed by Hosking and Wallis (1993; 1997) was found to identify erroneous or inconsistent data and was used to screen all the data used.

Regionalisation based only on site characteristics resulted in, after a few subjective relocations of stations, in 15 relatively homogeneous clusters in South Africa. The cluster analysis was found to be sensitive to the scaling of the site characteristics, and the best results were obtained when the scales of all the site characteristics were within the same range (0,100).

For each cluster and duration, the mean of AMS (L_1) for each duration was used as the index value when estimating regional growth curves which relate the ratio, of the design and index values, to return period. Hence, with the regional growth curve and the index value for a particular site, design rainfalls may be estimated for the site. The index value (L_1) may be estimated from reliable observed data, if available, or at ungauged sites by means of multiple linear regression relationships of site characteristics.

The accuracy of the regional growth curves was assessed at one "hidden" site (N23) in Cluster 3, which was not used in the regionalisation or the estimation of the regional growth curve, and also by means of Monte Carlo type simulations at all the clusters. Thus confidence or error bands were estimated for the regional growth curves and these were translated into confidence limits of design rainfalls at selected sites.

A number of tests were employed to determine the most appropriate probability distribution to use at all clusters. The GEV was found to be an acceptable distribution by most tests and at most clusters and hence is recommended for general use in South Africa. This finding was based only on 24 h duration rainfall events and it is hypothesised that it will apply to shorter durations as well. This assumes that the probability distributions of rainfall depth are the same at different time scales, i.e. the property of scale invariance noted by, *inter alia,* Gupta and Waymire (1990) and Burlando and Rosso (1996). This approach is also supported by

Wallis (1997). However, it is recommended that further tests be applied for selected durations shorter than 24 h.

The RLMA has been successfully applied in 15 relatively homogeneous clusters in South Africa. Ungauged sites or where the observed data are unreliable may be assigned to the cluster which has the closest Euclidean distance of site characteristics to the ungauged site. Thus the regional growth curve for the cluster is applicable to the site and in this manner design storms can be estimated at any location in South Africa where the index storm can be estimated.

In Chapter 6 the concept of scaling the moments of the AMS with respect to duration is investigated as another approach to estimating short duration design storms using an inadequate database.

CHAPTER 6

SCALING OF DEPTH-DURATION-FREQUENCY RELATIONSHIPS

In an effort to compensate for the low reliability of much of the data contained within the short duration rainfall database, three approaches to estimating design storms from the database were evaluated. The first approach, with results presented in Chapter 5, used a regional frequency analysis. The second approach, with results presented in this chapter, investigated scaling relationships of the moments of the Annual Maximum Series (AMS) and the third approach, with results presented in Chapter 7, used a stochastic intra-daily model to generate synthetic rainfall series. A common theme in all three approaches was the development of techniques to estimate short duration design storms from the daily rainfall totals, as measured by standard, non-recording raingauges for fixed 24 h periods ending at 08:00 each day, and not from the break-point digitised rainfall data where the highest 24 h period of rainfall may not correspond with the 08:00 - 08:00 period.

The scaling concepts used in this chapter were introduced in Section 2.4.3. The assumption was made, based on observations, that storm rainfall is characterised by the property of scale invariance (Gupta and Waymire, 1990), which implies that the probability distributions of rainfall depth is the same at different time scales.

Previous investigations of scaling properties of rainfall (e.g. Gupta and Waymire, 1990; Menabde *et al.*, 1998) have utilised ordinary product moments. In Section 6.1, as an innovation, the scaling properties of extreme rainfall depths are investigated using L-moments. It is shown that L-moments generally scale better with duration, i.e. are essentially linear on a log-log plot of moment vs duration, and scale over a wider range of durations than product moments.

Given the sparsity of recording raingauges in South Africa and low reliability of much of the digitised rainfall data available from the SAWB, six hypotheses are proposed in Section 6.2

and evaluated at selected sites and clusters in Section 6.5. The hypotheses utilise the scaling properties of extreme rainfall and scaled regional average L-moments and, as far as possible, can be estimated from the widely available and generally reliable daily rainfall data which are recorded manually for fixed 24 h intervals at 08:00 every day.

In order to apply some of the hypotheses described in Section 6.2, it is necessary to estimate the slope on a log-log plot of the linear relationship between L-moments and event duration. The estimation of this slope as a function of site characteristics is developed for each cluster in Section 6.3, thus enabling the estimation of the slope at ungauged sites. Similarly, one of the hypotheses requires estimates of the 24 h mean of the AMS (L_1) to be computed from the daily rainfall data. Hence, regional relationships are developed in Section 6.4 to convert the fixed increment L_1 value, computed from the daily rainfall database, into a continuous time value, equivalent to that computed from the digitised rainfall database.

6.1 ADVANTAGES OF SCALING USING L-MOMENTS

In Figure 43 the first and second order product moments and L-moments are presented for stations in different geographic and climatic locations in South Africa. Included in Figure 43 are the linear regressions for the moments estimated using event durations ranging from 15 min to 5760 min (4 days). By definition the first order L-moment (L_1) and conventional moment (mean) are the same and exhibit nearly linear scaling over this range at most sites in Figure 43. However, as evident in Figure 43 for all the stations shown, the second L-moment (L_2) tends to scale more linearly over a wider range of durations than the conventional second order (variance) moment. The advantages of scaling using L-moments are illustrated in Figure 44 where the deviations from linear scaling evident in Figure 43 are quantified as the Mean Absolute Relative Error ($MARE$), computed as shown in Equation 65, for event durations ranging from 15 min to 4 days.

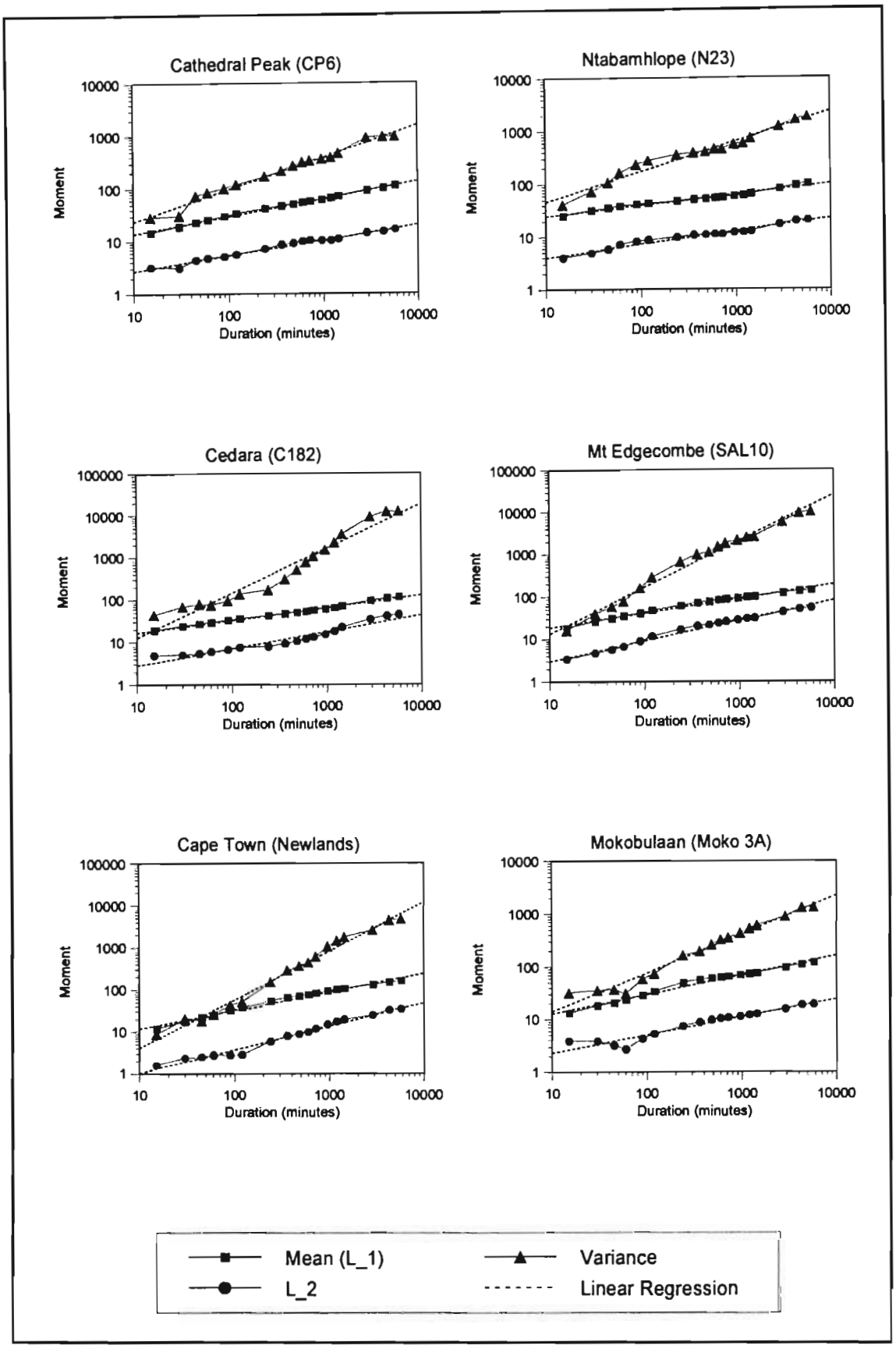


Figure 43 Scaling of conventional product moments and L-moments at selected sites in different climatic and geographic regions in South Africa

$$MARE = \frac{100}{N_D} \times \sum_{k=1}^{N_D} \left(\frac{\left| E_{(k)} - O_{(k)} \right|}{O_{(k)}} \right) \qquad \ldots 65$$

where

$MARE$ = mean absolute relative error (%),

N_D = number of event durations (17),

$E_{(k)}$ = estimated moment using linear regression for k-th duration, and

$O_{(k)}$ = observed moments for k-th duration.

For all stations shown in Figure 44, the *MARE* of the estimated *L_2* values are substantially lower than the *MARE* of the estimated second order product moments (variance), indicating more linear scaling of *L_2*. Thus, further efforts at developing techniques to estimate design storms using scaling principles were focussed on the use of L-moments, although all the methods developed could be applied equally to ordinary product moments.

6.2 DESCRIPTION OF HYPOTHESES

Design rainfall values estimated for specified durations are defined as the rainfall magnitude associated with a specified probability of being equalled or exceeded for the required event duration. The conventional approach to estimating design rainfall values is to use the L-moments, computed directly from the AMS of the observed data for the required duration, to estimate the parameters of an appropriate distribution. Design events for specified exceedance probabilities are then estimated using the fitted distribution. In the light of the low reliability of much of the digitised database, and hence dubious quality of L-moments estimated from the digitised database, other means of estimating the L-moments were hypothesised and evaluated.

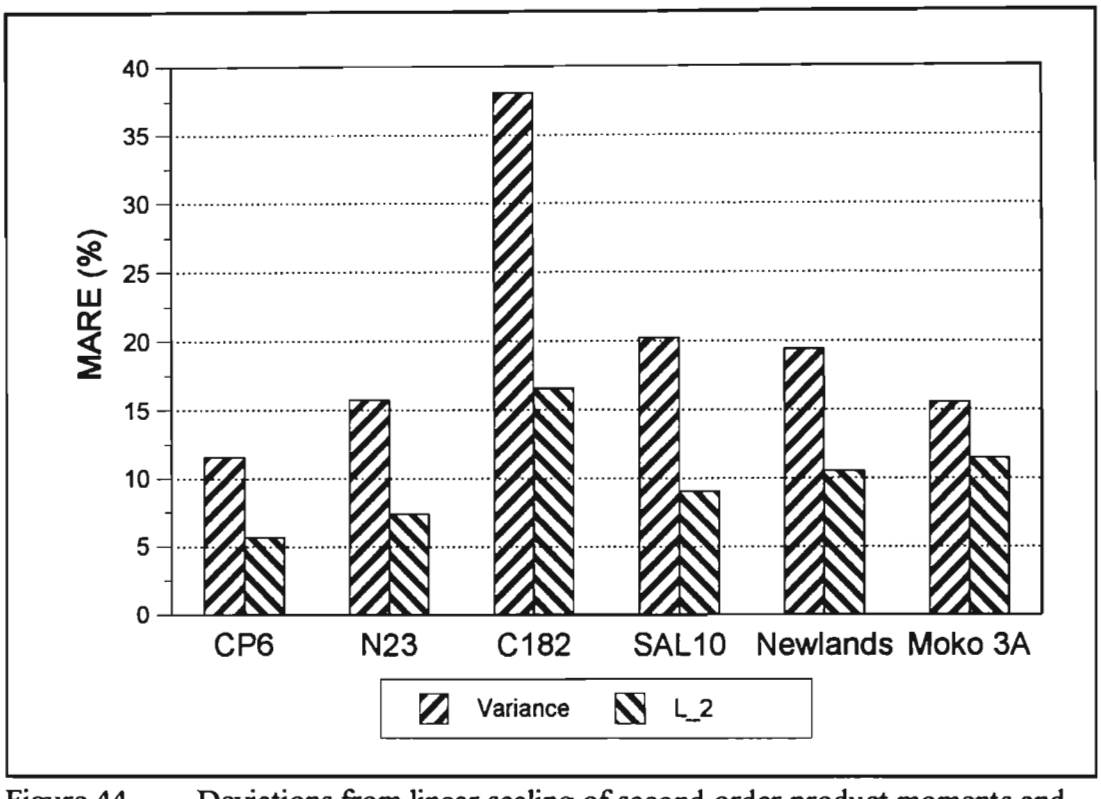


Figure 44 Deviations from linear scaling of second order product moments and L-moments at selected sites in South Africa

In this section, six hypotheses are proposed for estimating L-moments for durations ≤24 h and the hypotheses are evaluated at selected clusters and stations in Section 6.5. The hypotheses are based on the scaling properties of the moments of the AMS and the common distribution of the scaled L-moments of the AMS within a homogeneous region, with the objective of utilising only the daily rainfall data recorded manually at 08:00 every day for the preceding 24 h period.

6.2.1 Hypothesis 1

Hypothesis 1 proposed that the log-transformed L_1 and L_2:duration relationships are linear and that the first and second order L-moments of the AMS for durations < 24 h can be estimated from the 24 h and 48 h values, computed from the digitised data. In order to

estimate the parameters for distributions which have more than two parameters, the moments for orders higher than two are estimated from the mean of the at-site 24 h and 48 h values. As shown in Equation 66 and schematically in Figure 45, the *L_1* and *L_2* values are estimated by linear extrapolation from the 24 h and 48 h values.

$$\log\left(\hat{L\_x}_{(i,D)}\right) = \log(L\_x_{(i,24)}) - \alpha_{(x,i)} \times \left(\log(1440) - \log(D \times 60)\right)$$

$$\alpha_{(x,i)} = \frac{\left(\log(L\_x_{(i,48)}) - \log(L\_x_{(i,24)})\right)}{\left(\log(2880) - \log(1440)\right)} \qquad \ldots 66$$

where

$\hat{L\_x}_{(i,D)}$ = estimated first (x=1) and second (x=2) L-moment at site i for duration D hours and $D \leq 24$,

$\alpha_{(x,i)}$ = slope of log-transformed L-moment vs duration relationship at site i, and

D = duration (h).

Hence, if Hypothesis 1 is valid, the moments for durations < 24 h can be estimated from the 1 and 2 day values computed from the daily rainfall database, after appropriate scaling to account for the differences between extreme events extracted using a fixed ("clock time") and non-fixed ("break-point digitised") time increment.

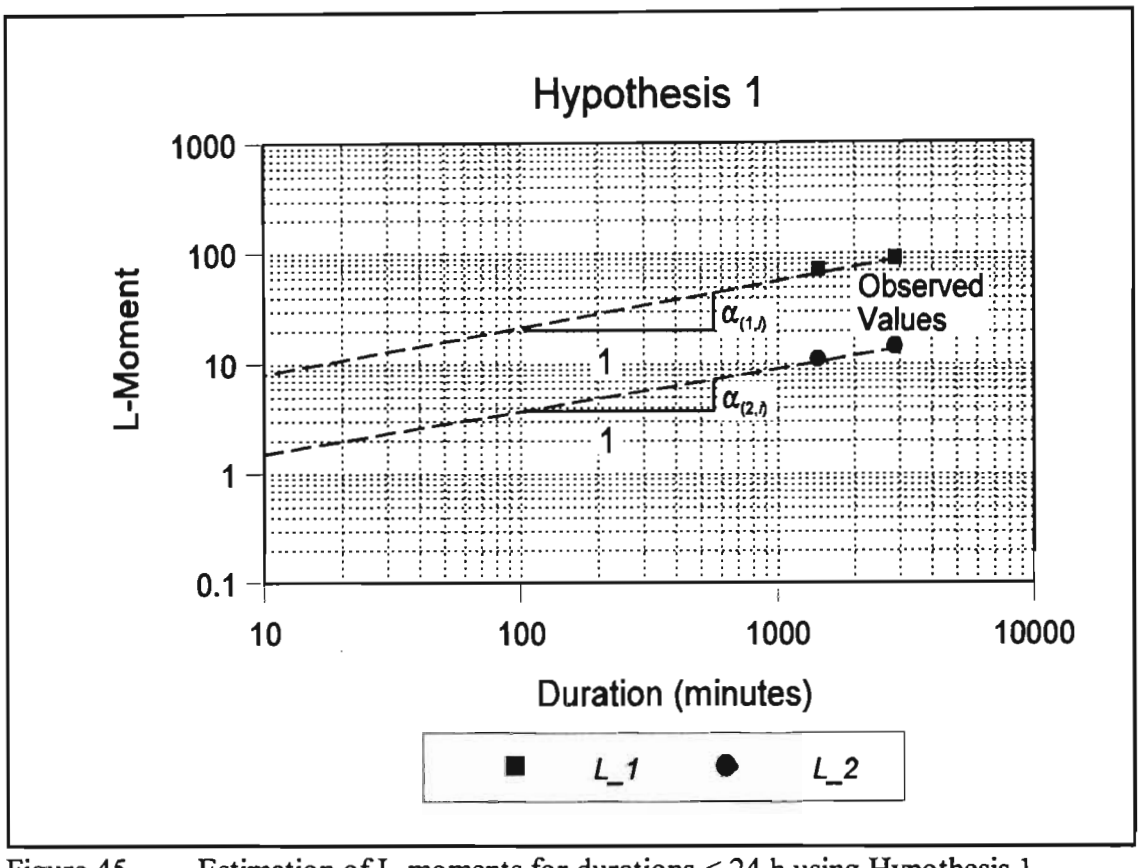


Figure 45 Estimation of L-moments for durations < 24 h using Hypothesis 1

Thus, at a site where rainfall data recorded only at a daily interval are available, the 1 and 2 day L-moments can be computed from the daily data and converted into equivalent 24 and 48 h values. L-moments for durations < 24 h can then be estimated using linear extrapolation of the 24 and 48 h values.

6.2.2 Hypothesis 2

Hypothesis 2 hypothesised that the slopes ($\alpha_{(x,i)}$) of the relationships between the log of the first and second order moments and log of event duration are linear, as shown in Figure 43, for durations ranging from 1 to 24 h and that the slopes of the relationships can be regionalised and estimated from site characteristics. The slope at site *i* of the log-transformed L-moment:duration relationship, estimated as a function of site characteristics

for each relatively homogeneous cluster identified in Chapter 5, is termed the Regional Slope ($RS_{(x,i)}$). Hypothesis 2 is illustrated schematically in Figure 46 and is implemented using the following algorithm:

(i) The slopes of the log of the L_1 ($\alpha_{(1,i)}$) and L_2 ($\alpha_{(2,i)}$) moments vs log of duration are estimated for each site i in each cluster using observed $L\_1_{(i,D)}$ and $L\_2_{(i,D)}$ values for durations ranging from 1 to 24 h.

(ii) Using multiple linear regression analysis, the $\alpha_{(1,i)}$ and $\alpha_{(2,i)}$ values are regressed against site characteristics and hence the $RS_{(1,i)}$ and $RS_{(2,i)}$ values, estimated using the regression equation for the relevant cluster and site characteristics for site i, are estimates of $\alpha_{(1,i)}$ and $\alpha_{(2,i)}$ respectively. The relationships for estimating $RS_{(1,i)}$ and $RS_{(2,i)}$ are presented in Section 6.3.

(iii) $L\_1_{(i,D)}$ and $L\_2_{(i,D)}$ for $D < 24$ h are estimated using Equation 67, where $L\_x_{(i,24)}$ is estimated directly from the observed digitised data.

$$\log\left(\hat{L\_x}_{(i,D)}\right) = \log(L\_x_{(i,24)}) - \left(RS_{(x,i)} \times \left(\log(1440) - \log(D \times 60)\right)\right) \qquad ...67$$

The slopes of the log-transformed L_1 and L_2:duration relationships are estimated using $RS_{(1,i)}$, $RS_{(2,i)}$ and the site characteristics. L_1 and L_2 moments for durations less than 24 h are computed from the estimated slope and observed 24 h L-moments as shown in Equation 67, with the 24 h L-moments computed from the digitised rainfall data. Thus Hypothesis 2 is applicable only at sites which have short duration rainfall data available. In order to fit distributions which have three or more parameters, Hypothesis 2 assumes that these higher order moments (≥ 3) can be estimated using regional record length weighted L-moment ratios.

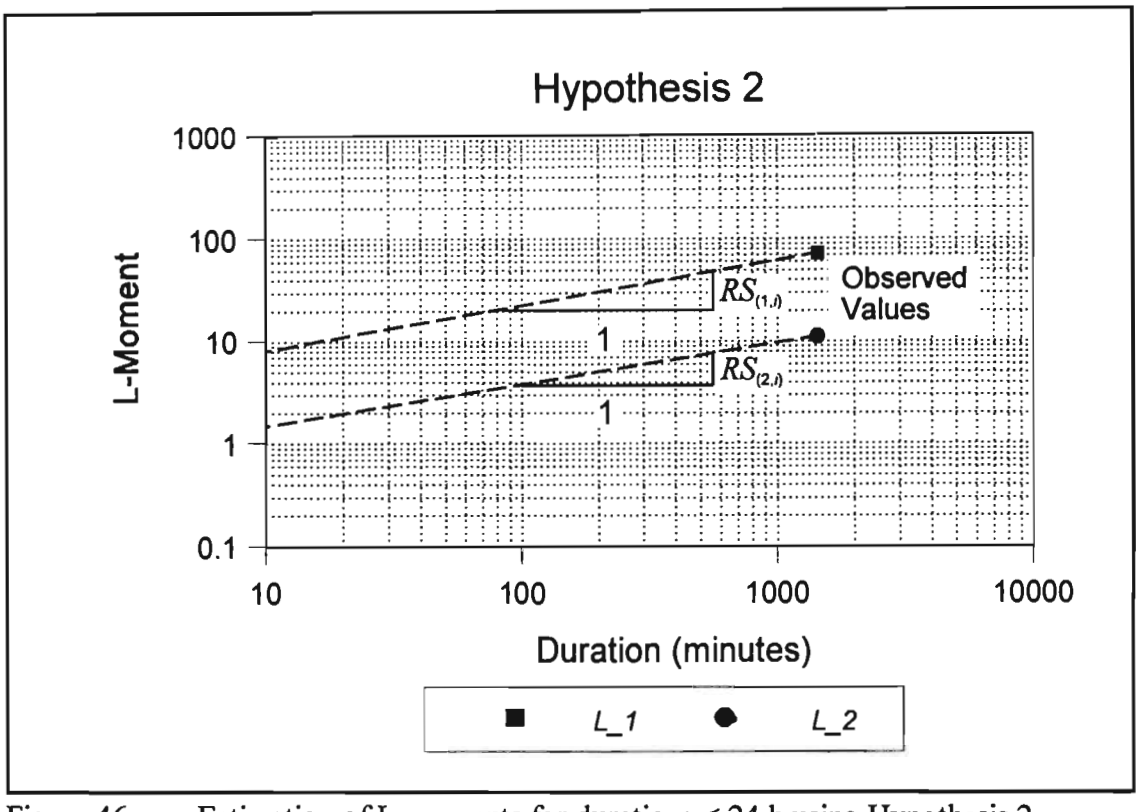


Figure 46 Estimation of L-moments for durations < 24 h using Hypothesis 2

6.2.3 Hypothesis 3

Hypothesis 3 assumed that the Regional L-Moment Algorithm (RLMA), as described in Section 2.2.3 is applicable. The RLMA is an index value procedure, which is commonly referred to as an index "flood" procedure. This assumes that the distribution of the dimensionless values $q_{(i,j,k)} = Q_{(i,j,k)} / L\_1_{(i,k)}$ is common to all sites within a relatively homogeneous region, where $Q_{(i,j,k)}$ is the j-th element of the AMS of k hour duration events at site i and $L\_1_{(i,k)}$ is the mean of the AMS of k hour duration events at site i. The algorithm for the implementation of Hypothesis 3 in each relatively homogeneous cluster is:

(i) Calculate the L-moments ($L\_x_{(i,k)}$) of the AMS ($Q_{(i,j,k)}$) for k hour duration events at each site i in the region.

(ii) Calculate regional average, record length weighted, L-moments as shown in Equation 68.

$$L\_x^R{}_{(k)} = \left[\sum_{i=1}^{N} n_i \times \frac{L\_x_{(i,k)}}{L\_1_{(i,k)}}\right] / \left[\sum_{i=1}^{N} n_i\right] \qquad ...68$$

where

$L\_x^R{}_{(k)}$ = regional average x-th order L-moment for duration k hours, $x \leq 5$,

$L\_x_{(i,k)}$ = x-th order L-moment for duration k hours at site i,

N = number of sites in region, and

n_i = record length at site i.

(iii) The first and second order L-moments at site i are estimated as shown in Equation 69, where $L\_1_{(i,k)}$ is computed directly from the observed digitised rainfall data.

$$\hat{L\_x}_{(i,k)} = L\_x^R{}_{(k)} \times L\_1_{(i,k)} \qquad ...69$$

where

$\hat{L\_x}_{(i,k)}$ = x-th order L-moment at site i, $x \leq 2$.

(iv) In order to fit distributions which have more than two parameters, third and higher order moments are estimated from the Regional Average L-Moment ($L\_x^R{}_{(k)}$) computed as shown in Equation 68.

Thus the first order regional average L-moment, $L\_1^R{}_{(k)} = 1$. In Hypothesis 3, the regional average L-moments are re-scaled using $L\_1_{(i,k)}$ estimated from the observed digitised data and thus observed short duration rainfall data are required to implement Hypothesis 3.

6.2.4 Hypothesis 4

Hypothesis 4 combines the approaches used in Hypotheses 2 and 3. The algorithm, as detailed below, utilises the multiple linear regression equations of site characteristics to estimate $L\_1_{(i,24)}$, as described in Section 5.4, and with $RS_{(1,i)}$ enable the estimation of $L\_1_{(i,D)}$

values for durations < 24 h, which in turn are used to dimensionlise the regional average L-moments.

(i) Use Equation 59 with parameters appropriate for the cluster and characteristics of site *i* to estimate $L\_1_{(i,24)}$.

(ii) Estimate $RS_{(1,i)}$ as a function of site characteristics as described for Hypothesis 2. These relationships are developed for each cluster in Section 6.3.

(iii) Use Equation 67 and the estimated $L\_x_{(i,24)}$ from (i) and $RS_{(1,i)}$ from (ii) to estimate $L\_1_{(i,D)}$ for site *i*, where $D < 24$ h.

(iv) Use $L\_1_{(i,D)}$ estimated in (i) for 24 h durations and in (iii) for durations < 24 h to dimensionalise the first and second order regional average L-moments, as computed in Equation 68.

(v) Third and higher order L-moments are assumed to be equal to the regional average L-moment ratios.

Thus the implementation of Hypothesis 4 does not require any observed data and can be implemented at any ungauged location in South Africa.

6.2.5 Hypothesis 5

Hypothesis 5 is similar to Hypothesis 4 but the mean of the 24 h AMS ($L\_1_{(i,24)}$), estimated as a function of site characteristics in Part (i) of Hypothesis 4, is replaced in Hypothesis 5 by $L\_1_{(i,24)}$ estimated from the daily rainfall data. This hypothesis was introduced to investigate and illustrate discrepancies between the digitised and daily rainfall data. Parts (ii) to (v) in the algorithm for Hypothesis 4 also apply to Hypothesis 5.

No adjustment is made to convert the moments computed from daily rainfall data, commonly referred to as fixed time increment or "clock" time, into equivalent 24 h values extracted from digitised data. Thus differences between the 24 h L_1 values calculated from the digitised and daily data are highlighted by this hypothesis.

6.2.6 Hypothesis 6

Similar to Hypothesis 5, Hypothesis 6 hypothesised that the 24 h regional average L-moments, as calculated in Equation 68, could be re-scaled using 24 h L_1 values computed from the daily rainfall database and adjusted to account for the difference between the 24 h (from the digitised database) and 1 day (from the daily rainfall database) L_1 values. Hence, $L\_1_{(i,24)}$ in part (i) of Hypothesis 4 is estimated from the daily rainfall database and, increased by the mean, for each cluster, of the ratio of 24 h to 1 day L_1 values. The relationship between the L-moments computed from continuous time (digitised) and fixed time increment data (daily), developed for each relatively homogeneous rainfall cluster in South Africa, are presented in Section 6.4. Parts (ii) to (v) in the algorithm for Hypothesis 4 also apply to Hypothesis 6. The six hypotheses evaluated are summarised in Table 48.

Table 48 Summary of hypotheses

Hypothesis	Method for Estimation of first and second L-Moments for durations < 24 h
0	Historical data
1	Multiple Scaling from 24 h and 48 h values
2	$RS_{(x,1)}$ = f(region, site characteristics) and observed $L\_x_{(i,24)}$
3	$L\_x^R_{(D)}$ re-scaled with observed $L\_1_{(i,D)}$
4	$L\_x^R_{(D)}$ re-scaled with $L\_1_{(i,D)}$ estimated using $L\_1_{(i,24)}$ = f (region, site characteristics) and $RS_{(1,1)}$ = f(region, site characteristics)
5	$L\_x^R_{(D)}$ re-scaled with $L\_1_{(i,D)}$ estimated using $L\_1_{(i,24)}$ computed from daily rainfall data and $RS_{(1,1)}$ = f(region, site characteristics)
6	$L\_x^R_{(D)}$ re-scaled with $L\_1_{(i,D)}$ estimated using $L\_1_{(i,24)}$, computed from daily rainfall data and adjusted using regionalised 24 h : 1 day ratios, and $RS_{(1,1)}$ = f(region, site characteristics)

The site characteristics and cluster locations of all the stations used in the cluster analysis are listed in Appendix A. The results from the estimation of the $RS_{(i,24)}$ for each cluster using site characteristics are presented in the following section.

6.3 ESTIMATION OF REGIONAL L-MOMENT:DURATION SLOPE

In order to estimate $\alpha_{(1,i)}$ and $\alpha_{(2,i)}$, the respective slopes of the linear relationship between the log of the first and second order L-moments and log of event duration at an ungauged site *i*, multiple linear regressions were developed for each cluster between $\alpha_{(1,i)}$ and $\alpha_{(2,i)}$ and the characteristics of each site *i* in the cluster. The values of $\alpha_{(1,i)}$ and $\alpha_{(2,i)}$ estimated at site *i* using the regression equations and characteristics of site *i* are termed the Regional Slope, $RS_{(1,i)}$ and $RS_{(2,i)}$ respectively. The form of the regression developed for each relatively homogeneous cluster is shown in Equation 70 and the results of the multiple linear regression analyses, with the objective of maximising R^2, are presented in Table 49.

$$RS_{(x,i)} = \left(\sum_{i=1}^{m} Var_i \times Cof_i \right) + Cons \qquad ...70$$

where

$RS_{(x,i)}$ = slope between the log of the first (x=1) and second (x=2) L-moments and log of event duration, estimated as a function of site characteristics,

m = number of variables (7),1=Latitude, 2=Longitude, 3=MAP, 4=Altitude, 5=Seasonality, 6=Precipitation Concentration (Ppt. Conc.), 7=Distance to Sea (Dist. Sea),

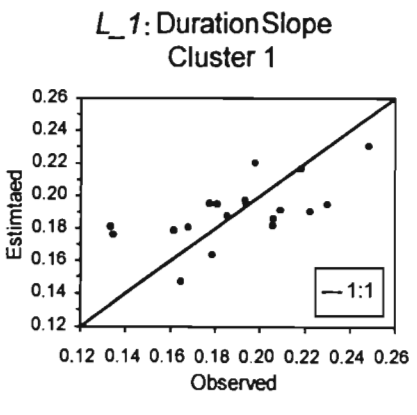
Var_i = i-th variable,

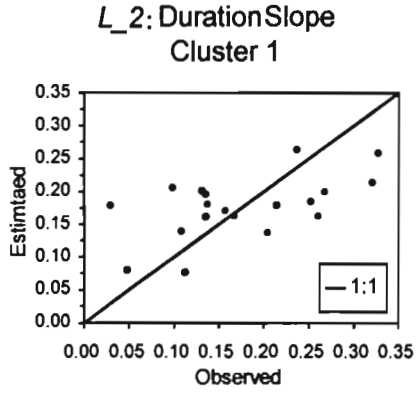
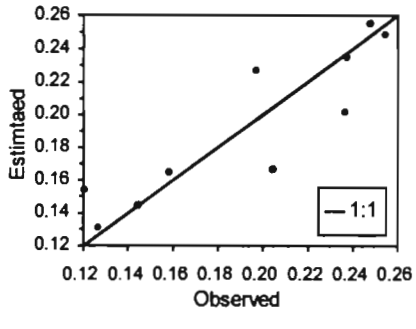
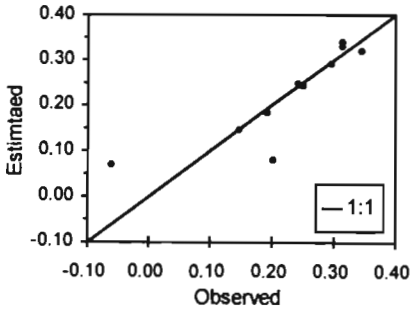
Cof_i = coefficient for i-th variable, and

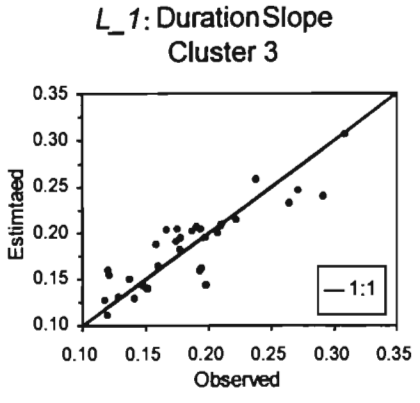
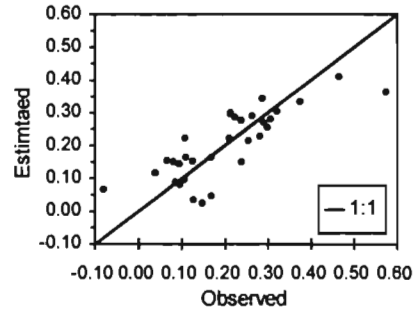
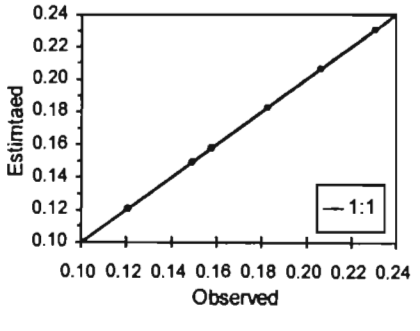
$Cons$ = constant.

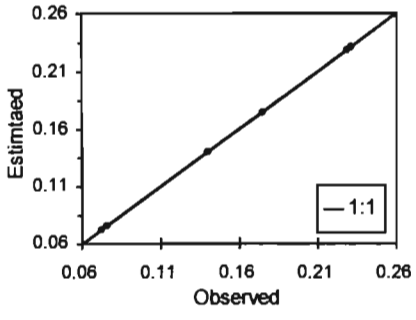
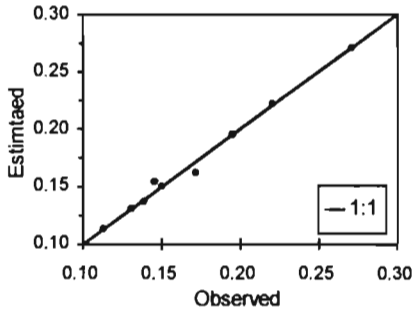
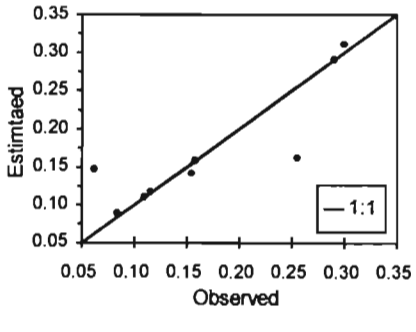
The limitations of the regression analysis as a result of the selection of independent variables and insufficient degrees of freedom in some clusters, as pointed out in Section 5.4, are also applicable to the analysis in this section. As shown in Table 49, with the exception of Clusters 1 and 11, the slope of the log-transformed *L_1*:duration and *L_2*:duration curves can be estimated relatively well using linear relationships of the individual site characteristics. Generally, an inverse relationship is evident between R^2 and the number of sites (N) where, as expected, high R^2 values are obtained for regions with fewer sites, particularly when $N \leq$ number of variables. Clusters 1 and 11, as shown in Figure 36, are adjacent clusters with the centre of the "cloud" of stations comprising the two stations located in Gauteng Province, and the clusters extending into the Free State, North-West, Northern and Mpumalanga Provinces. The H heterogeneity test-statistic, shown in Table 40, is low for both clusters indicating relatively homogeneous clusters. High intensity short duration thunderstorms dominate in these areas and hence it is probable that the contrast in the AMS between shorter and longer durations may explain the poor regressions obtained in these two clusters.

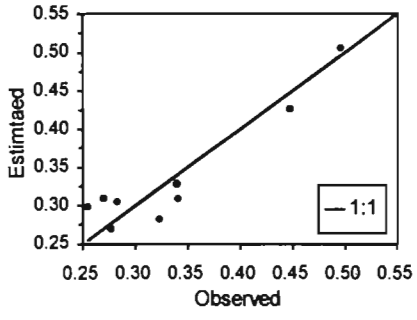
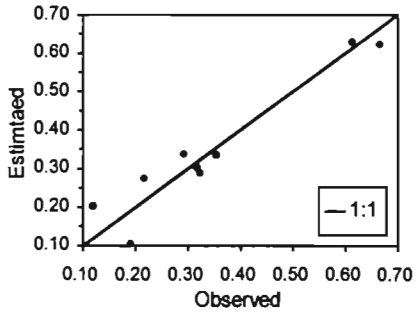
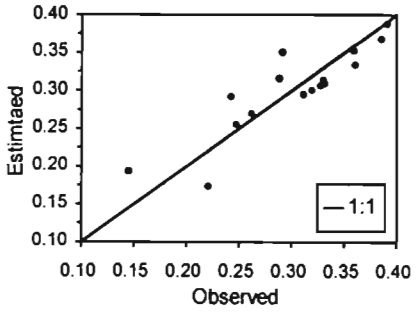
Table 49 Estimation of $RS_{(1,i)}$ and $RS_{(2,i)}$, the slopes between the log of the first and second order L-moments and log of event duration at site i, as function of site characteristics

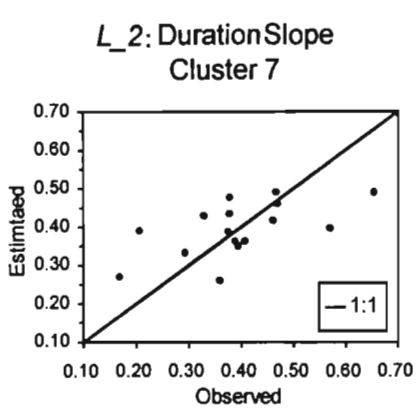
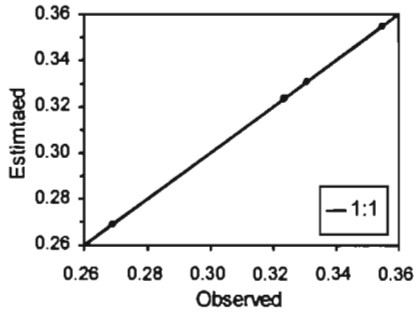
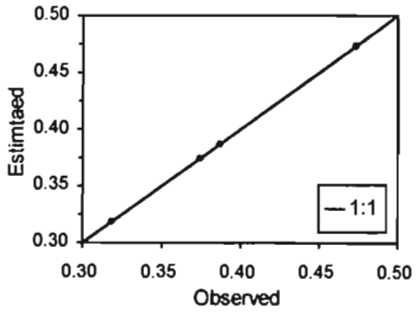
Cluster (No. of Stations)	L-moment	Regression			Scatter Plot
		Variable	Coefficient	R^2	
1 (19)	*L_1*	Intercept	3.32692878	0.40	
		Latitude	-0.02261845		
		Longitude	-0.08166781		
		MAP	0.00034695		
		Altitude	-0.00004258		
		Seasonality	0.01321386		
		Ppt. Conc.	-0.00182635		
		Dist. Sea	-0.00000081		

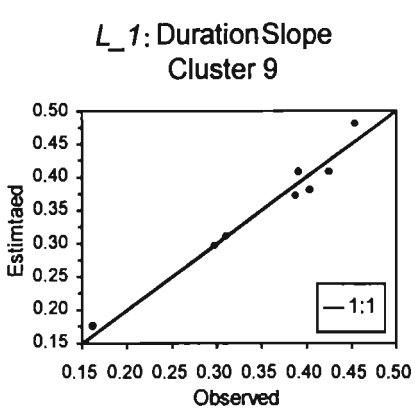
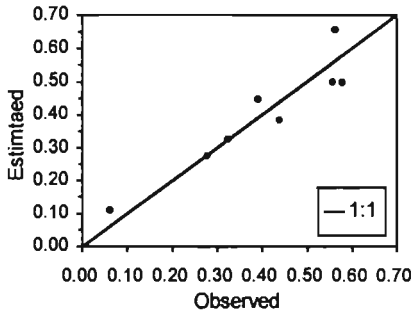
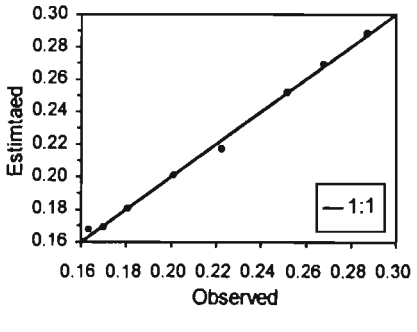
Cluster (No. of Stations)	L-moment	Regression			Scatter Plot
		Variable	Coefficient	R^2	
1 (19)	*L_2*	Intercept	9.29410811	0.32	
		Latitude	-0.12444888		
		Longitude	-0.17344196		
		MAP	0.00061569		
		Altitude	0.00011165		
		Seasonality	-0.00390009		
		Ppt. Conc.	-0.01448632		
		Dist. Sea	-0.00000148		
2 (10)	*L_1*	Intercept	-7.78895453	0.80	
		Latitude	0.00152912		
		Longitude	0.22013555		
		MAP	-0.00023471		
		Altitude	0.00085703		
		Seasonality	0.16458970		
		Ppt. Conc.	0.00388954		
		Dist. Sea	0.00000032		
2 (10)	*L_2*	Intercept	29.64861821	0.73	
		Latitude	-0.53645250		
		Longitude	-0.47785172		
		MAP	-0.00049777		
		Altitude	0.00042750		
		Seasonality	0.02259341		
		Ppt. Conc.	-0.00028607		
		Dist. Sea	-0.00000629		

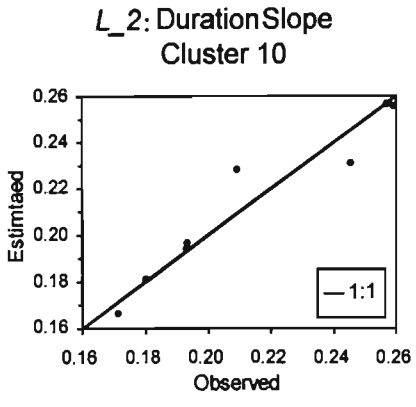
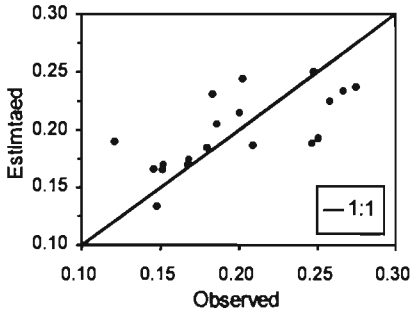
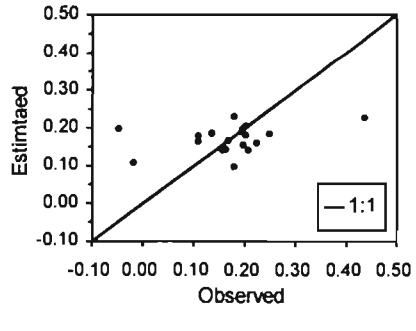
Cluster (No. of Stations)	L-moment	Regression			Scatter Plot
		Variable	Coefficient	R^2	
3 (32)	*L_1*	Intercept	17.08102924	0.77	
		Latitude	-0.25883251		
		Longitude	-0.30461415		
		MAP	0.00025878		
		Altitude	0.00011395		
		Seasonality	0.04999653		
		Ppt. Conc.	-0.00475038		
		Dist. Sea	-0.00000402		
3 (32)	*L_2*	Intercept	8.15777083	0.64	
		Latitude	-0.19290908		
		Longitude	-0.09167649		
		MAP	0.00005257		
		Altitude	0.00041928		
		Seasonality	0.20818367		
		Ppt. Conc.	-0.00888683		
		Dist. Sea	-0.00000379		
4 (6)	*L_1*	Intercept	5.33677312	1.00	
		Latitude	-0.04184360		
		Longitude	0.16042363		
		Altitude	-0.00213594		
		Ppt. Conc.	-0.06928734		
		Dist. Sea	-0.00000255		

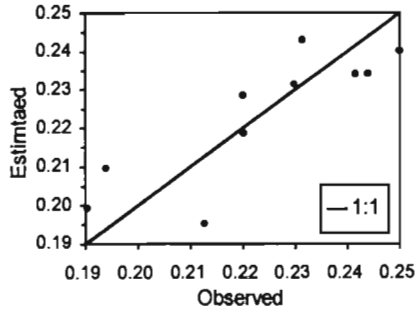
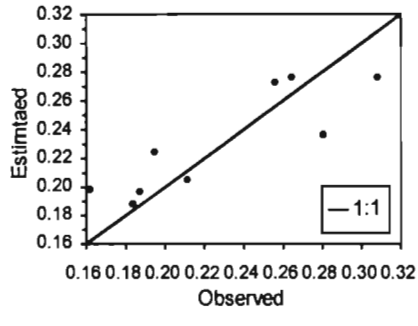
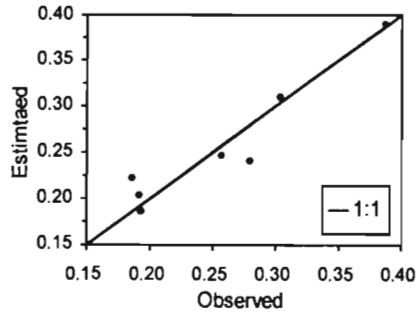
Cluster (No. of Stations)	L-moment	Regression			Scatter Plot
		Variable	Coefficient	R^2	
4 (6)	*L_2*	Intercept	6.36626997	1.00	
		Longitude	0.07042516		
		MAP	-0.00005494		
		Altitude	-0.00239427		
		Ppt. Conc.	-0.07937381		
		Dist. Sea	-0.00000119		
5 (9)	*L_1*	Intercept	-16.47806110	1.00	
		Latitude	0.47191420		
		Longitude	0.36886178		
		MAP	-0.00166477		
		Altitude	-0.00062601		
		Seasonality	0.04468570		
		Ppt. Conc.	-0.08963167		
		Dist. Sea	0.00000437		
5 (9)	*L_2*	Intercept	-40.71804086	0.75	
		Latitude	1.09692050		
		Longitude	0.83446100		
		MAP	-0.00373279		
		Altitude	-0.00122770		
		Seasonality	0.09706486		
		Ppt. Conc.	-0.16575464		
		Dist. Sea	0.00000970		

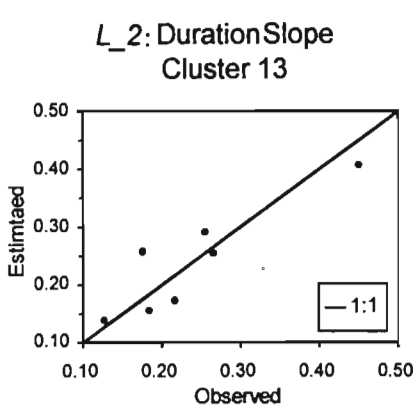
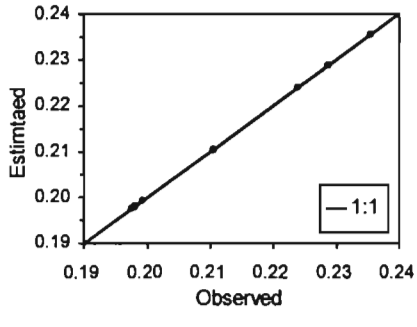
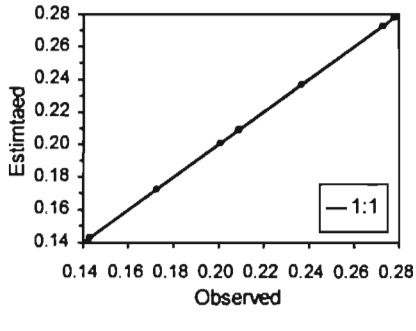
Cluster (No. of Stations)	L-moment	Regression			Scatter Plot
		Variable	Coefficient	R^2	
6 (9)	*L_1*	Intercept	-11.44424675	0.90	
		Latitude	0.35069325		
		Longitude	0.00907092		
		MAP	0.00010472		
		Altitude	0.00046264		
		Seasonality	-0.00774464		
		Ppt. Conc.	-0.00000223		
		Dist. Sea			
6 (9)	*L_2*	Intercept	3.69512007	0.91	
		Latitude	0.10607396		
		Longitude	-0.29091669		
		MAP	0.00018111		
		Altitude	0.00081886		
		Seasonality	-0.03326975		
		Ppt. Conc.	-0.00000671		
		Dist. Sea			
7 (16)	*L_1*	Intercept	84.61922144	0.78	
		Latitude	-1.34861054		
		Longitude	-1.39643644		
		MAP	-0.00011515		
		Altitude	0.00086027		
		Seasonality	-0.01415125		
		Ppt. Conc.	-0.02388912		
		Dist. Sea	-0.00002637		

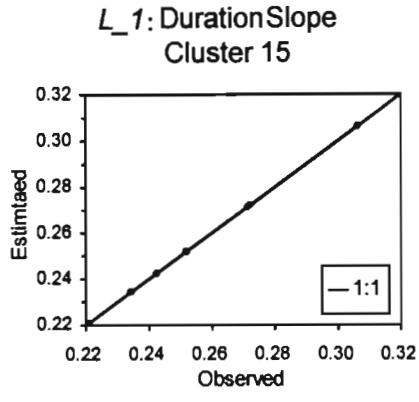
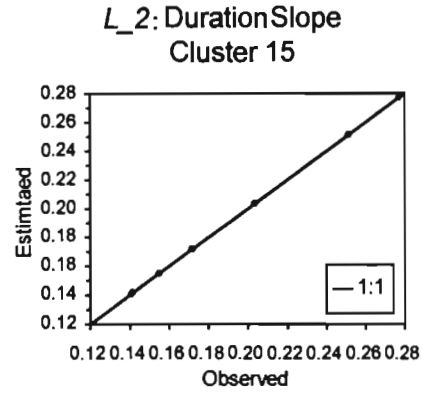
Cluster (No. of Stations)	L-moment	Regression			Scatter Plot
		Variable	Coefficient	R^2	
7 (16)	*L_2*	Intercept	49.24991585	0.34	
		Latitude	-1.05906274		
		Longitude	-0.58962188		
		MAP	0.00048030		
		Altitude	0.00088894		
		Seasonality	-0.00054789		
		Ppt. Conc.	-0.00267573		
		Dist. Sea	-0.00002351		
8 (4)	*L_1*	Intercept	-0.00808927	1.00	
		Altitude	0.00157935		
		Seasonality	0.05292950		
		Dist. Sea	-0.00000008		
8 (4)	*L_2*	Intercept	2.48097019	1.00	
		Latitude	-0.05842225		
		MAP	-0.00044648		
		Dist. Sea	0.00001162		

Cluster (No. of Stations)	L-moment	Regression			Scatter Plot
		Variable	Coefficient	R^2	
9 (8)	*L_1*	Intercept	-14.40664443	0.96	
		Latitude	0.43109168		
		Longitude	0.03477834		
		MAP	-0.00284800		
		Altitude	0.00096018		
		Seasonality	0.09852561		
		Ppt. Conc.	-0.00001039		
		Dist. Sea			
9 (8)	*L_2*	Intercept	-16.84573892	0.88	
		Latitude	0.53415471		
		Longitude	0.03061470		
		MAP	-0.00570828		
		Altitude	0.00179464		
		Ppt. Conc.	0.18046385		
		Dist. Sea	-0.00002478		
10 (8)	*L_1*	Intercept	-145.61846424	0.99	
		Latitude	4.28644196		
		Longitude	-0.16005348		
		MAP	0.00442788		
		Altitude	0.00166128		
		Ppt. Conc.	0.09483149		
		Dist. Sea	0.00002825		

Cluster (No. of Stations)	L-moment	Regression			Scatter Plot
		Variable	Coefficient	R^2	
10 (8)	*L_2*	Intercept	82.71929376	0.93	
		Latitude	-2.42799604		
		Longitude	0.09531528		
		MAP	-0.00219901		
		Altitude	-0.00101570		
		Ppt. Conc.	-0.06082819		
		Dist. Sea	-0.00001472		
11 (19)	*L_1*	Intercept	0.09289562	0.46	
		Latitude	0.01222547		
		Longitude	-0.01378212		
		MAP	0.00015475		
		Altitude	0.00001024		
		Seasonality	0.02587884		
		Ppt. Conc.	-0.00071644		
11 (19)	*L_2*	Intercept	-1.11917566	0.12	
		Latitude	0.05371173		
		Longitude	-0.00593565		
		MAP	0.00074981		
		Altitude	-0.00012576		
		Seasonality	-0.06973502		
		Ppt. Conc.	-0.00106577		

Cluster (No. of Stations)	L-moment	Regression			Scatter Plot
		Variable	Coefficient	R^2	
12 (10)	*L_1*	Intercept	-0.00494577	0.70	
		Latitude	0.00028621		
		Longitude	0.02728788		
		MAP	-0.00031068		
		Altitude	-0.00008019		
		Seasonality	-0.00430803		
		Ppt. Conc.	-0.00565048		
		Dist. Sea	0.00000021		
12 (10)	*L_2*	Intercept	1.09420567	0.72	
		Latitude	-0.02548151		
		Longitude	0.07102855		
		MAP	-0.00076443		
		Altitude	-0.00022772		
		Seasonality	0.07619555		
		Ppt. Conc.	-0.04309336		
		Dist. Sea	0.00000137		
13 (7)	*L_1*	Intercept	0.02753683	0.90	
		Longitude	0.01992899		
		Ppt. Conc.	-0.00801313		

Cluster (No. of Stations)	L-moment	Regression			Scatter Plot
		Variable	Coefficient	R^2	
13 (7)	*L_2*	Intercept Latitude MAP	-5.70354704 0.19549067 -0.00053035	0.81	
14 (7)	*L_1*	Intercept Latitude Longitude Altitude Seasonality Ppt. Conc. Dist. Sea	-1.28232278 0.07785424 -0.04873901 -0.00041840 0.02962118 0.00441116 0.00000095	1.00	
14 (7)	*L_2*	Intercept Latitude Longitude MAP Altitude Seasonality Dist. Sea	6.59444579 -0.06864796 -0.16175284 -0.00175746 -0.00041191 0.10600035 0.00000032	1.00	

Cluster (No. of Stations)	L-moment	Regression			Scatter Plot
		Variable	Coefficient	R^2	
15 (7)	*L_1*	Intercept	0.44054040	1.00	
		Latitude	0.01546503		
		Longitude	-0.05951068		
		MAP	0.00014943		
		Altitude	-0.00001561		
		Ppt. Conc.	0.00657321		
		Dist. Sea	0.00000124		
15 (7)	*L_2*	Intercept	1.49862860	1.00	
		Latitude	-0.10552150		
		Longitude	0.14966819		
		MAP	0.00172692		
		Altitude	-0.00000724		
		Ppt. Conc.	-0.01305506		
		Dist. Sea	-0.00000529		

Hypothesis 6 requires that the *L_1* value calculated from the daily rainfall data be adjusted into a continuous 24 h value, as would be computed from digitised data for a continuous 24 h period. Regionalised 24 h : 1 day ratios for each cluster in South Africa are presented in the following Section 6.4.

6.4 CONTINUOUS : FIXED TIME *L_1* RATIOS

Hypothesis 6 assumes that the mean of the 24 h AMS at site *i* ($L\_1_{(i,24)}$), normally computed using a continuously moving 24 h "window" in the digitised rainfall data, can be estimated from the mean of the 1 day AMS extracted from the daily rainfall data recorded at fixed 24 h intervals. Thus it is required to convert the 1 day (fixed time) extreme values into equivalent 24 h values (continuous time).

Values reported in the literature for South Africa suggest that the fixed time interval extreme values should be increased by between 10 and 20% (Adamson, 1981; Schulze, 1984; Alexander, 1990). More recently, Dwyer and Reed (1995) showed that, based on theoretical considerations, the correction factor should be 1.33, but recommend a value of 1.16, which is based on rainfall data from the United Kingdom and Australia.

Ratios of the mean ($L\_1_{(i,24)}$) of the 24 h AMS and 1 day AMS were computed for each station *i* in each cluster and averages of these ratios were computed for each cluster. The results of the analysis are presented in Table 50 which contains the average ratios and their standard errors for each cluster. As noted, for example, in Sections 4.2, 4.3 and 4.4, discrepancies are evident between the digitised and daily rainfall data for most SAWB stations. Hence, in order to ensure consistency of data sets, the 1 day values used in this analysis were derived by extracting the AMS from the digitised data based on a fixed 24 h incremental period and the actual 1 day data measured by standard raingauges were not used. As shown in Table 50, the average 24 h :1 day ratios range from 1.15 to 1.28 in South Africa. These ratios are slightly larger than the values reported in Chapter 2 for South Africa which range from 1.11 to 1.20, but which were computed from the 24 h and 1 day design rainfall depths, which may incorporate bias due to the selection of distribution used to estimate the design rainfalls. Unexpectedly high values were consistently obtained for the Eastern and South Eastern Cape regions (Clusters 9 and 13 as shown in Figure 36). The values presented in Table 50, which are the average ratios for all the stations in each cluster, were used to estimate the 24 h *L_1* values from the 1 day *L_1* values, computed from the daily rainfall record.

Table 50 Ratios of 24 h :1 day *L_1* values

Cluster	Mean	Std. Error	Cluster	Mean	Std. Error
1	1.20	0.05	9	1.26	0.11
2	1.21	0.06	10	1.19	0.09
3	1.19	0.07	11	1.20	0.09
4	1.21	0.09	12	1.19	0.04
5	1.20	0.10	13	1.28	0.14
6	1.17	0.06	14	1.24	0.06
7	1.15	0.05	15	1.25	0.10
8	1.20	0.03			

Techniques for estimating the *RS* for both the *L_1* and *L_2* values, and 24 h : 1 day *L_1* ratios have been presented. In the following Section 6.5, the effect on L-moments and design storms estimated using the six hypotheses described in Section 6.2 are investigated.

6.5 EVALUATION OF SIX HYPOTHESES FOR ESTIMATING SHORT DURATION *L_1* AND *L_2* VALUES

Rainfall data from selected stations used in the delineation of relatively homogeneous clusters, as shown in Figure 36, were utilised in the evaluation of the six hypotheses. The performance of the six hypotheses, detailed in Section 6.2 and summarised in Table 48, were evaluated by the mean absolute relative deviation:

- between the L-moments estimated by the hypotheses and the L-moments computed from the observed digitised rainfall data, and
- between design rainfall events estimated using the GEV distribution fitted to the L-moments estimated by the hypotheses and fitted to the L-moments computed from the observed digitised rainfall data.

The errors found throughout the SAWB digitised rainfall database and the inconsistencies between the digitised and daily rainfall data at SAWB raingauges, as detailed in Chapter 4, has led to the assumption that the majority of the SAWB digitised rainfall data are of low reliability. The frequency of errors found in all non-SAWB digitised rainfall data was nearly zero. However, most of the data from these non-SAWB stations were digitised from autographic rainfall charts that were changed at weekly intervals. Hence, consistency checks between the digitised and daily rainfall totals could not be performed, as was done for the SAWB stations. However, the data collection procedures followed, for example by the DAEUN, do include routine consistency checks between the total rainfall measured for the duration of the chart and the rainfall digitised from the chart and data are flagged when discrepancies are noted. Hence, although the consistency checks for non-SAWB stations could not be performed as part of this study, the very few digitising errors and knowledge of data collection procedures at some non-SAWB raingauges, led to the supposition that the non-SAWB digitised rainfall data are generally relatively more reliable than the SAWB digitised rainfall data.

Detailed evaluation of the hypotheses at selected sites are presented in Section 6.5.1 for Cluster 3 which has the most non-SAWB data and which are assumed to be more reliable than the SAWB data. Thereafter, summarised results are presented for Cluster 6 (Sections 6.5.2) and for one of two selected sites in clusters located in different geographic and climatic regions of South Africa (Section 6.5.3).

6.5.1 Cluster 3

Thirty-two stations are contained in Cluster 3, of which 16 stations are SAWB stations and the remaining stations are operated by the DAEUN (15) and FORESTEK (1). Hence 50% of the stations in Cluster 3 are non-SAWB stations. In the regression analyses performed, data from SAWB Station 0476131 were omitted as the data did not appear to be consistent with the rest of the data in the region.

6.5.1.1 Cathedral Peak

The results of evaluating the six hypotheses outlined in Table 48 to estimate the first two L-moments at Cathedral Peak (CP6) are shown in Figure 47. As evident from Figure 47 all hypotheses, with the exception of Hypothesis 5, estimate the *L_1* and *L_2* values computed from the observed data extremely well over the range of 2 h - 24 h duration events. Since the 24 h regional average L-moment are re-scaled by the unadjusted 1 day value in Hypothesis 5, it is not unexpected that Hypothesis 5 should estimate lower *L_1* values. Each hypothesis estimates *L_1* and *L_2* values for all durations, and the third order L-moment used is either the mean of the 24 h and 48 h values (Hypothesis 1) or the regional record length weighted value (Hypotheses 2-6). The estimates of the first three L-moments for each duration and hypothesis were used to determine the parameters of the GEV distribution. The design storm depths computed using the GEV probability distribution fitted to the L-moments estimated by Hypotheses 1-6 are shown for the 20 year return period event in Figure 48. Similar results were obtained for other return periods.

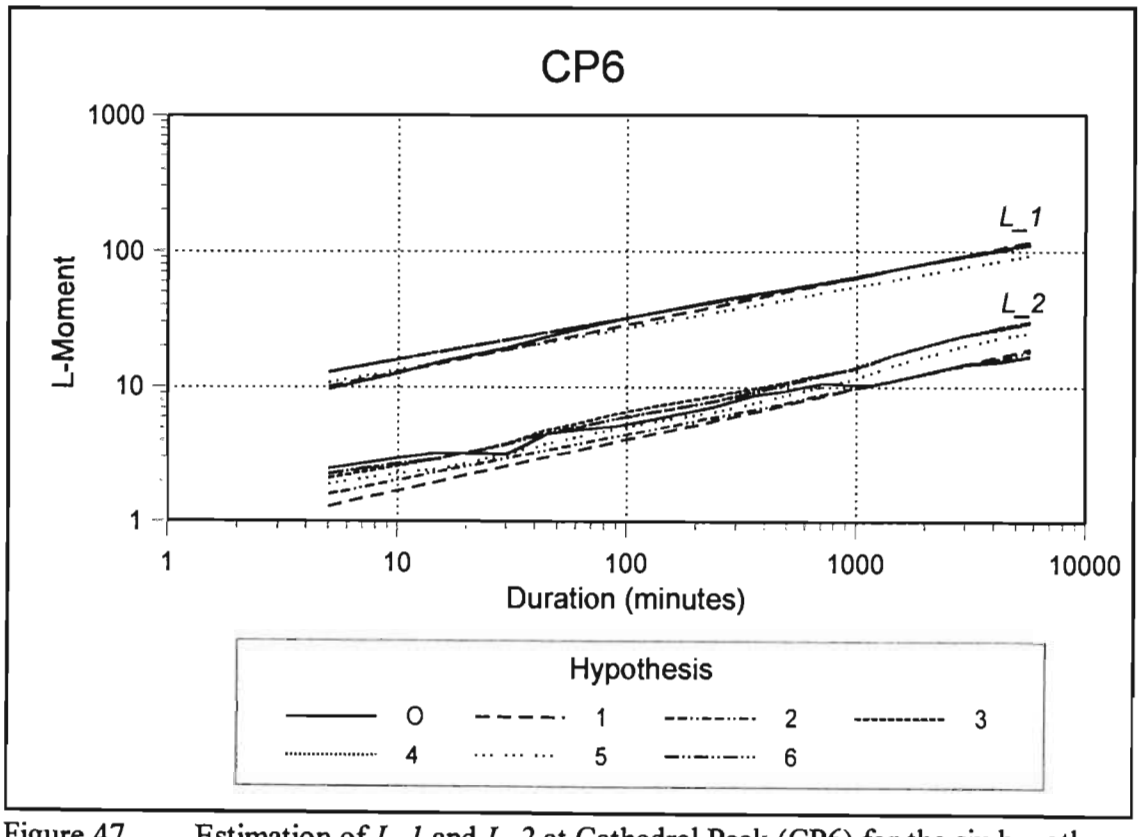


Figure 47 Estimation of *L_1* and *L_2* at Cathedral Peak (CP6) for the six hypotheses summarised in Table 48 (O=Observed, 1-6= Hypotheses)

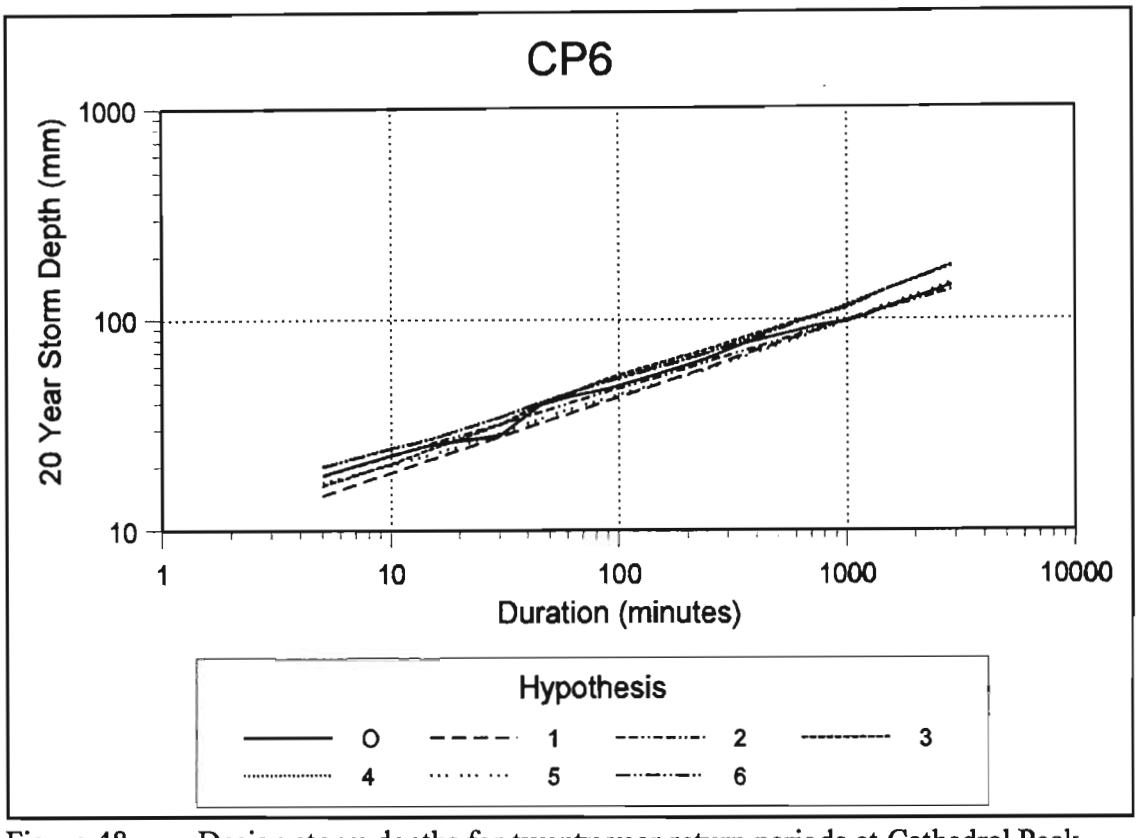


Figure 48 Design storm depths for twenty year return periods at Cathedral Peak (CP6) estimated from the observed data and for the six hypotheses summarised in Table 48 (O=observed, 1-6=Hypotheses)

The results contained in Figure 48 for the 20 year return period design storms may be further summarised by the Mean Absolute Relative Error (*MARE*) between design storm depths computed from the historical data and from each of the six hypotheses for return periods of 2, 5, 10, 20, 50 and 100 years, computed using Equation 71 and shown in Figure 49 for CP6.

$$MARE_j = \frac{100}{N_{RP}} \times \sum_{k=1}^{N_{RP}} \left(\frac{\left| S_{(j,k)} - O_{(j,k)} \right|}{O_{(j,k)}} \right) \qquad \ldots 71$$

where

$MARE_j$ = mean absolute relative error of j-th hour design rainfall (%),

$S_{(j,k)}$ = k-th return period, j-th hour annual maximum design rainfall computed using hypothesis,

$O_{(j,k)}$ = k-th return period, j-th hour design rainfall computed from observed data, and

N_{RP} = number of return periods (2, 5, 10, 20, 50 and 100).

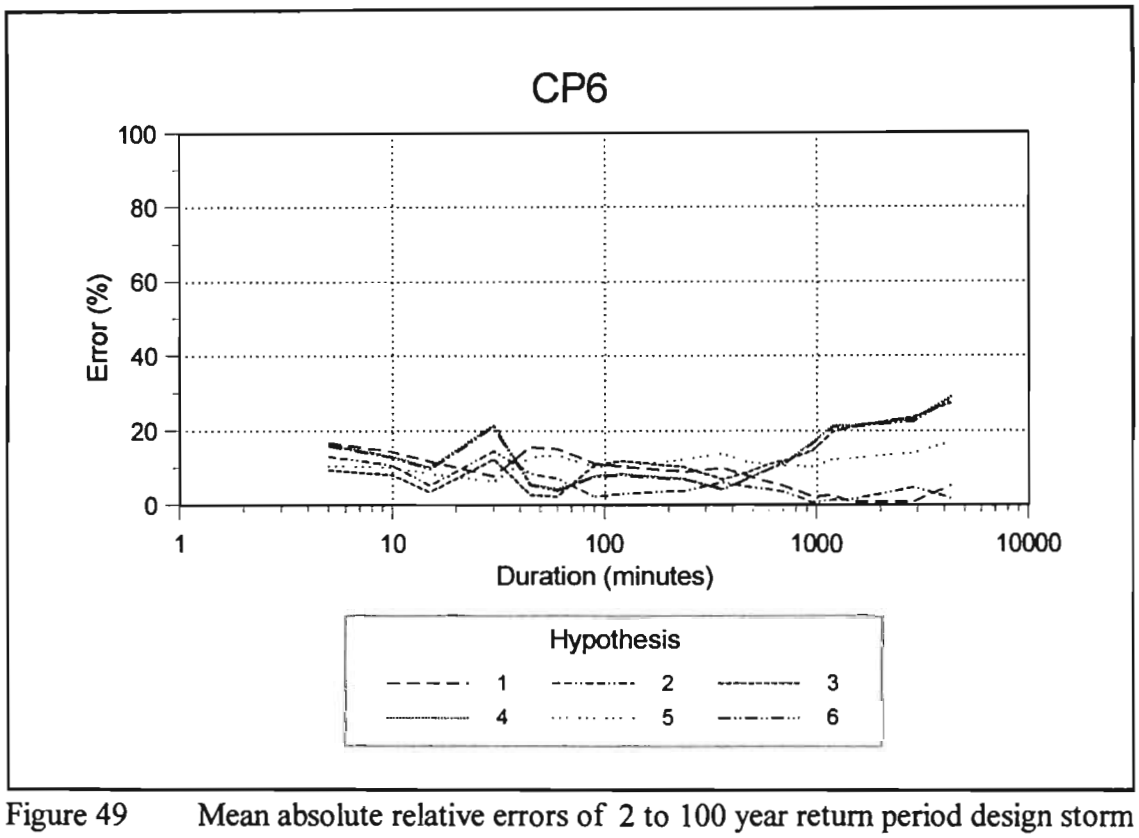


Figure 49 Mean absolute relative errors of 2 to 100 year return period design storm depths estimated at Cathedral Peak (CP6) for the six hypotheses summarised in Table 48

From Figures 48 and 49 it is evident that all the *MARE*s between design storms computed from the historical data and for each of the 6 hypotheses are less than 20% (deemed to be acceptable) for durations ≤ 24 h and the mean error is generally < 10%. Hence all six hypotheses appear to be able to produce similar L-moments and design storms at Cathedral Peak (CP6). Thus, in the event of only daily rainfall data being available at this site, reasonably accurate design storms for durations ≤ 24 h could be estimated using only data from a standard non-recording raingauge. Hypothesis 1, which is the simplest of the hypotheses evaluated and assumes multiple linear scaling of the *L_1*:duration and *L_2*:duration relationships for durations < 24 h and up to 48 h, appears to be applicable at Cathedral Peak (CP6).

The results contained in Figure 49 can be further summarised as shown in Equation 72.

$$AV - MARE = \frac{100}{N_D}\sum_{j=1}^{N_D} MARE_j \qquad ...72$$

where

$AV\text{-}MARE$ = average $MARE_j$ (%), computed from N_D durations.

The *AV-MARE* values were computed for durations ≤ 1 h and for durations ranging from 2 - 24 h for CP6 as shown in Figure 50. Hypothesis 2 resulted in the best estimation of the design storm depths at CP6 for all the periods shown in Figure 50. The next best design storm depths for durations of 2 h - 24 h were estimated by Hypothesis 1. However, the estimation of design storms at CP6 were acceptable (i.e. errors < 20%) for all hypotheses evaluated.

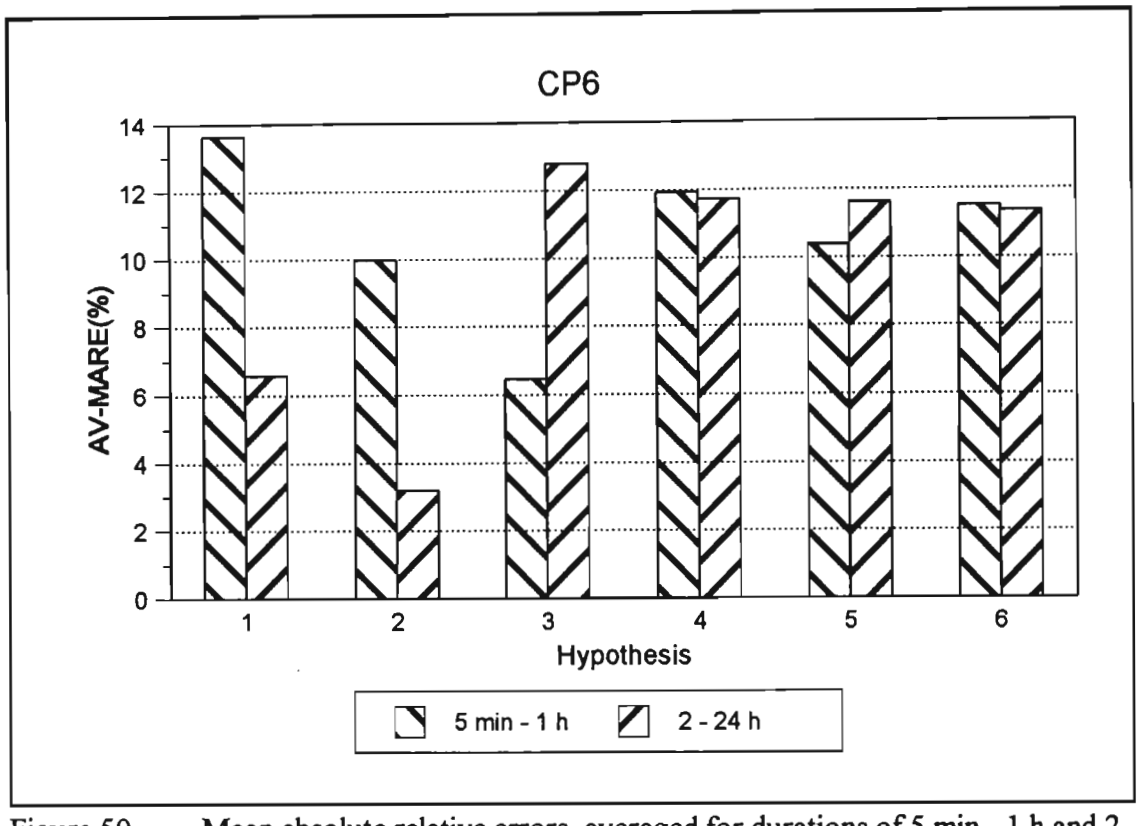


Figure 50 Mean absolute relative errors, averaged for durations of 5 min - 1 h and 2 - 24 h, of 2 to 100 year return period design storm depths estimated at Cathedral Peak (CP6) for the six hypotheses summarised in Table 48

6.5.1.2 Ntabamhlope

The DAEUN monitors and maintains a dense network of raingauges in the Ntabamhlope/De Hoek hydrological research catchments near Estcourt. One of these raingauges, N23, was not used in the establishment of homogeneous regions using cluster analysis, or in any of the regression analyses to estimate the 24 h L_1 value or in the regression analyses to estimate the regional slope of the L_1:duration relationship. Hence this site presents a good and relatively long (31 years) record to evaluate the hypotheses. The results of estimating the L-moments at Station N23 using the six hypotheses are shown in Figure 51. It is evident from Figure 51 that changes in the slope of the L_1:duration relationship occur at event durations of approximately 1 h and 24 h. Hence Hypothesis 1 is not valid at this site and the AMS for durations of 1 and 2 days cannot be used to estimate the L-moments for shorter durations. The breaks in linear scaling at approximately 1 h and 24 h is a characteristic

displayed by all the data from raingauges at Ntabamhlope. The break in linear scaling at approximately 1 h could be a result of historical periods when weekly drum-type autographic charts were used at Ntabamhlope, where each 1 mm on the chart represents approximately 0.5 h. Whilst the resolution of chart digitisation may theoretically be as good as 0.5 mm, in practice the effective resolution of the digitiser is probably closer to 1 mm. Hence, the data for durations shorter than 0.5 h when the weekly drum type charts were used, are expected to be relatively unreliable and the break in scaling at approximately 1 h may be the result of the temporal resolution of the digitisation process. However, for more than half of the 31 years of data, strip-type autographic charts were used which have a time resolution of as little as 2 minutes. Hence these breaks in linear scaling, which are also observed at most other sites located in summer rainfall regions in South Africa, may not be caused by the data measurement system. Again as expected, Hypothesis 5 which uses the 1 day *L_1* value to scale the RGC and to estimate the *RS*, underestimates the at-site L-moments. The *AV-MAREs* of the design storms computed from the estimated L-moments for the six hypotheses are shown in Figure 52.

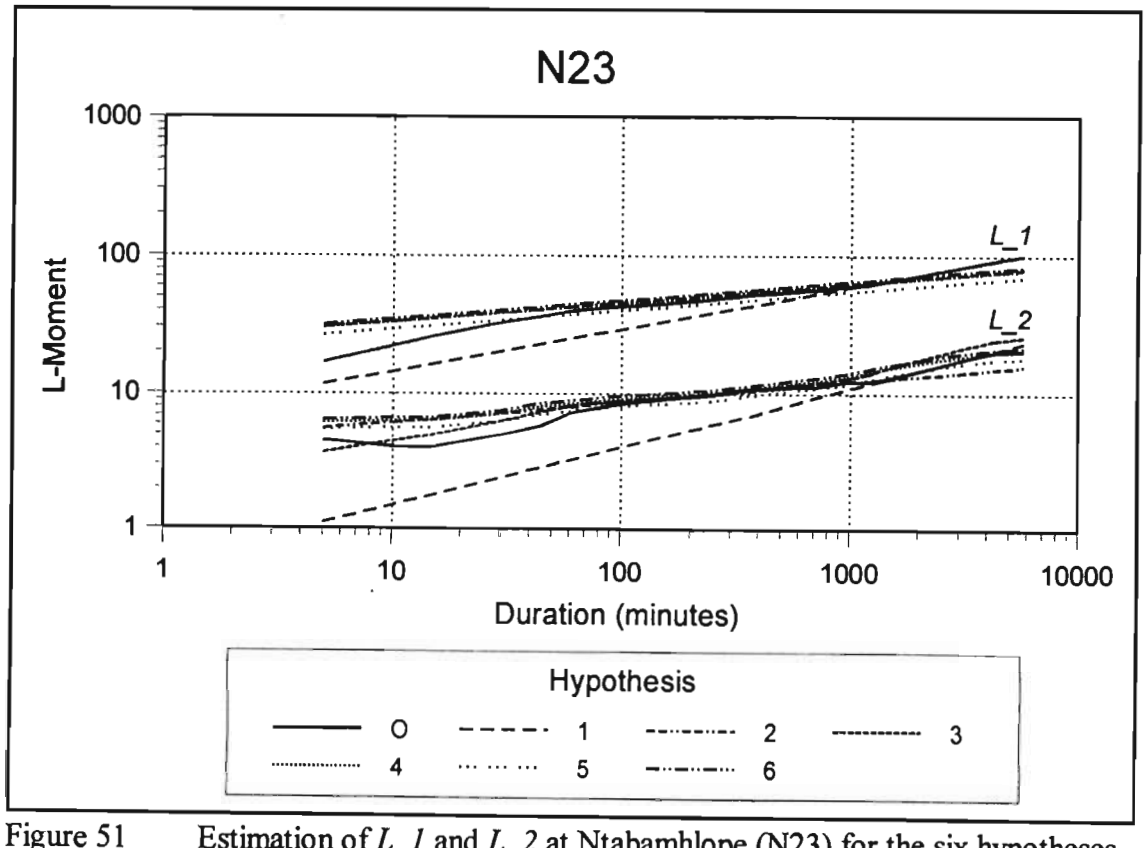


Figure 51 Estimation of *L_1* and *L_2* at Ntabamhlope (N23) for the six hypotheses summarised in Table 48

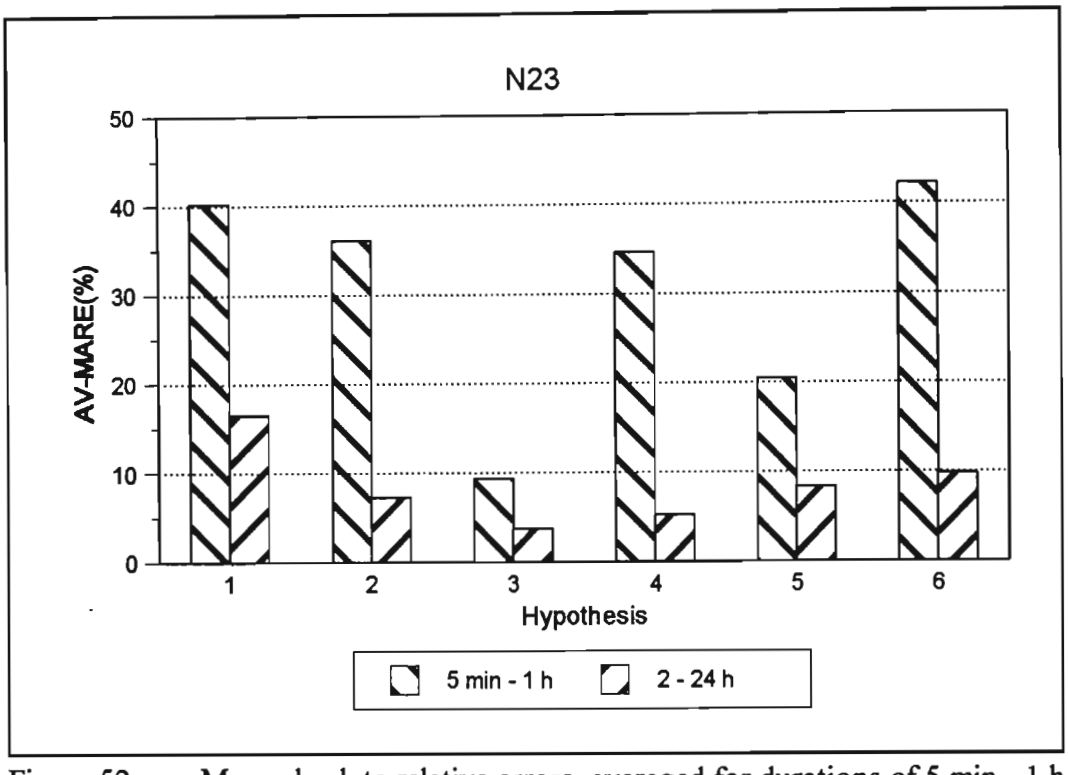


Figure 52 Mean absolute relative errors, averaged for durations of 5 min - 1 h and 2 - 24 h, of 2 to 100 year return period design storm depths estimated at Natabamhlope (N23) for the six hypotheses summarised in Table 48

With the exception of Hypothesis 1 and exclusion of Hypothesis 5, all the other hypotheses are able to estimate the design storm depths extremely well for durations ranging from 2 to 24 h at N23, as shown in Figure 52. For durations ≤ 1 h, only Hypothesis 3 resulted in acceptable design storms. Thus design storms for durations > 1 h and up to 24 h may be estimated at Ntabamhlope using only the regional average L-moments, scaled either with observed (if available) or estimated 24 h L_1 values, in conjunction with the regional slope of the log transformed L_1:duration and L_2 duration relationships.

6.5.1.3 Cedara

The DAEUN also monitors and maintains a dense network of raingauges in the Cedara hydrological research catchments near Pietermaritzburg in KwaZulu-Natal. In addition, an

official SAWB station (0239482) is located at the Cedara Agricultural Research Station. The L-moments estimated by the six hypotheses at Stations C182 and 0239482 are shown in Figures 53 and 54 respectively. At Station C182, a distinct change in the scaling of the L-moments at approximately 24 h is evident and hence Hypothesis 1 was not valid, while Hypotheses 4 and 6 slightly overestimated the *L_1* values computed from the observed values. It is noted that the 24 h *L_1* value used in Hypothesis 5, which is the 1 day *L_1* value, is less than the observed 24 h *L_1* value at C182. This is not the case for SAWB Station 0239482, where the 1 day *L_1* value (from the standard raingauge) is greater than the 24 h *L_1* value (from digitised data). Hence at SAWB Station 0239482, it is postulated that the unreliability of the data, particularly the number of missing extreme events, has resulted in the mean of the 24 h AMS to be less than the mean of the 1 day AMS. This trend is noted at many SAWB stations, reinforcing previous comments regarding the reliability of the SAWB data and the need to develop techniques to estimate design storms based on the daily, non-recording raingauge network. For durations less than 30 min the data at Station 0239482 are not consistent with the rest of the data nor with regional trends and are thus assumed to be problematic.

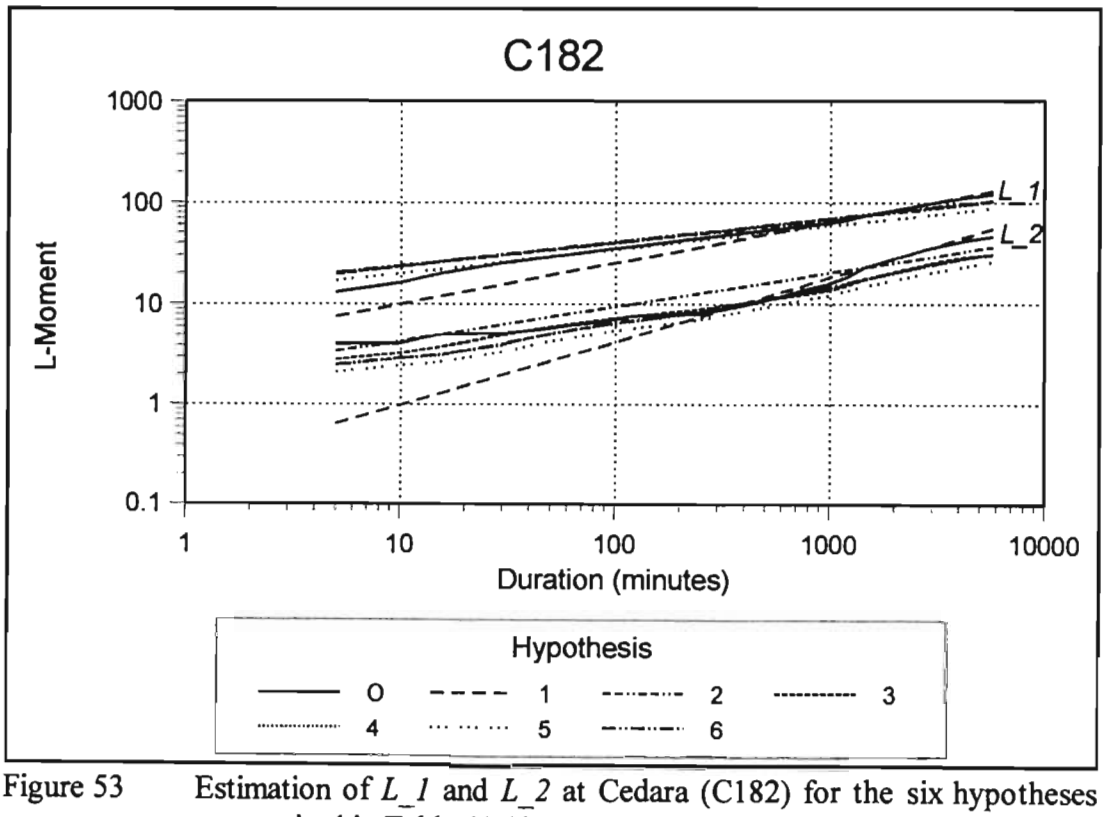


Figure 53 Estimation of *L_1* and *L_2* at Cedara (C182) for the six hypotheses summarised in Table 48 (O=Observed, 1-6=Hypotheses)

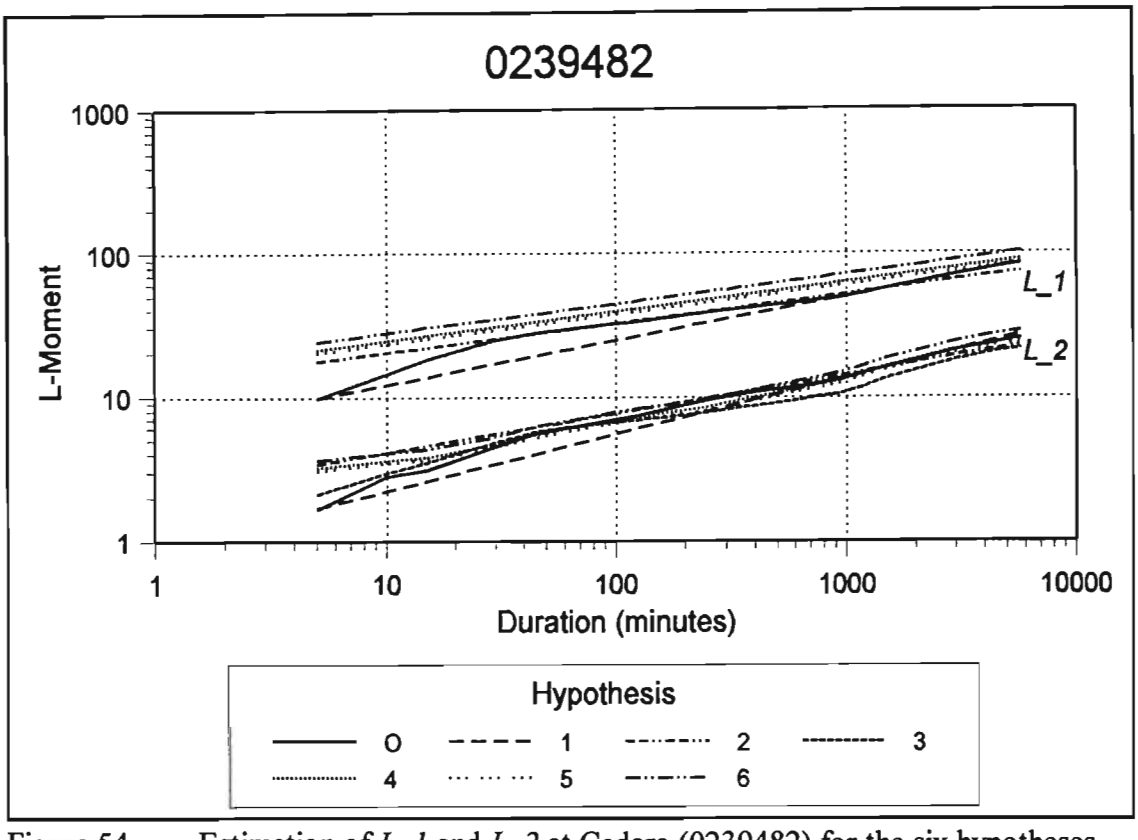


Figure 54 Estimation of *L_1* and *L_2* at Cedara (0239482) for the six hypotheses summarised in Table 48

The *AV-MARE* of design storm depths estimated for Stations C182 and 0239482 are shown in Figures 55 and 56 respectively. At C182, application of Hypotheses 3 - 6 result in the estimation of acceptable design storms for durations ranging from 5 min to 24 h, whilst design storms estimated using Hypotheses 1 and 2 exceed the "acceptable" 20% error level. The opposite trends are evident in Figure 56 where, for Station 0239482, the largest errors appear to result from Hypotheses 4-6. In Hypothesis 4 the 24 h *L_1* value is estimated using regional regressions of site characteristics, Hypothesis 5 uses the daily rainfall data to estimate the 24 h *L_1* value and Hypothesis 6 uses an adjusted daily *L_1* value to estimate the 24 h *L_1* value. Thus, in the light of the inconsistency between the digitised and daily rainfall databases at Station 0239482, it is postulated that Hypothesis 4, which utilises information from the entire region, and Hypothesis 6, which adjusts the *L_1* value extracted from the daily rainfall data into an equivalent 24 h *L_1* value, are both more reliable estimates of the true 24 h *L_1* value than the value computed directly from the digitised data, as used in Hypotheses 2 and 3. Therefore, it is postulated that the

discrepancies between design storms estimated using Hypotheses 4-6 and from the digitised data, as shown in Figure 56, are not "real" errors and merely reflect the errors in the digitised data. Thus, it is postulated that Hypotheses 4 and 6 result in the most reliable estimates of design storms at Station 0239482.

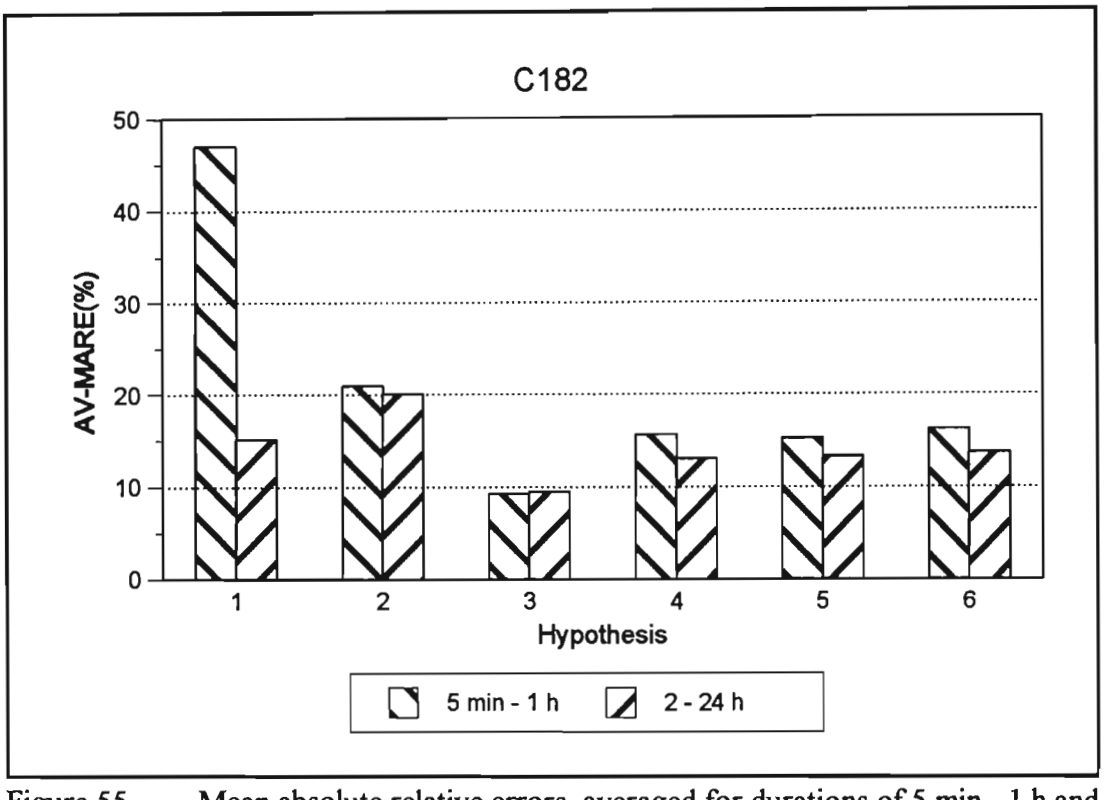


Figure 55 Mean absolute relative errors, averaged for durations of 5 min - 1 h and 2 - 24 h, of 2 to 100 year return period design storm depths estimated at Cedara (C182) for the six hypotheses summarised in Table 48

6.5.1.4 Comparison between selected stations

A detailed analysis of the performance of the six hypotheses in estimating the first and second L-moments and design storms have been presented for raingauges located at Cathedral Peak, Ntabamhlope and Cedara, all of which are located in Cluster 3. In this section the *AV-MARE*s of the design storms estimated using the hypotheses at selected stations in Cluster 3 are compared. The *AV-MARE* values of design storms at selected sites and for durations of 2 - 24 h are shown in Figure 57. In addition to the stations in Cluster

3 already discussed, Figure 57 includes results from Kokstad (0180722) and Piet Rietief (0444540).

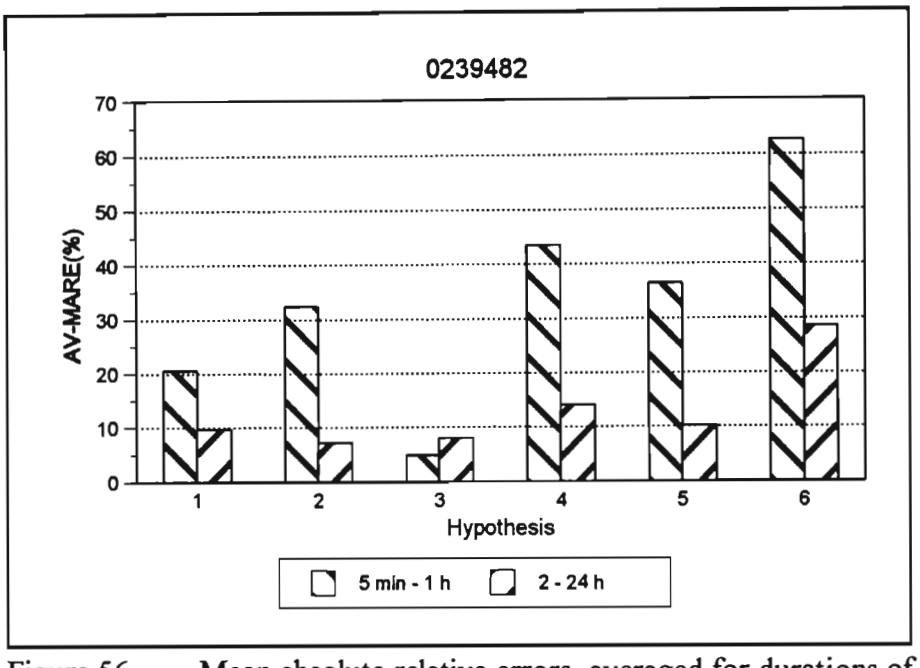


Figure 56 Mean absolute relative errors, averaged for durations of 5 min - 1 h and 2 - 24 h, of 2 to 100 year return period design storm depths estimated at Cedara (0239482) for the six hypotheses summarised in Table 48

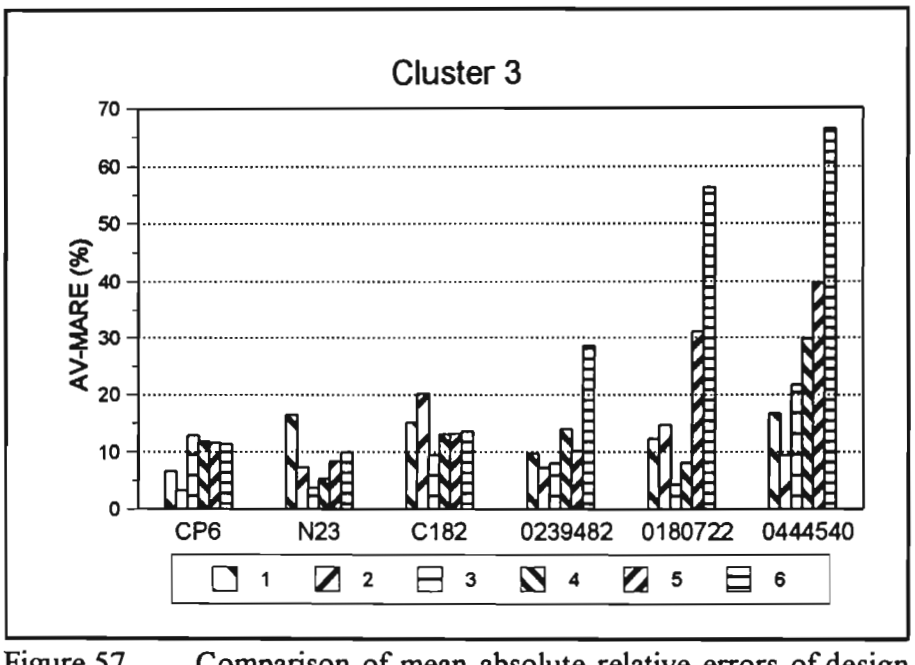


Figure 57 Comparison of mean absolute relative errors of design storms, averaged for durations of 2 - 24 h and for return periods of 2 - 100 years, estimated at selected sites in Cluster 3 for the six hypotheses summarised in Table 48

As evident in Figure 57, Hypotheses 4 and 6 performed consistently well at the non-SAWB stations (CP6, N23 and C182), but resulted in some of the largest errors at the SAWB stations (0239482, 0180722 and 0444540). As shown in Figure 58 the 24 h *L_1* values extracted from the digitised data correctly exceed the values extracted from the daily data at non-SAWB stations, and the adjusted daily value, as used in Hypothesis 6, is similar to the value extracted from the digitised data. However, at all the SAWB stations the *L_1* values extracted from the digitised data are less than those extracted from the daily rainfall data, indicating inconsistencies in the two sets of data. The limitations of the regional regression relationships which estimate the 24 h *L_1* value as a function of site characteristics, as developed in Section 5.4 and used by Hypothesis 4, are evident in Figure 58. The estimated 24 h *L_1* values tend to mimic the observed 24 h *L_1* values extracted from the digitised data, which were used in the development of the regression equations and which have been shown to be unreliable at some SAWB stations. Hence, as before, it is postulated that the best estimate of the 24 h *L_1* value is the adjusted value extracted from the daily rainfall data, as used in Hypothesis 6, and thus design storms based on L-moments estimated using Hypothesis 6 are deemed to be the most reliable in Cluster 3. Based on this assumption and on results from Station 0444540, design storms estimated directly from the digitised rainfall data may underestimate, on average over durations ranging from 2 - 24 h, the true values by as much as 65% at some sites in Cluster 3.

6.5.2 Cluster 6

Nine stations are contained within Cluster 6, of which six are SAWB stations, one is a CSIR station (Jnk 19A) and the remaining two stations (Newlands, Athlone) are operated by the Cape Town City Engineers' Department. All the data for Athlone and Newlands and some of the data for Jnk 19A were digitised by the DAEUN. The data from these three stations had no digitising errors and are assumed to be relatively reliable, although no control daily rainfall data were available to check the consistency of the data. The *AV-MARE* of design storms estimated at selected sites in Cluster 6 and for durations ranging from 2 - 24 h are shown in Figure 59.

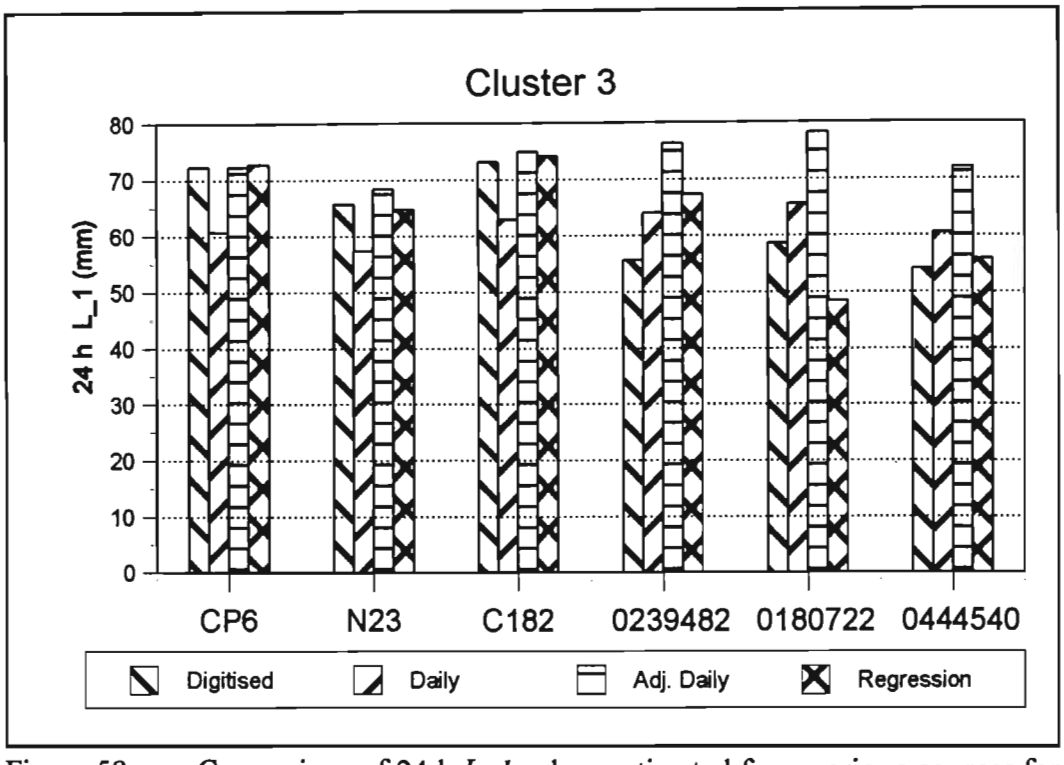


Figure 58 Comparison of 24 h L_1 values estimated from various sources for selected sites in Cluster 3

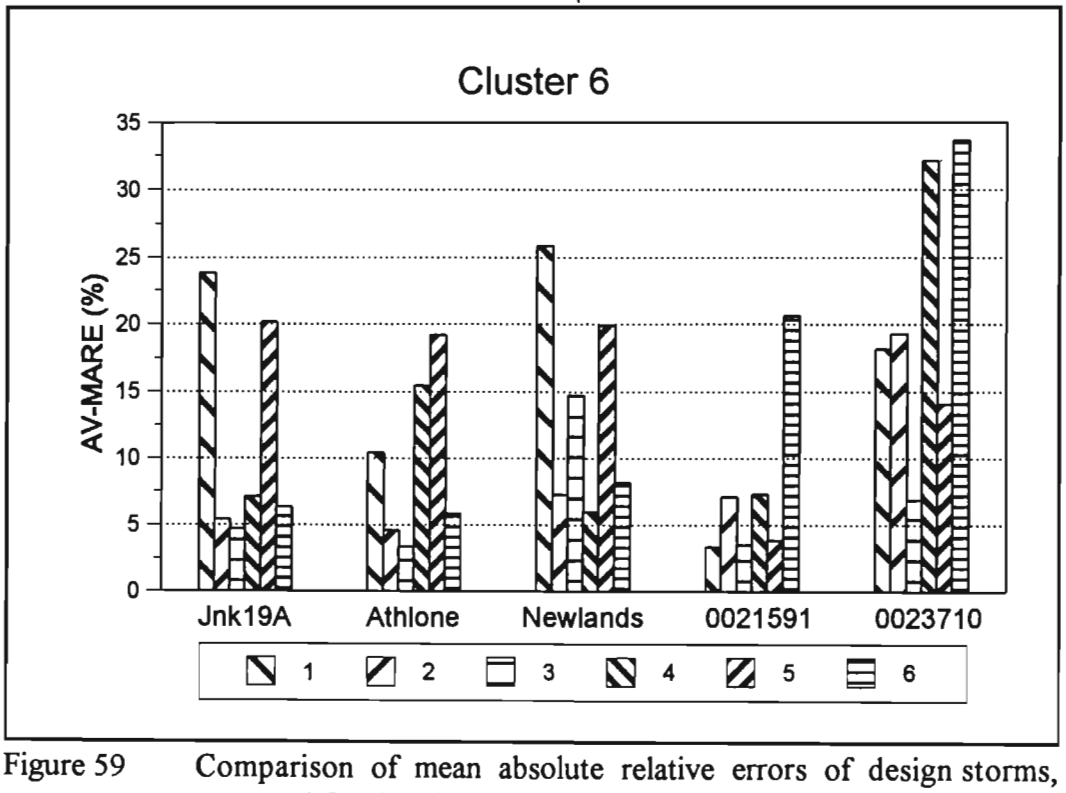


Figure 59 Comparison of mean absolute relative errors of design storms, averaged for durations of 2 - 24 h and for return periods of 2 - 100 years, estimated at selected sites in Cluster 6 for the six hypotheses summarised in Table 48

Similar to results presented for Cluster 3, design storms resulting from the application of Hypotheses 4 and 6 generally result in the smallest deviation from design storms estimated from the digitised rainfall data at most non-SAWB stations and the converse is true at SAWB stations. Again this may be explained by the results contained in Figure 60, which indicate that discrepancies exist between the digitised and daily SAWB rainfall data (0023710, 0021591). It is also noted that Hypothesis 1, which is the simplest of the hypotheses considered, is not valid at most sites considered in Cluster 6. Hence, it is proposed that the estimation of the first and second L-moments using Hypothesis 6 results in the most reliable estimates of design storms in Cluster 6.

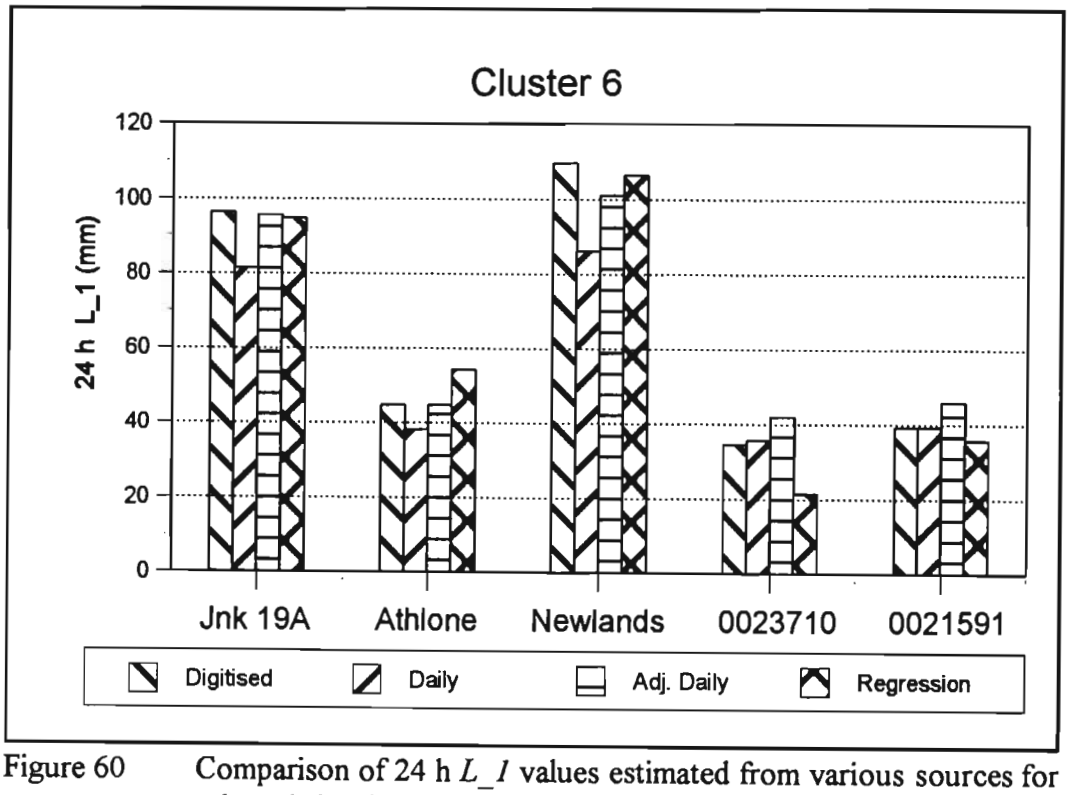


Figure 60 Comparison of 24 h L_1 values estimated from various sources for selected sites in Cluster 6

6.5.3 Selected Other Clusters

Results from selected stations in different geographic and climatic regions are presented in this section. These include stations in the central (Cluster 11), North-Eastern (Cluster 2),

Northern Cape (Cluster 7) and West coast (Cluster 14) regions of South Africa, as indicated in Figure 36.

The CSIR stations at Mokobulaan, which were not used in the clustering procedure or regression analyses, fall geographically on the boundary of Clusters 2 and 11, but are closer to the Euclidian mean of site characteristics of Cluster 11 than of Cluster 2. Hence these stations provide an opportunity for an independent validation of the hypotheses. In addition, SAWB Station 0476398 (Johannesburg International Airport) also located in Cluster 11 is considered. The station selected for Cluster 2 is SAWB Station 0596179 (Skukuza), for Cluster 7 is SAWB Station 0258213 (Drieplotte) and for Cluster 14 is SAWB Station 0106880 (Vrendendal). The *AV-MARE* values for these stations are shown in Figure 61 and a comparison of the *L_1* estimated from various sources for the same stations is shown in Figure 62. The discrepancies at the SAWB between the *L_1* values estimated from the digitised and daily rainfall data again indicate that the most reliable design storms are estimated when Hypothesis 6 is used to estimate short duration L-moments.

The relatively high average deviation of 20% in design storms estimated using Hypothesis 6 at Moko 3A for durations ranging from 2 h - 24 h reduces to an acceptable 12% if the range is reduced to 4 h - 24 h, thus indicating a break in linear scaling for shorter durations. This trend is evident in the observed and estimated L-moments for Moko 3A shown in Figure 62. Similar breaks in linear scaling at durations ranging from 1 to 4 h were also noted at other sites, for example, Ntabamhlope, Cedara, Kokstad, Piet Rietief, Johannesburg, Skukuza and Drieplotte, which are all located in the summer rainfall region where short duration, intense events resulting from thunderstorm activity is the predominant rainfall generating mechanism. The breaks in linear scaling at shorter durations were not evident at, for example, Jonkershoek, Cape Town or Vredendal, which are all in the winter rainfall region and generally experience low intensity, longer duration frontal type rainfall events. An anomaly to this explanation is Cathedral Peak, which is in a summer rainfall region and experiences thunderstorm activity, but scales linearly to durations as short as 5 minutes.

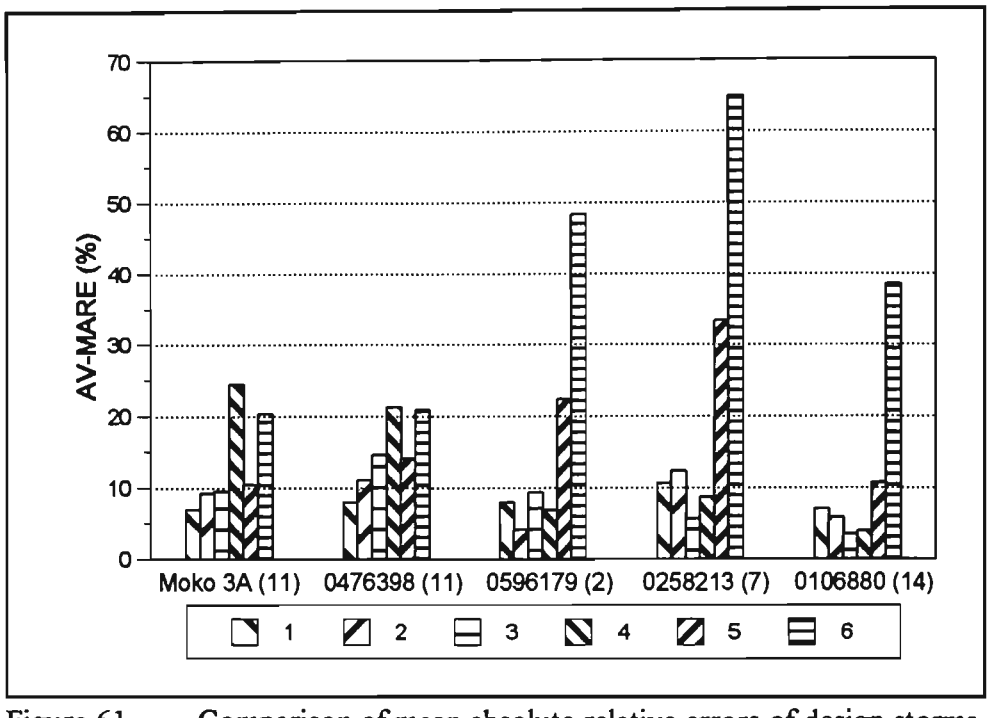


Figure 61 Comparison of mean absolute relative errors of design storms, averaged for durations of 2 - 24 h and for return periods of 2 - 100 years, estimated at selected sites and clusters for the six hypotheses summarised in Table 48
() indicates cluster number

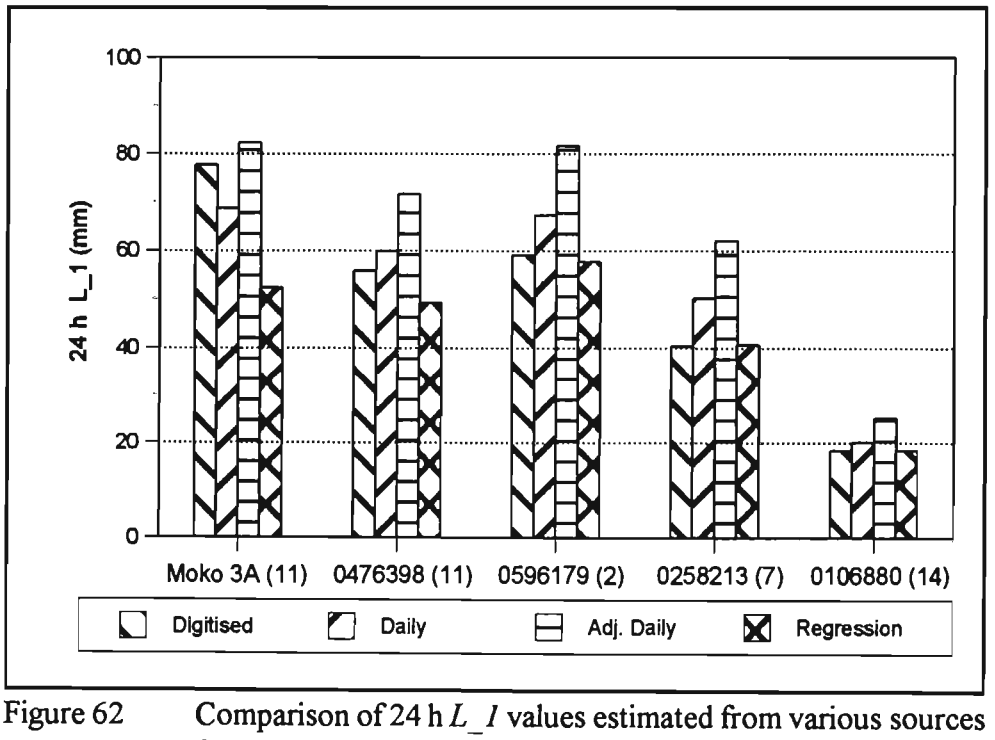


Figure 62 Comparison of 24 h L_1 values estimated from various sources for selected sites and clusters

6.6 CHAPTER CONCLUSIONS

The focus of this chapter has been on estimating design storms for durations < 24 h from daily rainfall data. Six hypotheses for estimating design storm depths for durations up to 24 h were examined. In most cases the stations evaluated were also used in the regression analyses. However, this was unavoidable owing to the limited number of stations which have recording raingauges. Where possible stations have been "hidden" (e.g. N23 and Moko 3A) and used to evaluate the hypotheses.

Of the six hypotheses evaluated, the simplest and intuitively most attractive to adopt is Hypothesis 1 which assumes that the L_1 and L_2 values for durations < 24 h can be derived directly from the at-site 24 h and 48 h values. However, while this hypothesis resulted in good estimates of design storms at some sites, it was shown not to be valid at other sites and is therefore not recommended for general use.

It was evident for stations within a relatively homogeneous cluster, that the slope of the L_1:duration and L_2:duration relationships, when plotted on log scales (i.e. power law relationships), were similar at the sites within the cluster. Hence equations based on multiple linear regression relationships of site characteristics were developed to estimate this slope, and the slope estimated using the regression analyses and site characteristics was termed the Regional Slope (*RS*). The estimation of the slope from site characteristics proved to be feasible for most clusters, except for Cluster 1 and 11. Even in clusters where relatively weak relationships were obtained (e.g. Clusters 7 and 11), reasonable design storm depths were estimated. Thus, given an index point such as the 24 h values used in this analysis, the L_1 and L_2 moments can be estimated for durations shorter than 24 h. Scaling of the site characteristics prior to the regression analysis would have resulted in more reasonable coefficients in the multiple linear regression equation and it is recommended that this should be done in future work of this nature.

The use of the regional average L-moments, which are record length weighted averages of L-moments of the AMS scaled by the at-site mean of the AMS (L_1) for each different

duration (Hypothesis 3), also proved to be successful within the limitations of the reliability of the majority of the digitised data. Thus, in the derivation of the regional average L-moments, the at-site L_1 value for the D h duration is used as an index storm. Observed L_1 values were used to re-scale the regional average L-moments in Hypothesis 3.

Hypothesis 4 re-scaled the 24 h regional average L-moments using 24 L_1 values estimated from site characteristics and regressions developed for each cluster as reported in Chapter 5. D h L_1 values ($D < 24$) were then be derived from the 24 h L_1 value estimated in this manner in conjunction with the *RS*, estimated using the regression equations and site characteristics for the site in question. Again, within the limitations of the data, Hypothesis 4 generally performed well at most sites evaluated. Although Hypothesis 4 incorporates information from the region via the regional average L-moments and *RS* and thus should compensate for limited amounts of unreliable data, the large amount of unreliable data in some clusters used in this study resulted in the compensation by Hypothesis 4 to be relatively ineffective.

It was very evident that from the results presented in this and other chapters that the SAWB digitised and daily rainfall data sets are not consistent. The inconsistency between the 1 day and 24 h L_1 values resulted in Hypothesis 6 being developed. The 24 h L_1 value used in Hypothesis 6 was calculated from the daily rainfall data and converted into a continuous time value using the regionalised ratios developed in Section 6.4. The regionalised slope of the log-transformed L_1:duration relationship was then used in conjunction with the estimated 24 h L_1 value to estimate L_1 for other durations, which are then used to re-scale the regional average L-moments. The GEV distribution was fitted to the estimated L-moments for durations $\leq$ 24 h and hence design storms are estimated for these durations. Hypothesis 6 is thus eminently suitable for application at sites which have daily, but not shorter, duration data available. The use of the daily rainfall data and the regionalised continuous: fixed time L_1 ratios to estimate the true 24 h L_1 values thus attempt to compensate for any bias that may be contained in the 24 h L_1 computed from the digitised data as a result of, for example, missing data either caused by instrument malfunction or incorrect digitisation of charts. The use of Hypothesis 6 indicates that design events

estimated directly from at-site digitised rainfall data obtained from the SAWB would, at some sites, have underestimated the "true" design value by up to 65% on average over durations ranging from 2 - 24 h.

The use of regional average L-moments, particularly when scaled as in Hypothesis 6 with a better estimate of the true 24 h *L_1*, performed well in all clusters. In particular, the use of a regional record length weighted T3 (third L-moment ratio ≡ skewness) value, as the third moment for the fitting of the GEV distribution resulted in reasonable design rainfall estimates at all sites.

Hypotheses 4 to 6 assume that the L-moment:duration relationship is linear when plotted as log-transformed values. This power law function appears to hold true for most clusters over the range from 1 to 24 h. However, a change in the linear relationship at durations ranging from 1 to 4 h was noted at most sites which experience summer rainfall (e.g. Ntabamhlope, Cedara, Kokstad, Mokobulaan and Drieplotte), where thunderstorms are the predominant rainfall generating mechanism. In the winter rainfall region (e.g. Jonkershoek, Cape Town and Vredendal), where frontal rainfall systems predominate, the deviation in linear scaling at a particular duration is not as marked. Although deficiencies in the temporal resolution of the rainfall measurement and digitisation processes cannot entirely be discounted as the cause of the change in linear scaling, it is postulated that the phenomenon is mainly the result of the predominant rainfall generating system. The durations at which the breaks occur at a particular site are hypothesised to be related to the typical duration of thunderstorm activity.

Regional ratios of 24 h : 1 day *L_1* values were used to estimate the 24 h value from the daily rainfall data for each site in each cluster. When the standard error of the mean ratios for each cluster are considered, it is noted that the mean value (=1.20) for all clusters falls within one standard deviation of the mean value for all clusters. Hence a generalised value of 1.20 may be adopted for use in South Africa.

Hypotheses 4 to 6 assume that the slope of the log-transformed L-moment:duration relationship used is correct even though it has been pointed out that the majority of the SAWB digitised rainfall data were not reliable, as result of numerous digitising errors and inconsistencies between the digitised and daily rainfall data. The limited amount of non-SAWB digitised rainfall data resulted in the use of some SAWB data in the regional analyses to estimate the *RS*, although it is conceded that some of this data was unreliable. It was shown in Chapter 2 that the errors in the daily totals of rainfall computed from the digitised database occurred over a wide range of values. It is probable that the wide range of event totals where errors occurred is associated with a wide range of event durations. Thus, it is postulated that the slopes are probably reasonable estimates of their "true" values, as events over a range of durations were affected by the periods of missing data. It is thus assumed that missing events affect all durations equally and thus that the "true" slope and the slope derived from the data are similar.

Of the hypotheses considered in this chapter, Hypothesis 6 performed consistently well at sites where no discrepancies were noted between the digitised and daily rainfall data. At sites where inconsistencies were noted, it is postulated that Hypothesis 6 compensates for deficiencies in the digitised data. Thus Hypothesis 6, which combines a regional index value approach to design storm estimation and the scaling properties of the extreme rainfall events, is recommended for estimating design storms in South Africa for durations ranging from 2 h to 24 h. However, Hypothesis 6 should be used with caution for durations < 2 h and further research into estimating design storms for these shorter durations is recommended.

Hypothesis 6 can only be applied at sites which have daily rainfall data. It is recommended that regional relationships be developed to estimate the at-site 1 day L_1 value, computed from the daily rainfall data, as a function of site characteristics as reported in Section 6.3 for the 24 h L_1 values, which were computed from the digitised rainfall data. This relationship in conjunction with the regionalised 24 h : 1 day L_1 ratios and *RS* would enable reliable estimation of design storms for durations ≤ 24 h at any site in South Africa.

In this chapter, regional average L-moments have been combined with the power law relationship between the first and second order L-moments and duration to give a technique for estimating short duration design rainfall values at ungauged sites or at sites where only daily rainfall data are available. In Chapter 7, the use of stochastic intra-daily rainfall models to estimate short duration design rainfall values is investigated.

CHAPTER 7

MODELLING RAINFALL AND ESTIMATING SHORT DURATION DESIGN STORMS IN SOUTH AFRICA USING THE BARTLETT-LEWIS RECTANGULAR PULSE MODEL

The Bartlett-Lewis (BL) stochastic rainfall models described in Chapter 3 were applied to rainfall data from various locations in South Africa. In this chapter the methodology of determining and optimising the parameters for the models, measures of performance and the results from applying the models to selected stations are presented. The performances of both the Modified Bartlett-Lewis Rectangular Pulse Model (MBLRPM) and the Bartlett-Lewis Rectangular Pulse Gamma Model (BLRPGM), as described in Chapter 3, were assessed for various sets of historical moments used to determine model parameters. The assessments include comparisons between observed and both analytical as well as simulated moments and between design rainfall depths computed from the observed data and from the synthetic rainfall series generated by the models.

In addition to establishing whether the performances of the models were adequate, and in the light of the low reliability of much of the SAWB digitised rainfall data, the focus was also on determining model parameters using readily available daily rainfall values, and on inferring shorter duration statistics using statistics computed from the daily data. Most of the selected case studies presented use data from sources considered reliable and/or which were digitised by the DAEUN. The locations of the stations used in this chapter are illustrated in Figure 63. Within the limits of available reliable data, the models were evaluated in different regions of the country. In regions where no reliable data were available, data which were deemed to be of low reliability were used to illustrate some of the inconsistencies in the data.

The method of estimating the parameters for the models is described in Section 7.1. A goodness-of-fit index and sets of moments to be used for parameter estimation are proposed in Section 7.2 and the performance of the models in terms of moments, event characteristics

and extreme values are evaluated at selected sites in Sections 7.7 and 7.8 for parameters determined using the different sets of moments. Section 7.3 addresses the estimation of monthly moments from the observed data. In order to estimate the parameters of the models at sites which have only daily rainfall data, a technique was developed to estimate short duration variances from the daily data, and this technique is explained in Section 7.4. One of the problems noted with the use of the BLRPMs was the difficulty in identifying model parameters, and the correlation between model parameters is investigated in Section 7.5 and based on these correlations, a parameter search strategy was developed as detailed in Section 7.6. The performance of the models with respect to the temporal distribution of storms is evaluated at selected sites in Section 7.9. In order to identify better parameters for the models the results of various parameter optimisation strategies are presented in Section 7.10. In two interesting case studies presented in Section 7.11, the problem of estimating design rainfall depths from a short period of record is addressed.

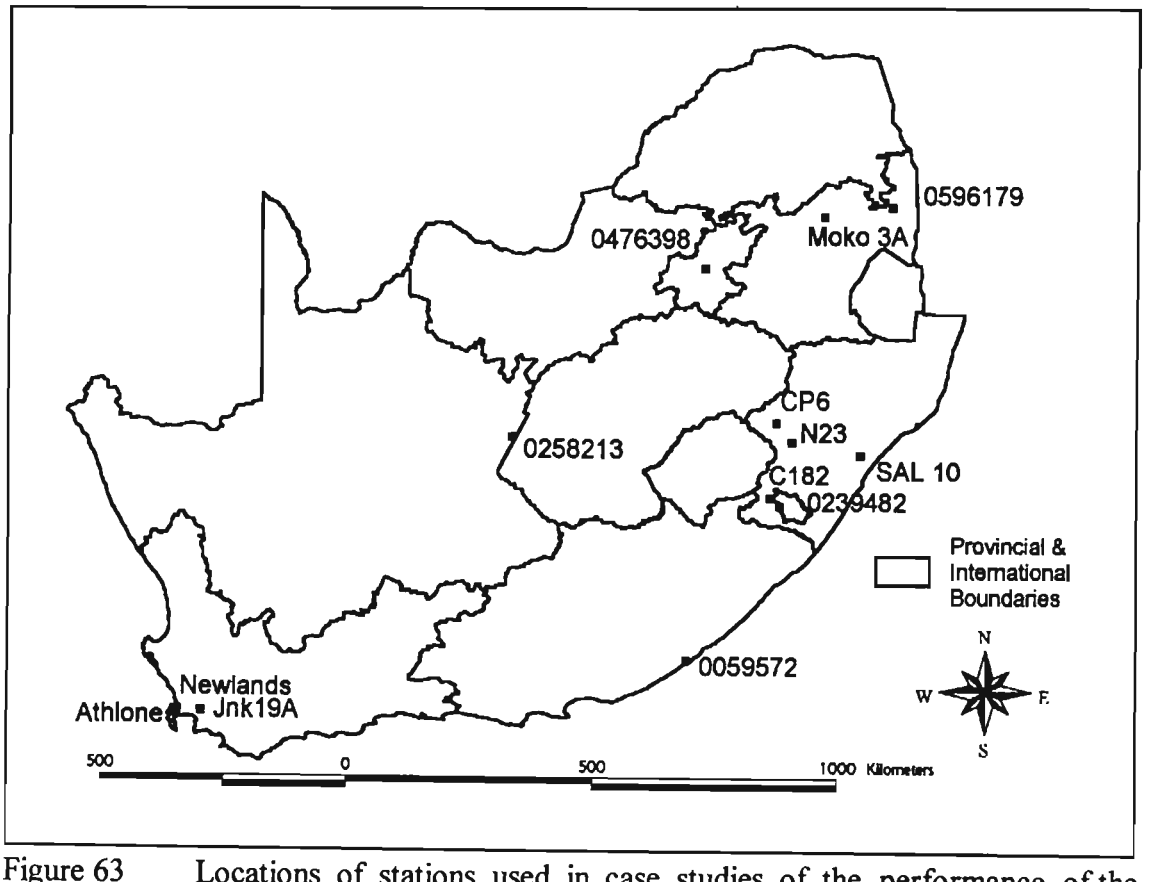


Figure 63 Locations of stations used in case studies of the performance of the MBLRPM and the BLRPGM

7.1 PARAMETER ESTIMATION

The parameters for the models were determined by the method of moments which equates moments computed from observed data (historical moments) with the equivalent analytical expressions of the moments derived for the model. The resulting set of non-linear equations were solved using a quasi-Newton routine to minimise the objective function given in Equation 54 and repeated here as Equation 73, constrained such that the parameters were ≥ 0 .

$$Z = \min\left[\sum_{i=1}^{N} W_i \left(\frac{F_i(X)}{F_i'} - 1\right)^2\right] \qquad ...73$$

where

X = parameter vector with 6 elements for the MBLRPM and 7 elements for the BLRPGM,

$Z(X)$ = goodness-of-fit statistic or residual of least square function,

$F_i(X)$ = model expression for statistic i at specified level of aggregation (duration) computed using parameter vector X,

F_i' = statistic i estimated from historical data at the same level of aggregation,

m = number of statistics and different levels of aggregation used in parameter determination, and

W_i = weight assigned to i-th statistic (set = 1 for all statistics in this study).

The parameters were transformed as shown in Equation 74 such that each parameter was constrained to fall within defined ranges. The transformation generally aided the estimation of the parameters when the range of the transformed values was limited to (0:1), i.e. $YMAX_i = 1$ and $YMIN_i = 0$.

$$Y_i = \left[\left(\frac{X_i - XMIN_i}{XMAX_i - XMIN_i}\right) \times (YMAX_i - YMIN_i)\right] + YMIN_i \qquad ...74$$

where

Y_i = transformed value for i-th parameter (X_i),

$YMAX_i$ = required maximum transformed value for i-th parameter (usually 1),

$YMIN_i$ = required minimum transformed value for i-th parameter (usually 0),

$XMAX_i$ = maximum allowable value for i-th parameter, and

$XMIN_i$ = minimum allowable value for i-th parameter.

7.2 SELECTION OF MOMENTS

As shown in Table 10 in Chapter 3, the choice of the combinations of statistical moments used in the estimation of parameters affects the values of the parameters and could also influence the performance of the model. Hence a comparison was made between the statistical moments and other storm characteristics (e.g. dry probability, event duration and number of events) computed from the observed data (historical moments) and those computed using the estimated parameters and derived moment expressions (analytically derived moments), both at the levels of aggregation of the moments used in the estimation of the parameters and at other levels. A goodness-of-fit statistic was computed as the deviation between the analytical and historical moments, expressed as a percentage of the historical moments for different levels of aggregation and moments and averaged over all months as shown in Equation 75.

$$GOF = \frac{1}{(12 \times N_L \times N_M)} \times \sum_{m=1}^{12} \sum_{i=1}^{N_L} \sum_{j=1}^{N_M} 100 \times \left(\frac{\left| A_{(m,i,j)} - H_{(m,i,j)} \right|}{H_{(m,i,j)}} \right) \qquad ...75$$

where

GOF = goodness-of-fit mean scaled absolute deviation (%),

$A_{(m,i,j)}$ = analytical moment for month m, i-th aggregation level and j-th moment,

$H_{(m,i,j)}$	=	historical moment for month m, i-th aggregation level and j-th moment,
N_L	=	number of aggregation levels used, and
N_M	=	number of moments used.

The above *GOF* was computed for different sets of moments in order to establish an optimum set to use in the derivation of model parameters. Two approaches were used in the selection of sets of moments to use. In the first approach the *GOF* was evaluated assuming that reliable short duration rainfall data were available and thus moments for any level of aggregation could be used in the composition of parameter sets. The sets of moments evaluated by this approach are termed Set 1 in Table 51. The second approach attempted to derive the model parameters based only on moments and storm characteristics that could be derived or estimated from the daily rainfall data and are denoted as Set 2 in Table 51. Thus the 24 h and 48 h values referred to in Table 51 are derived from the digitised data for Parameter Set 1 and from the daily rainfall data for Parameter Set 2. The method of deriving the variance for durations < 24 h from the daily rainfall data, as required in Set 2f, is outlined in the Section 7.4. The computation of moments from the data is discussed in Section 7.3.

7.3 ESTIMATION OF MOMENTS

In the literature two approaches have been adopted in the estimation of moments from the historical data. One option is to pool the data for each calendar month and to calculate the moments from the pooled data. The second approach computes the moments from the individual months of data and then pools the moments for each calendar month. Pooling the data into a continuous series could result in some erroneous moments (e.g. variance and autocorrelation) as a result of the moments computed for the period from the end of one month to the beginning of the next month. Problems are also encountered in the computation of the autocorrelation when periods of missing data are encountered in the pooled data. In the pooled moments approach, the moments for the month are excluded if

any part of the month has missing data. Hence the pooling of moments approach was adopted in this study, although the differences in the moments computed using the two approaches were generally found to be small.

Table 51 Definition of sets of statistics used for estimating model parameters
( ) indicates that the moment was used for the BLRPGM only
[] indicates values are estimated from daily rainfall data

Set No.	Level of Temporal Aggregation of Moment / Event Characteristics Used (h)			
	Mean	Variance	Lag-1 Auto-Correlation	Dry Probability
1a	1	1, 24	1, (24)	1, 24
1b	1	1	1, 24	1, 24
1c	1	1,24	1,24	1
1d	1	1, 6	1, (6)	1, 24
1e	1	1, 6	1(6)	1, 6
1f	24	1, 6, 12, 24, 48	24	24, 48
1g	1, 6, 12, 24	1, 6, 12, 24	1, 6, 12, 24	1, 6, 12, 24
2a	24	24, (48)	24, 48	24, 48
2b	24	24, 48	24	24, 48
2c	24	24, 48	24, 48	24
2d	24	24, 48	24, 48	24, 48
2e	24, 48	24, 48	24, 48	24, 48
2f	24	[1, 6, 12], 24, 48	24	24, 48

A problem encountered with the digitised rainfall data was the apparent digitisation of spurious periods of very low intensity rainfall. The linear interpolation between adjacent data points within the breakpoint digitised rainfall data can result in very small amounts of apparent rainfall when totals of rainfall for fixed time increments (e.g. 15 minutes) are

extracted from the data. For example, if two consecutive digitised points have a time difference of 24 h between them and have a slightly different rainfall depth, then the linear interpolation between data points would result in the extraction of a small amount of rainfall for each of the intervals within the 24 h period and would thus appear as continuous rainfall in the extracted data. Hence apparent rainfall rates of less than 1 mm per day or 0.01 mm per 15 min increment were assumed to be periods with zero rainfall.

7.4 ESTIMATION OF VARIANCES FOR SHORT DURATION RAINFALL

Analytically derived moments matched the historical moments better when historical moments for durations shorter than 24 h were included in the set of moments used to estimate the model parameters. Marked improvements in analytically derived moments resulted when second-order moments for shorter durations (1 to 24 h) were used in the estimation of parameters. Hence, in the absence of short duration data as assumed for Set 2 moments, which were based on daily rainfall data, or when the short duration rainfall are considered unreliable, it is necessary to estimate the shorter duration moments. Cowpertwait *et al.* (1996b) estimated the variances of rainfall for durations shorter than 24 h from the variances of daily rainfall totals, using a regionalised regression approach between the shorter duration and daily variances. In this study, insufficient reliable short duration data were available to estimate regional relationships. Hence an alternative approach had to be devised.

It was noted at sites where the short duration rainfall data were considered to be reliable, that the relationship between variance and duration, when plotted on a log scale, is nearly linear. This is depicted in Figure 64 for selected stations from different climatic regions and for selected months. Hence, assuming a linear relationship on a log-scale between variance at a particular aggregation level and duration, the variance for any duration can be estimated given the daily rainfall data. The results of estimating the variance for durations shorter than 24 h from the variance of 1 and 2 day daily rainfall data are shown in Figure 65 for selected stations and include results from all calendar months. As shown in Figure 65 by the

deviation of the estimated values from the 1:1 line, it was found that the estimated variances generally exceed the observed variances for values < 1 mm^2. The variance of hourly data is generally < 1 mm^2 for most stations. Hence the estimation method is deemed to be suitable for durations ≥ 1 h. Thus this method was used to estimate the historical variance for durations shorter than 24 h from daily rainfall data and enables the estimation of the historical moments in Set 2f to be derived entirely from daily rainfall data.

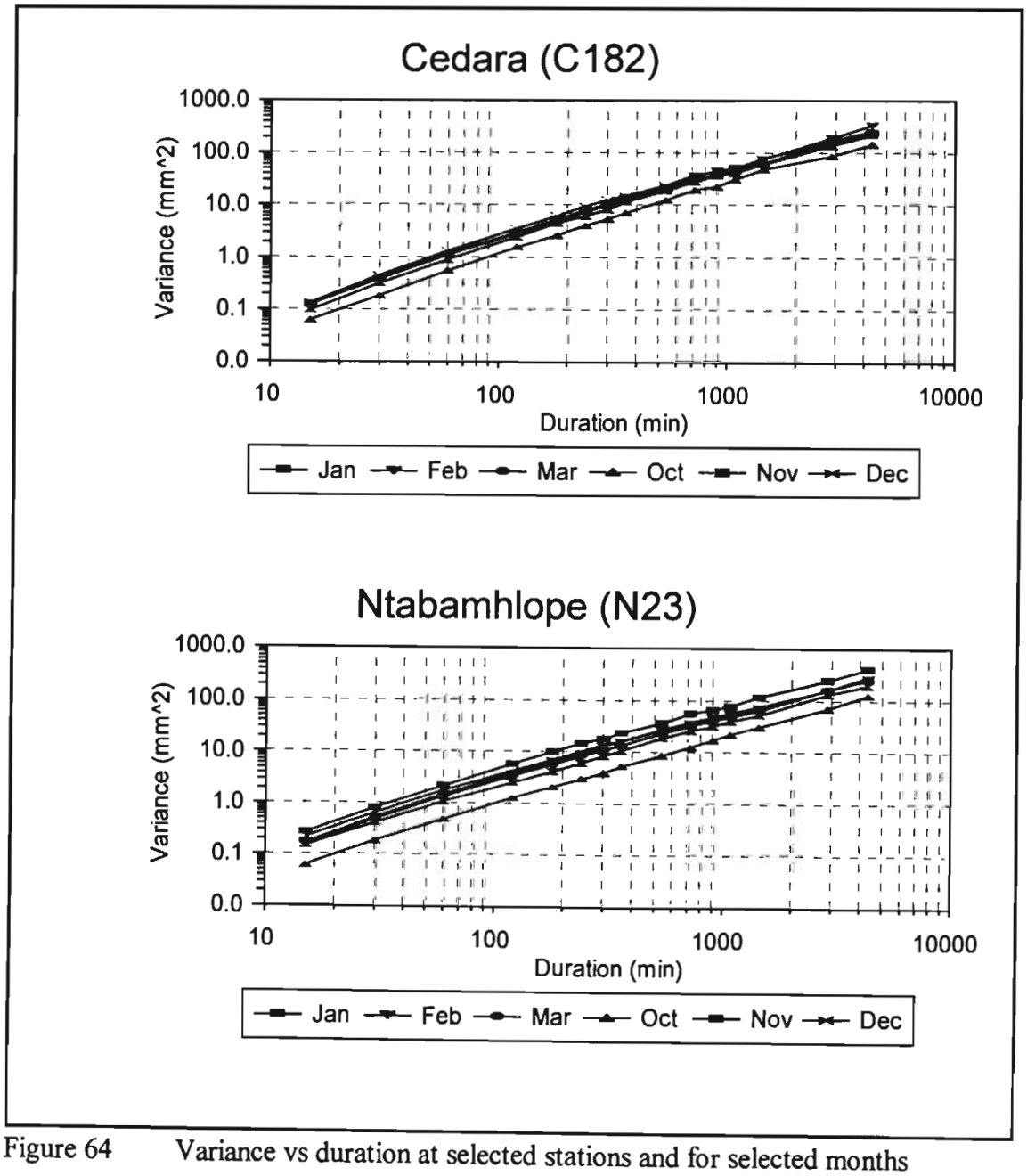


Figure 64 Variance vs duration at selected stations and for selected months

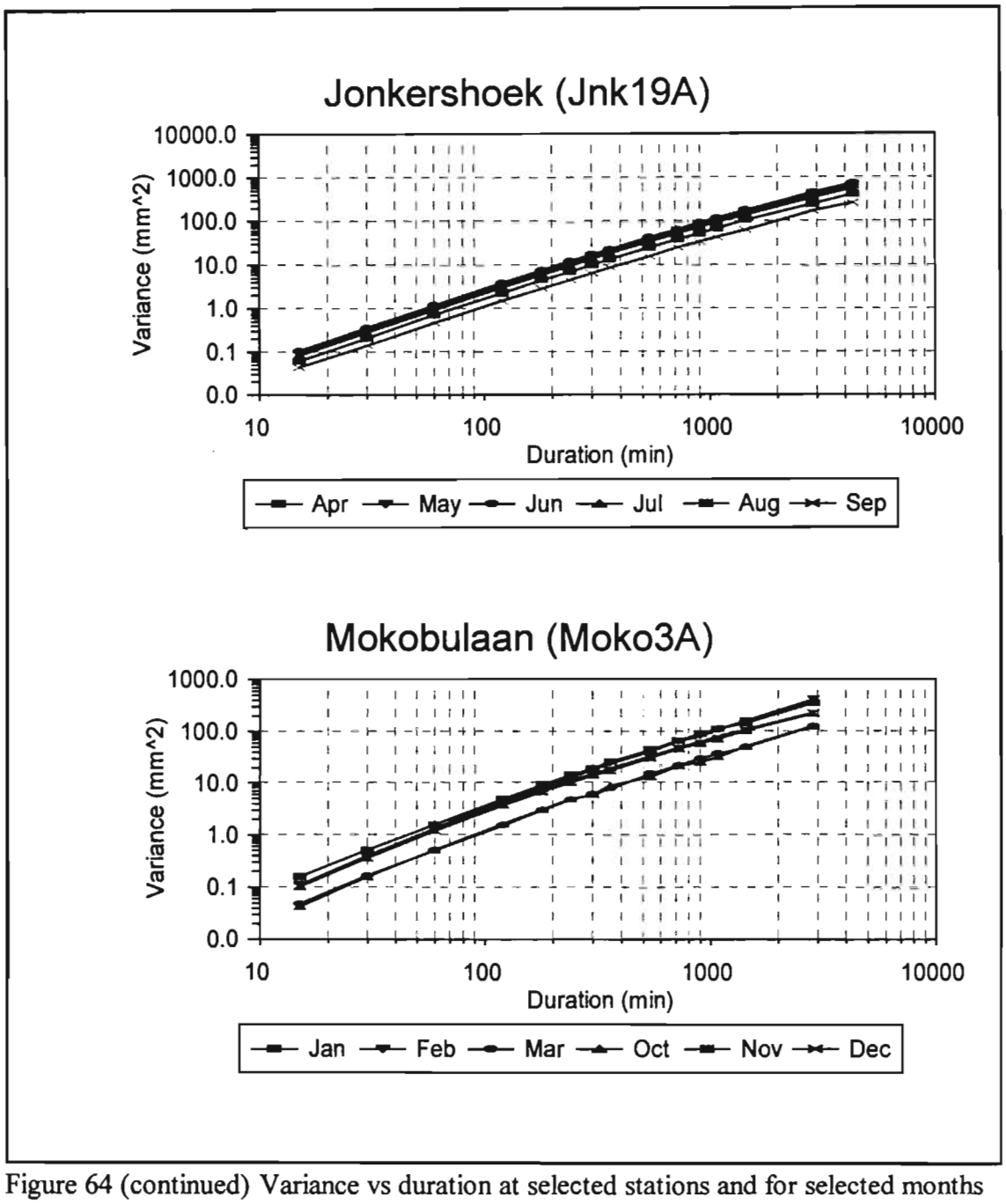


Figure 64 (continued) Variance vs duration at selected stations and for selected months

7.5 PARAMETER CORRELATION

The constrained minimisation of $Z(X)$ Equation 73 generally resulted in a satisfactory solution with the constraints on the parameters set to values such that the physical attributes of the parameters, such as inter-storm and storm duration, were realistic and/or to ensure

that the parameters generally fell within the bounds of parameters reported in the literature. However, on occasion with particular sets of historical moments at some sites, and generally with the Set 2 moment combinations, difficulties were encountered and the minimum of the objective function was frequently located at the limits set for one or more of the parameters. In addition, the relationships between the parameters of the models are not explicit and the quasi-Newton minimisation procedure gives no confidence interval to the estimated parameters.

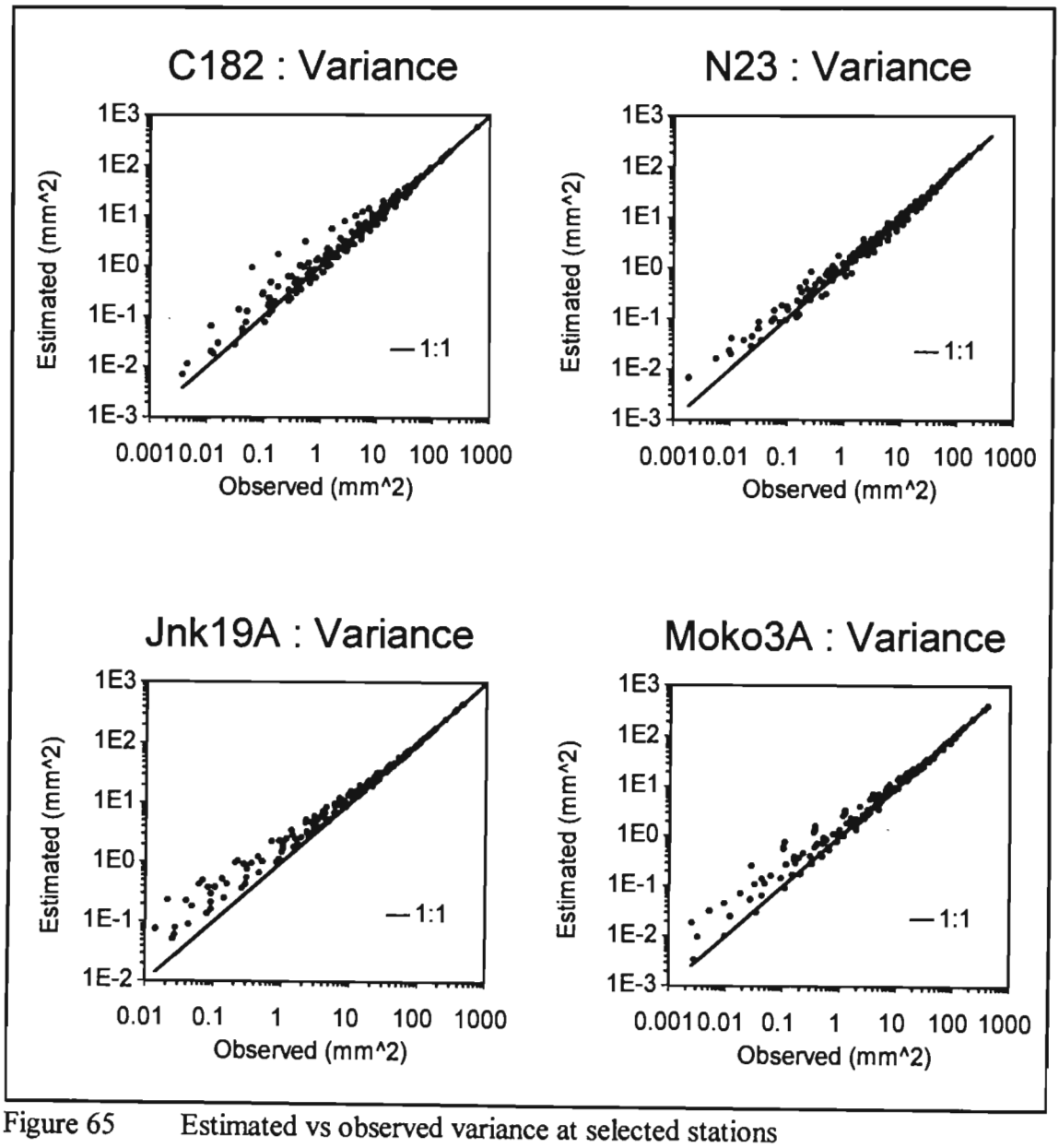


Figure 65 Estimated vs observed variance at selected stations

The sum-of-squares function in Equation 73 can be re-written as a set of m simultaneous equations with n parameters in matrix notation (matrix and vectors shown in bold typeface), as shown in Equation 76.

$$Z(\boldsymbol{X}) = \sum_{i=1}^{m}\left[r_i(\boldsymbol{X})\right]^2 = \boldsymbol{r}^T\boldsymbol{r} \qquad \text{....76}$$

where

$$\boldsymbol{r}^T \quad = \quad [r_1(\boldsymbol{X}), r_2(\boldsymbol{X}), ..., r_m(\boldsymbol{X})],$$

in which

$$r_i(\boldsymbol{X}) = \left(\frac{F_i(\boldsymbol{X})}{F_i^{'}} - 1\right)$$

and the first and second order derivatives of $Z(\boldsymbol{X})$ can be derived (Fletcher, 1987) as

$$\boldsymbol{Z}^{'}(\boldsymbol{X}) = 2\boldsymbol{A}\boldsymbol{r} \qquad \text{....77}$$

$$\boldsymbol{Z}^{''}(\boldsymbol{X}) = 2\boldsymbol{A}\boldsymbol{A}^T + 2\sum_{i=1}^{m} r_i\nabla^2 r_i \qquad \text{....78}$$

where

$$\boldsymbol{A}(\boldsymbol{X}) = \left[\nabla r_1, \nabla r_2 ..... \nabla r_m\right] \qquad \text{....79}$$

is the n x m Jacobian matrix the columns of which are the first derivative vectors ∇r_i of the components of $\boldsymbol{r}$ ($A_{ij} = \partial r_j / \partial X_i$), i.e.

$$\nabla r_i \quad = \quad [\partial r_i / \partial X_1, \partial r_i / \partial X_2, ..., \partial r_i / \partial X_n]^T \qquad \text{...80}$$

The Hessian matrix of second partial derivatives is

$$\boldsymbol{H}(\boldsymbol{X}) = \left[\nabla^2 r_1, \nabla^2 r_2 ..... \nabla^2 r_m\right] \qquad \text{...81}$$

where

$$\nabla^2 r_i = [\partial^2 r_i / \partial X_i \partial X_1, \partial^2 r_i / \partial X_i \partial X_2, ...,\partial^2 r_i / \partial X_i \partial X_n]^T . \quad ...82$$

When $Z(X)$ is minimised, the residual r_i values are generally small and hence the second term in Equation 78 can be ignored. Assuming that the residual (r_i) values are normally distributed with variance σ^2 and using this approximation, the variance-covariance matrix (V) may be estimated according to Fletcher (1987) as

$$V = (AA^T)^{-1}\sigma^2 \quad ....83$$

and the variance estimated as

$$\sigma^2 = \frac{Z(X')}{m-n} \quad ....84$$

where $Z(X')$ is the maximum likelihood sum of squares obtained by minimising $Z(X)$, m is the number of equations and n is the number of parameters.

The diagonal of the V matrix corresponds to the variance of the parameters and the off-diagonal elements correspond to the covariance between the parameters. Hence the correlation coefficient between parameter i and j may be computed as $\rho_{ij} = V_{ij}/(\sigma_i \sigma_j)$ where V_{ij} is the element in row i and column j in V (Stuart and Ord, 1987).

The variance-covariance matrix V may also be estimated from the Hessian (H) matrix for maximum likelihood functions as was performed by Woolhiser and Pegram (1979). In the case of least squares estimates, according to Fletcher (1987), the variance-covariance matrix V may also be estimated from the Hessian (H) matrix as

$$V = \sigma^2(\frac{1}{2}H^{-1}) \quad ...85$$

Thus the correlation matrix ($\boldsymbol{R}$) may, according to Woolhiser and Pegram (1979) and (Press *et al.*, 1992), be derived as

$$\boldsymbol{R} = \boldsymbol{S}^{-1}\left[\sigma^2\left(\frac{1}{2}\boldsymbol{H}^{-1}\right)\right]\boldsymbol{S}^{-1} \qquad \text{....86}$$

where $\boldsymbol{S}$ is a square matrix with σ_i (derived from the diagonal of $\boldsymbol{V}$) on the diagonal and the rest of the matrix void.

Using the above relationships, the standard deviation of the parameters can be estimated and relationships between the parameters can be investigated. The variance-covariance matrix was calculated with very similar results using both the Jacobian matrix and the Hessian matrix. For a well determined system ($m=n$), σ cannot be estimated using Equation 84 and hence σ was estimated as $Z(X')$ when $m=n$.

The estimates of the values of the parameters and the results from estimating the Standard Deviation of the estimates (SD), Coefficient of Variation (CV) as SD/estimate and the correlation between parameters of the MBLRPM, computed using Equation 86, are contained in Table 52 for raingauge N23 in the Ntabamhlope research catchments. The parameters of the MBLRPM were determined using moment Set 1b in Table 51 and are referred to as parameter Set 1b. Thus the term "parameter Set 1b" refers to the set of parameters derived for the model using the Set 1b moments referred to in Table 51.

From Table 52 it is evident that there is a high degree of correlation between parameters and that the parameters are not well defined. This is apparent from computing the CV, i.e. the ratio between the SD and parameter value. In particular, the κ, ϕ, υ and μ_x parameters are poorly defined. The most poorly defined parameter is υ and the results of fixing υ at a value determined by the constrained minimisation procedure, thus reducing the parameter space by one, are contained in Table 53.

Table 52 Estimated parameters, correlation matrix and goodness-of-fit of the MBLRPM, fitted to data for January from N23, using parameter Set 1b

Parameter				Correlation Coefficient						Z
Name	Value	SD	CV	λ	κ	ϕ	υ	α	μ_x	
λ	0.0380	0.0057	0.1511	1.0000	0.7702	0.8060	-0.8556	-0.8671	0.7819	0.0062
κ	0.0911	0.2594	2.8455	0.7702	1.0000	0.9900	-0.9571	-0.9319	0.8200	
ϕ	0.0861	0.1113	1.2924	0.8060	0.9900	1.0000	-0.9716	-0.9589	0.8435	
υ	0.9734	2.6474	2.9199	-0.8556	-0.9571	-0.9716	1.0000	0.9952	-0.9378	
α	4.5231	5.3119	1.1744	-0.8671	-0.9319	-0.9589	0.9952	1.0000	-0.9427	
μ_x	10.2520	3.5456	0.3458	0.7819	0.8200	0.8435	-0.9378	-0.9427	1.0000	

Table 53 Estimated parameters, correlation matrix and goodness-of-fit for the MBLRPM, fitted to data for January from N23, using parameter Set 1b and with υ fixed

Parameter				Correlation Coefficient						Z
Name	Value	SD	CV	λ	κ	ϕ	υ	α	μ_x	
λ	0.0380	0.0030	0.0782	1.0000	-0.3240	-0.2067		-0.3077	-0.1145	0.0062
κ	0.0909	0.0750	0.8255	-0.3240	1.0000	0.8779		0.7264	-0.7708	
ϕ	0.0860	0.0264	0.3066	-0.2067	0.8779	1.0000		0.3477	-0.8235	
υ	0.9763	Fixed								
α	4.5290	0.5206	0.1149	-0.3077	0.7264	0.3477		1.0000	-0.2753	
μ_x	10.2483	1.2303	0.1201	-0.1145	-0.7708	-0.8235		-0.2753	1.0000	

The effect of fixing the value of υ in the MBLRPM results in better defined parameters with lower inter-parameter correlations. However, the goodness-of fit (Z) is not improved using

this strategy. A similar analysis to the above was performed for the BLRPGM and the results for a selected month are contained in Tables 54 and 55.

Table 54 Estimated parameters, correlation matrix and goodness-of-fit for the BLRPGM, fitted to data for January from N23, using parameter Set 1f

Parameter				Correlation Coefficient							Z
Name	Value	SD	CV	λ	κ	ϕ	υ	α	ρ	δ	
λ	0.0344	0.0077	0.2223	1.0000	-0.2961	0.6942	-0.9153	-0.9153	0.5929	0.0079	0.0084
κ	0.1350	0.0544	0.4028	-0.2961	1.0000	0.0879	0.3056	0.3063	-0.8712	-0.7240	
ϕ	0.0708	0.0131	0.1857	0.6942	0.0879	1.0000	-0.7138	-0.7144	0.3516	-0.0089	
υ	45.3685	58.1870	1.2825	-0.9153	0.3056	-0.7138	1.0000	1.0000	-0.6176	0.0851	
α	105.1307	86.5885	0.8236	-0.9153	0.3063	-0.7144	1.0000	1.0000	-0.6181	0.0840	
ρ	0.3571	0.1135	0.3178	0.5929	-0.8712	0.3516	-0.6176	-0.6181	1.0000	0.6843	
δ	0.0717	0.0167	0.2325	0.0079	-0.7240	-0.0089	0.0851	0.0840	0.6843	1.0000	

From Table 54 it is evident that there is a large degree of correlation between some parameters of the BLRPGM and that the parameter υ is the least well defined. In this instance the α and υ parameters are completely correlated and fixing either of these parameters will fix the other parameter. The results of fixing υ, and thus reducing the parameter space by one, are contained in Table 55. Similar to the results from the MBLRPM, this strategy results in better defined parameters for the BLRPGM, but does not improve the fit (Z) of the model. A strategy to reduce the parameter space, and thus have more reliable estimates of the parameters of the model, while simultaneously improving the overall fit of the model is investigated in the following section.

Table 55 Estimated parameters, correlation matrix and goodness-of-fit for the BLRPGM, fitted to data for January at N23, using parameter Set 1f and with υ fixed

Parameter				Correlation Coefficient							Z
Name	Value	SS	CV	λ	κ	ϕ	υ	α	ρ	δ	
λ	0.0344	0.0028	0.0817	1.0000	-0.0425	0.1451		0.0110	0.0867	0.2137	0.0084
κ	0.1349	0.0473	0.3503	-0.0425	1.0000	0.4591		0.2558	-0.9113	-0.7905	
ϕ	0.0708	0.0084	0.1188	0.1451	0.4591	1.0000		-0.2668	-0.1621	0.0741	
υ	42.8269	Fixed									
α	99.3731	0.2354	0.0024	0.0110	0.2558	-0.2668		1.0000	-0.1720	-0.3607	
ρ	0.3574	0.0815	0.2281	0.0867	-0.9113	-0.1621		-0.1720	1.0000	0.9405	
δ	0.0717	0.0152	0.2116	0.2137	-0.7905	0.0741		-0.3607	0.9405	1.0000	

7.6 SEARCH STRATEGY FOR IMPROVING MODEL FIT

As shown in the previous section, the effect of fixing one or more of the least well defined parameters is to improve the confidence in the remaining non-fixed parameters, but with no decrease in the goodness-of-fit (Z). In order to determine the optimum value at which to set the fixed parameters(s), a search was performed between user-defined boundaries for the fixed parameter(s). Once the optimum value(s) for the parameter(s) being fixed had been established, the parameter(s) were set to these values and remaining parameters were determined using the constrained minimisation procedure. An example of the constrained minimisation procedure and parameter search is shown in Figure 66 where the least well defined parameter has been established as υ and a constrained minimisation procedure is implemented for each fixed value of υ. In order to determine better defined parameters, the constraints used in the minimisation procedure were such that the mean storm characteristics, as shown in Figure 67, made reasonable physical sense. Based on these and other successful improvements in Z, the search strategy was adopted for all model parameter estimation, with the exception of results in Section 7.10, where additional parameter optimisation techniques are evaluated.

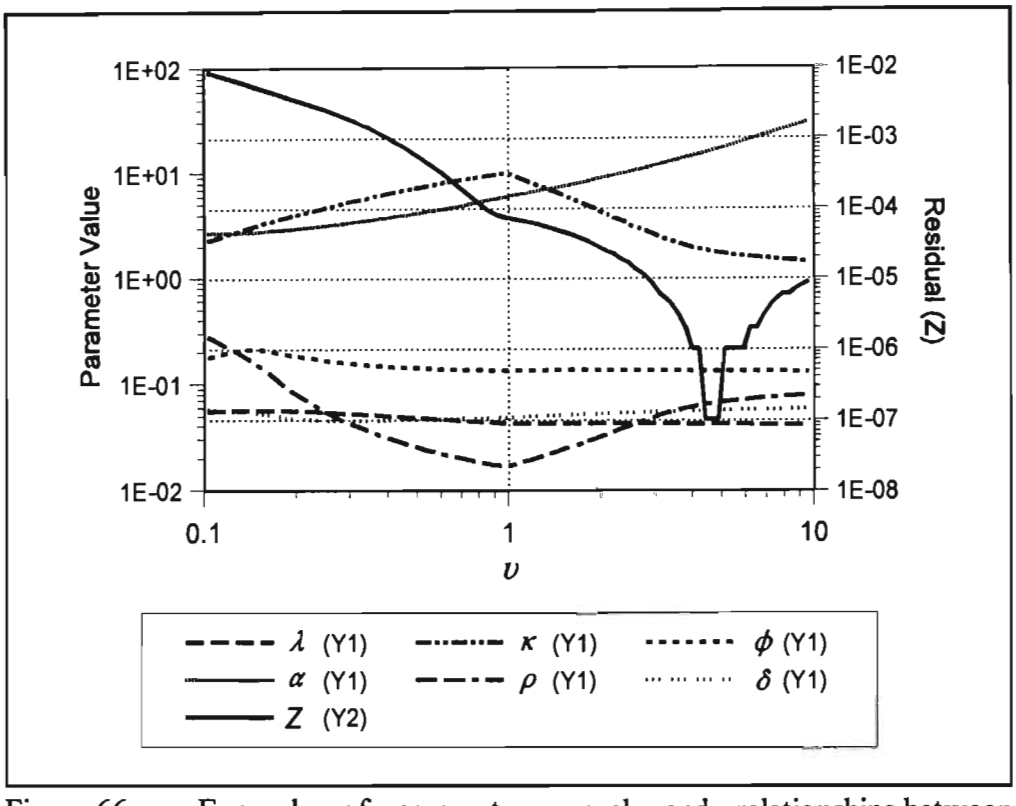


Figure 66 Example of parameter search and relationships between parameters: BLRPGM (Set 1e), Raingauge N23

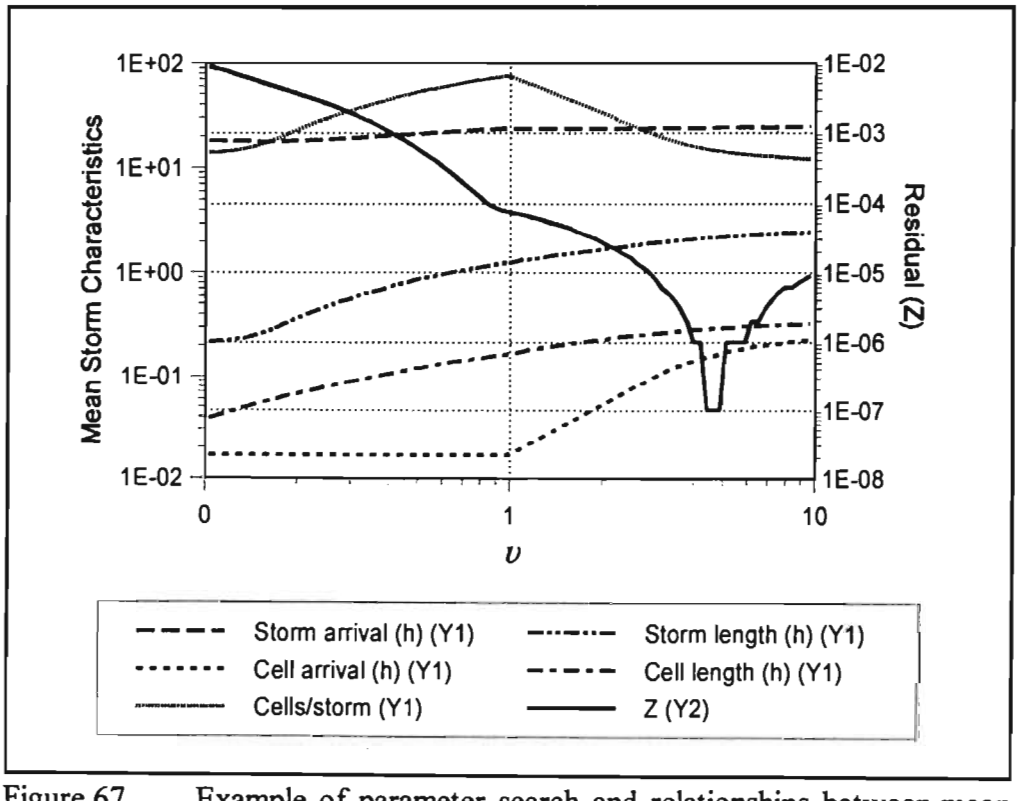


Figure 67 Example of parameter search and relationships between mean storm characteristics: BLRPGM (Set 1e), Raingauge N23

7.7 ANALYTICAL PERFORMANCE

The performance of the two models can be evaluated by comparing model moments and event characteristics with the historical values. Model moments may be computed either by using the analytical expressions for the moments and fitted model parameters, which are termed "analytical" moments, or by using the model to simulate rainfall and compute the "simulated" moments from the synthetic rainfall series by re-sampling. In this section the analytical performance of the models is investigated and the simulated performance is analysed in the following section.

Three *GOF* indices were computed for the analytical moments, using Equation 75. The first, termed "Fit Only", only incorporated the moments at the levels of aggregation used in the determination of the parameters i.e. as per list in Table 51. The second , termed "Lag-1 Only", used the mean, standard deviation, lag-1 autocorrelation, probability of dry periods and the duration and number of wet periods, computed at 16 levels of aggregation ranging from 15 min to 48 h, to compute the *GOF*. The third *GOF* computed was similar to the "Lag-1 Only", but included the lag-2 and lag-3 autocorrelations and is termed "Lag 1-3". As an example, these indices are shown in Figure 68 for both the MBLRPM and BLRPGM fitted to data from raingauge N23 in the Ntabamhlope catchments for the sets of moments used in parameter determination listed in Table 51.

From Figure 68 it is evident that the performances of both the MBLRPM and BLRPGM are affected by the set of moments used in the determination of parameters. Parameter Set 1e gave the best performance for both the MBLRPM and BLRPGM in the scenario that assumed that short duration rainfall data were available. If only daily data were available (i.e. Set 2), Set 2f resulted in the best performance for the MBLRPM when only the lag-1 autocorrelations were considered and similar performance was obtained from Sets 2a, 2d and 2e if the lag-2 and lag-3 autocorrelations were included. Similarly, for the BLRPGM and assuming only daily rainfall data were available, the best parameter set for the Lag-1 Only *GOF* was Set 2f, while Sets 2d and 2e resulted in similar values for the Lag 1-3 *GOF*. The relatively larger Fit Only *GOF* obtained with both models for Sets 2a, 2c, 2d and 2e is

a result of the inclusion of the 48 h lag-1 autocorrelation in these sets and which was negative for some months at raingauge N23. This does not appear to affect the overall performance of the analytical moments of these moment sets. For example, the Fit Only *GOF* of Set 2b, which does not include the 48 h lag-1 autocorrelation, is much smaller than the other Set 2 analytical moments, but the overall analytical moments obtained using Set 2b are not as good as that obtained using the other Set 2 moments.

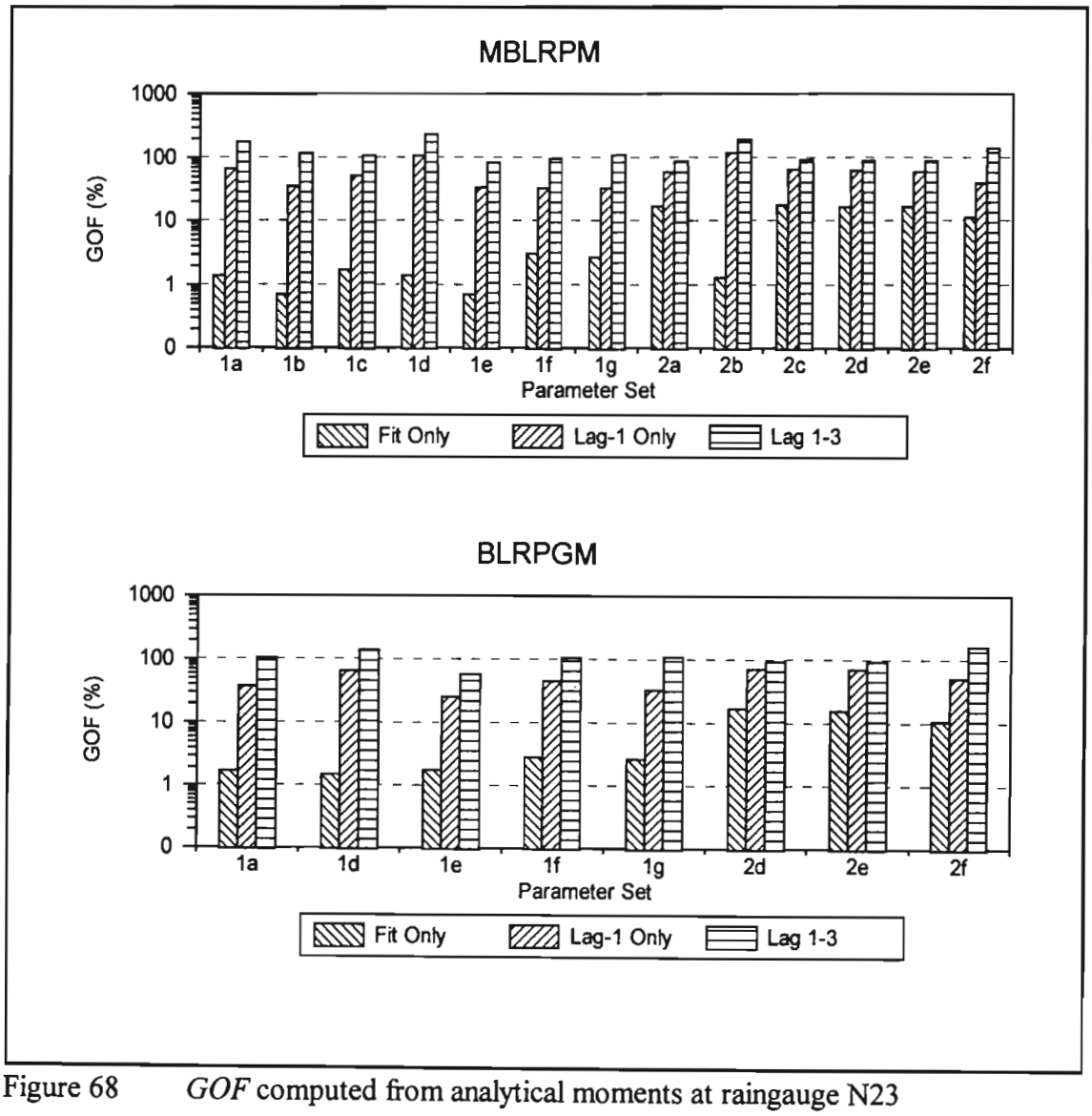


Figure 68 *GOF* computed from analytical moments at raingauge N23

Unexpectedly, the overall performance of the models did not improve when more than the minimum number of moments (Sets 1f , 1g and 2f) were used in the estimation of

parameters. Thus the expected improvement in *GOF* as a result of including more moments to be used for parameter estimation is off-set by the difficulty in estimating parameters with more degrees of freedom, as indicated by larger Fit Only *GOF* for Sets 1f and 1g.

A comparison of analytical and observed moments for selected durations in January at N23 using the MBLRPM and the BLRPGM, both with parameters derived using moment Set 1e, is shown in Figure 69. The relative error is the absolute difference between the analytical and historical moment expressed as a percentage of the historical moment. As illustrated in Figure 69 the analytical moments of the BLRPGM better represent the historical values, particularly for shorter durations. In addition, it is noted that both the mean and probability of no rain are better represented by the BLRPGM over all the range of durations shown.

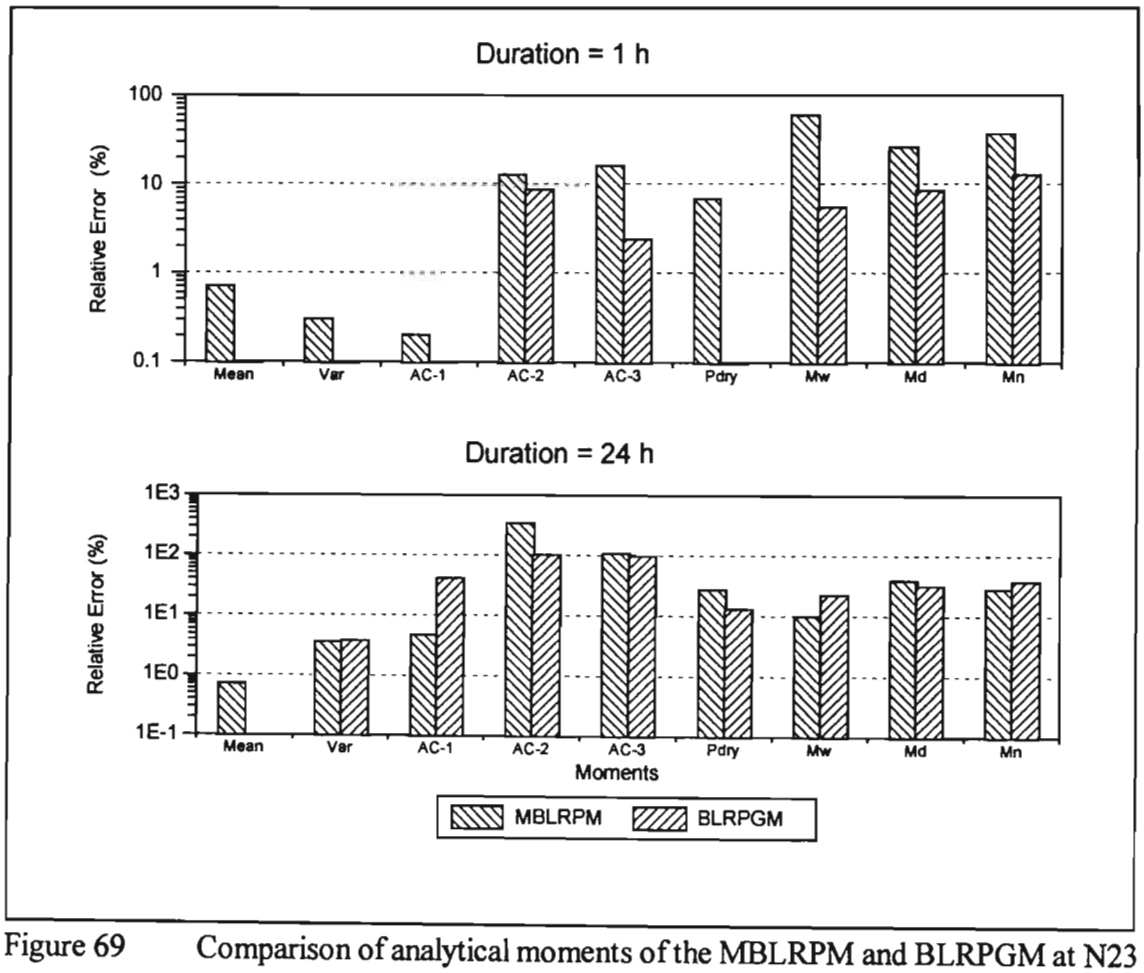


Figure 69 Comparison of analytical moments of the MBLRPM and BLRPGM at N23 during January (Var = variance; AC-n = lag-n autocorrelation; Pdry = dry probability; Mw = event duration; Md = dry duration; Mn = no. of events)

A comparison of the analytical moments using the Lag-1 Only *GOF* is shown for selected stations in Figure 70. For the MBLRPM and assuming that digitised data were available, then parameter Sets 1b, 1f and 1g gave the best fit to the historical values, whereas if only the daily data had been available, then parameter Set 2f resulted in the best fit at the stations shown. Similar fits to the historical values were obtained using the BLRPGM for the Set 1 parameters. However, if only daily rainfall data were available, then Set 2f resulted in the best analytical fit to the historical moments at the stations shown, except for Station C182. A comparison, for the same parameter set, of the fit to the historical moments of the analytical moments computed by the two models indicates that the BLRPGM, despite needing to estimate an additional parameter, generally performs better than the MBLRPM.

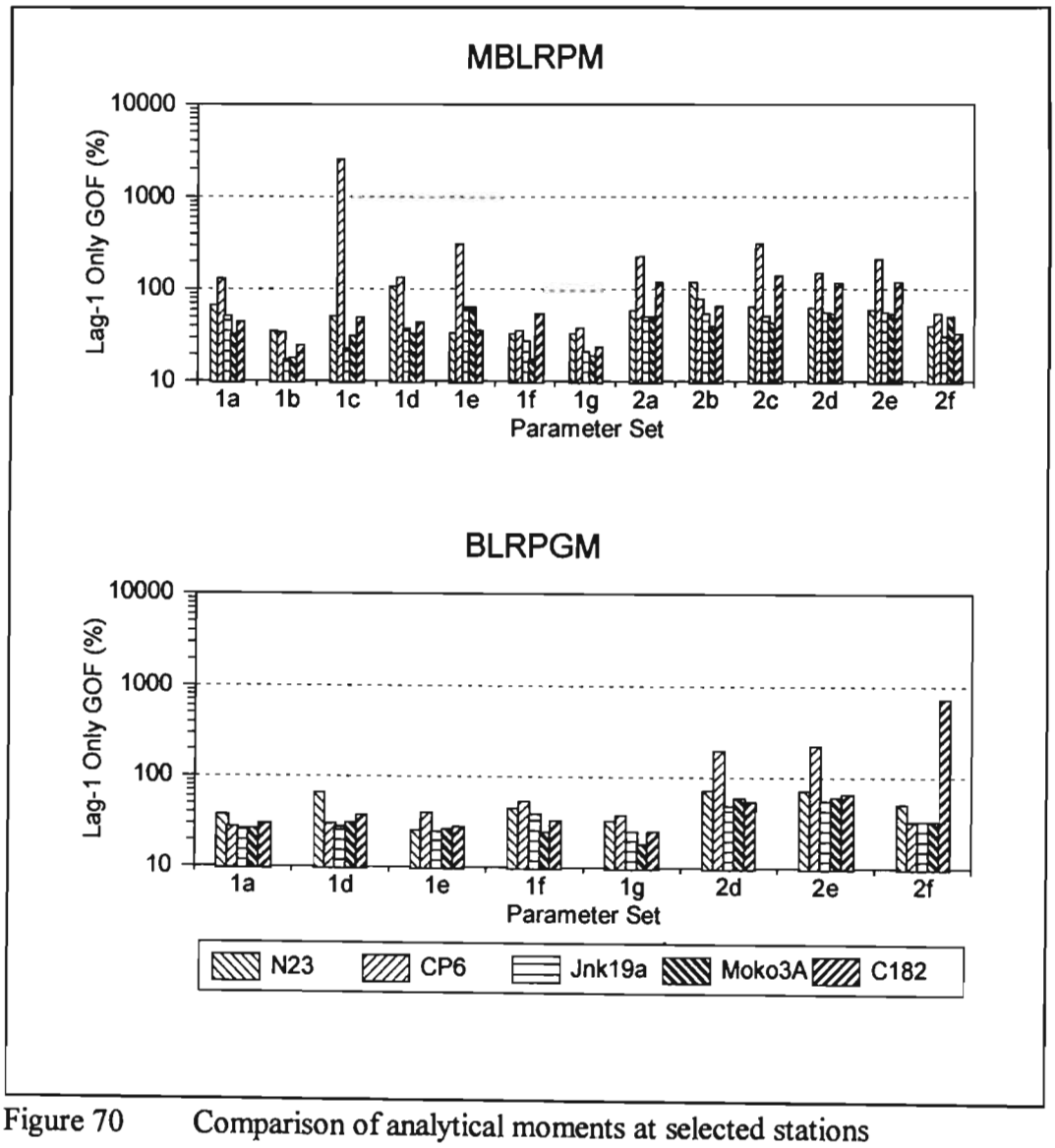


Figure 70 Comparison of analytical moments at selected stations

It has been established that, at the sites considered and which fall in different climatic regions in South Africa, the analytical moments computed using the BLRPGM generally fit the historical moments better than those computed from the MBLRPM, irrespective of which set of moments was used to determine the parameters of the models. Assuming that the short duration digitised data were available then, for the MBLRPM, moment Sets 1b, 1f and 1g resulted in the best fit to historical values when using analytical moments, whereas for the BLRPGM a similar performance was obtained for all the parameter sets used. Hence the fit of the BLRPGM analytical moments to the historical values appears to be less dependent than that of the MBLRPM on the set of moments used to derive the parameters. If only daily rainfall data are available then parameter Set 2f, which includes estimated variances for durations shorter than 24 h, generally resulted in the best fit for both models.

In the following section the simulated performance of the models, with parameters determined using different sets of moments, are examined.

7.8 SIMULATED PERFORMANCE OF THE MODELS

In the previous section the performance of the model was assessed relative to the analytical moments of the model. In order to quantify the simulated performance of the models, moments and other event characteristics computed from the simulated synthetic rainfall series are compared to the equivalent values computed from the observed data in Section 7.8.1. Similarly, design rainfall depths computed from the simulated synthetic rainfall series are compared to the equivalent values computed from the observed data in Section 7.8.2.

For each evaluation at a particular site, one hundred sets of synthetic rainfall series were generated, each with a record length equal to that of the historical data. The performance of the model is assessed using two measures. In the first measure, seven moments and statistics (mean, standard deviation, auto-correlation, dry probability, durations of wet and dry periods and the number of events) of the synthetic data are compared to the corresponding characteristics of the historical data. More emphasis is placed on the second

measure of performance of the model where design storms, computed from the synthetic series, are compared to those computed from the historical data.

The measures of performance of the models are both initially focussed on detailed results using data from raingauge N23 in the Ntabamhlope catchments, and are subsequently generalised and expanded to data from other raingauges in order to lead to some general conclusions.

7.8.1 Moments and Statistics

At each of the selected stations the stochastic variability of the BL models was simulated by generating 100 sets of synthetic rainfall series, each with the same length of record as the observed data, for each of the parameter sets outlined in Table 51. A frequency analysis for each statistic and for each duration was performed on the 100 sets of synthetic rainfall. High -Low bar graphs depicting the observed moments and 25-th and 75-th non-exceedance percentiles of the 100 synthetic data sets are used to graphically depict the adequacy of the models. For example, the results from generating synthetic rainfall series using the MBLRPM, fitted to the data from N23 using parameter Set 1b, are shown in Figure 71. For the moments and statistics shown in Figure 71, the MBLRPM with parameters derived using Set 1b generally simulates the observed values well, particularly for durations longer than 15 min.

In order to objectively assess the performance of the models, the Mean Absolute Relative Error (*MARE*), as calculated in Equation 87, is shown in Figure 72 for the MBLRPM fitted to data from raingauge N23 using moment Set 1b. The number of aggregation levels in Equation 87 (N_L) was set to 10 and the durations used were 2, 3, 4, 5, 6, 9, 12, 15, 18 and 24 h. For the summer months (Oct - Mar), when more than 80% of the rainfall occurs and when the extreme rainfall events usually occur, the *MARE* for the 10 levels of aggregation used are less than 10% for the mean, standard deviation and dry probability.

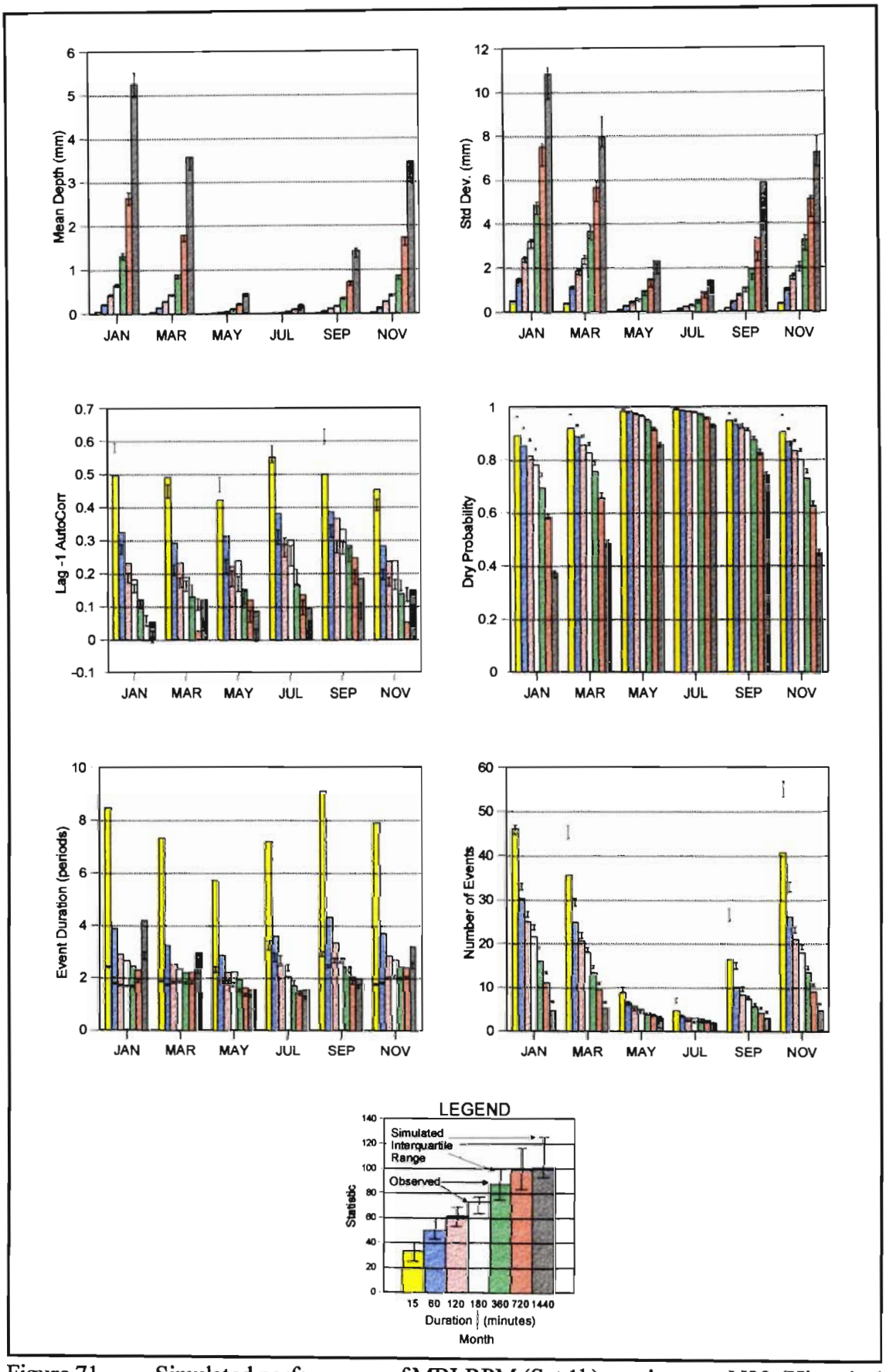


Figure 71 Simulated performance of MBLRPM (Set 1b) at raingauge N23 (Historical values in histograms. Interquartile range of 100 simulations in I-beams)

$$MARE_{(i,j)} = \frac{1}{N_L} \times \sum_{i=1}^{N_L} 100 \times \left(\frac{\left| S_{(i,j)} - O_{(i,j)} \right|}{O_{(i,j)}} \right) \qquad ...87$$

where

$MARE_{(i,j)}$ = mean absolute relative error (%) for month i, and statistic j (%), mean (j=1), standard deviation (j=2), autocorrelation lag-1 (j=3), dry probability (j=4), duration of wet periods (j=5), duration of dry periods (j=6) and number of wet periods (j=7),

$S_{(i,j)}$ = mean j-th statistic computed from the 100 synthetic rainfall series generated for month i,

$O_{(i,j)}$ = j-th statistic computed from observed data for month i, and

N_L = number of aggregation levels used.

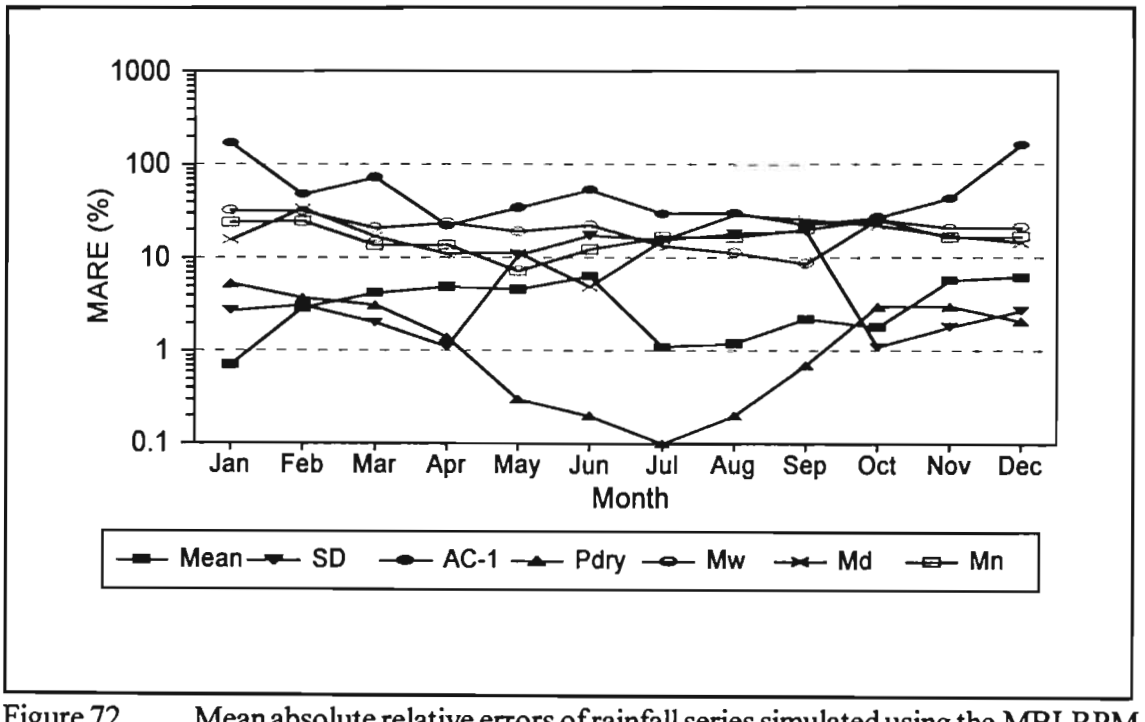


Figure 72 Mean absolute relative errors of rainfall series simulated using the MBLRPM (Set 1b) at raingauge N23 (SD = standard deviation; AC-1 = lag-1 autocorrelation; Pdry = dry probability; Mw = event duration; Md = dry duration; Mn = no. of events)

The *MARE* value shown in Figure 72 reflects the differences between the mean of the statistic computed from the 100 synthetic rainfall series and the corresponding statistic computed from the observed data, and does not reflect the stochastic variation within the 100 values computed from the 100 synthetic series. Thus a frequency analysis was performed on the 100 values for each of the statistics and the percentage of times (*EXC*) the observed statistic fell outside of the 25th and 75th percentile of simulated values was computed. The *MARE* value was adjusted using the *EXC* value for the statistic as shown in Equation 88. In addition, a mean adjusted *MARE* value (*STATS_INDEX*) was computed as the mean of the *MARE* values for individual months to form a composite index for the statistic for the 10 durations considered and for all the rainy season months.

$$STATS\_INDEX_{(j)} = \frac{1}{6} \times \sum_{i=1}^{6} MARE_{(i,j)} \times \left(1 + \frac{EXC_{(i,j)}}{50}\right) \quad ...88$$

where

$STATS\_INDEX_{(j)}$ = performance index for rainy season months for statistic j which includes 10 aggregation levels, and

$EXC_{(i,j)}$ = percentage of times the observed j-th statistic in month i fell outside the range of the 25th and 75th percentiles of 100 values computed from the synthetic series.

The *STATS_INDEX* values for both the MBLRPM and BLRPGM at raingauge N23, with parameters determined using the Set 1 moments, are shown in Figure 73. By comparing the *STATS_INDEX* for different parameter sets for the same statistic, it is evident that the seven moments computed from the synthetic rainfall series generated by the MBLRPM best fit the historical values when moments Sets 1b, 1f and 1g are used to determine the model parameters. These are the same findings as when the analytical moments were considered. Parameter Sets 1e and 1g resulted in the best fit of the simulated moments of the BLRPGM. When the *STATS_INDEX* computed from the MBLRPM and the BLRPGM are compared for the same parameter set shown in Figure 73, it is evident that moments computed from

the synthetic rainfall series generated by the BLRPGM fit the observed moments better than those from the MBLRPM.

Assuming only daily rainfall data to have been available at raingauge N23, the moments computed from the synthetic rainfall series generated by the MBLRPM best fitted the observed values, as shown in Figure 74, when parameter Set 2f was used. For the BLRPGM, very little difference in performance is noted between the Set 2 parameters, although Set 2f performed slightly better than either Sets 2d or 2e. A comparison between the performance of the MBLRPM and BLRPGM at raingauge N23, for the same Set 2 parameters, indicates that the moments computed from the synthetic rainfall series generated by the BLRPGM fit the observed moments better than those from the MBLRPM.

In order to compare the performance of the models at different stations, the mean value (*M_STATS_INDEX*) of the seven $STATS\_INDEX_{(j)}$ values were computed for each station and parameter set as shown in Equation 89.

$$M\_STATS\_INDEX = \frac{1}{7}\sum_{j=1}^{7} STATS\_INDEX_{(j)} \qquad ...89$$

where

M_STATS_INDEX = goodness-of-fit index of model to all seven moments for durations ranging from 2 h to 24 h and for all rainy season months

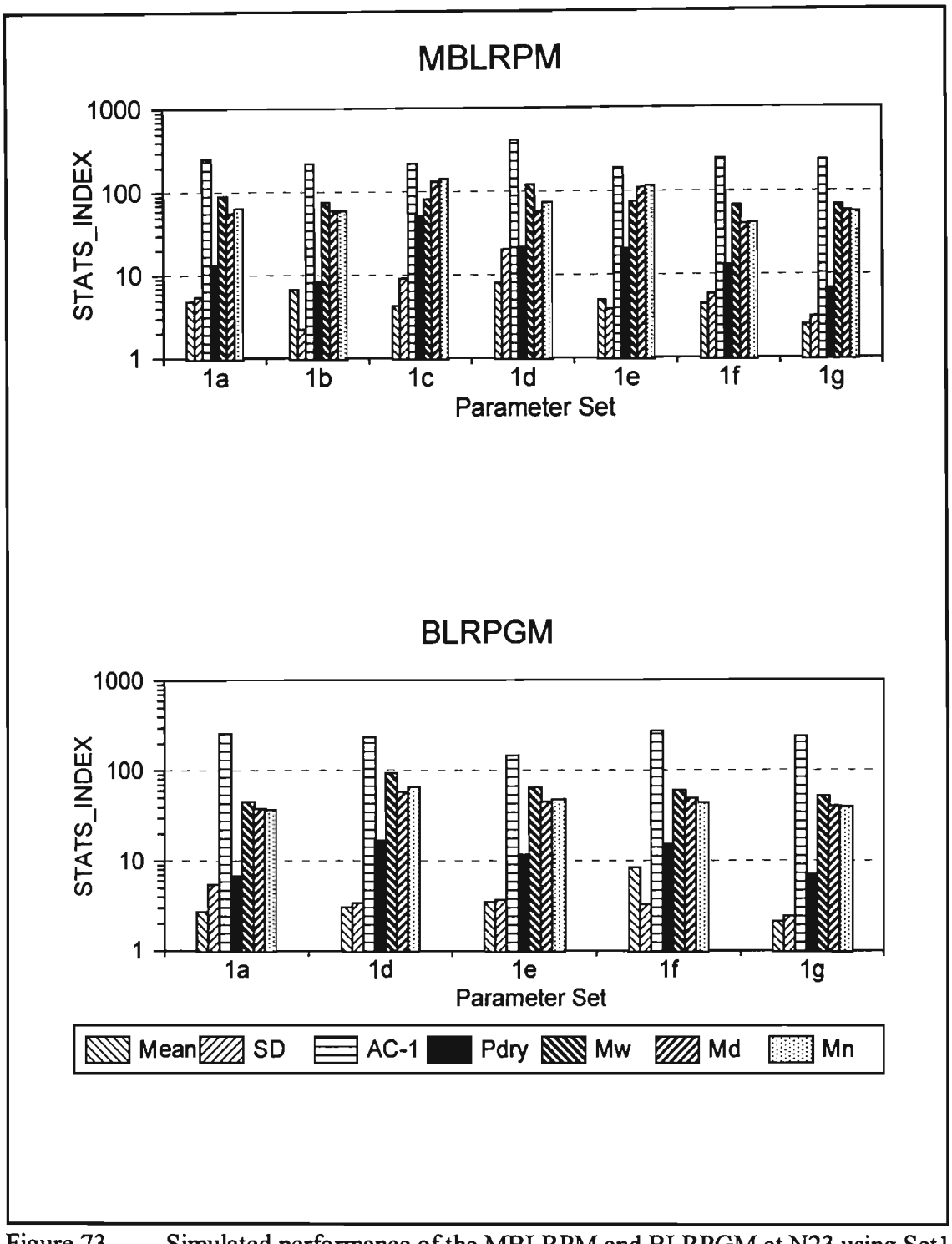


Figure 73 Simulated performance of the MBLRPM and BLRPGM at N23 using Set1 parameters for rainy season months and durations ranging from 2 h to 24 h

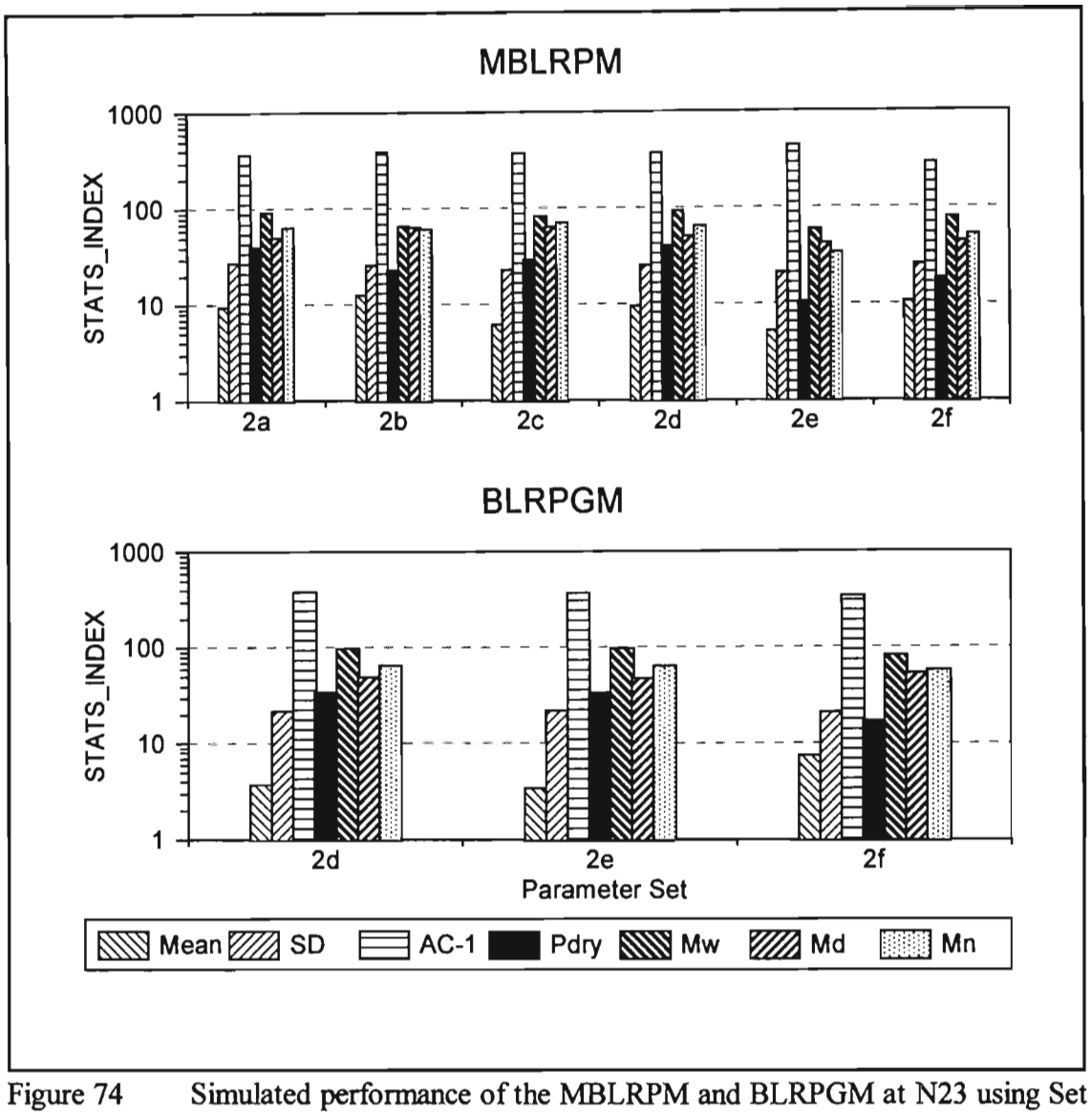


Figure 74 Simulated performance of the MBLRPM and BLRPGM at N23 using Set 2 parameters for rainy season months and durations ranging from 2 h to 24 h

The *M_STATS_INDEX* was computed at selected stations for both the MBLRPM and BLRPGM using all parameters sets. The results for the Set 1 parameters are shown in Figure 75 and for the Set 2 parameters in Figure 76. Assuming that short duration rainfall data were available at all the sites, then the best performance for the MBLRPM, relative to the seven statistics considered, was achieved with parameter Set 1f while for the BLRPGM the performance for all Set 1 parameters were similar. However, if the performance of MBLRPM and BLRPGM are considered for the same Set 1 parameters it is evident that the synthetic rainfall series generated by the BLRPGM fit the observed data better than the series generated by the MBLRPM. Assuming that only daily rainfall data are available at the

selected stations then, as shown by the results for Set 2 parameters in Figure 76, the performance of the two models is very similar for both the parameter sets and, with the exception of Station Moko3a, the best performance for both models is obtained using parameter Set 2f. These trends in the simulated performance of the models for the different parameters sets reflect the trends noted in the analytical performance of the models. With the focus of the study being the estimation of design rainfall values, the most important assessment of the models is how well the extreme events are modelled in the synthetic rainfall series.

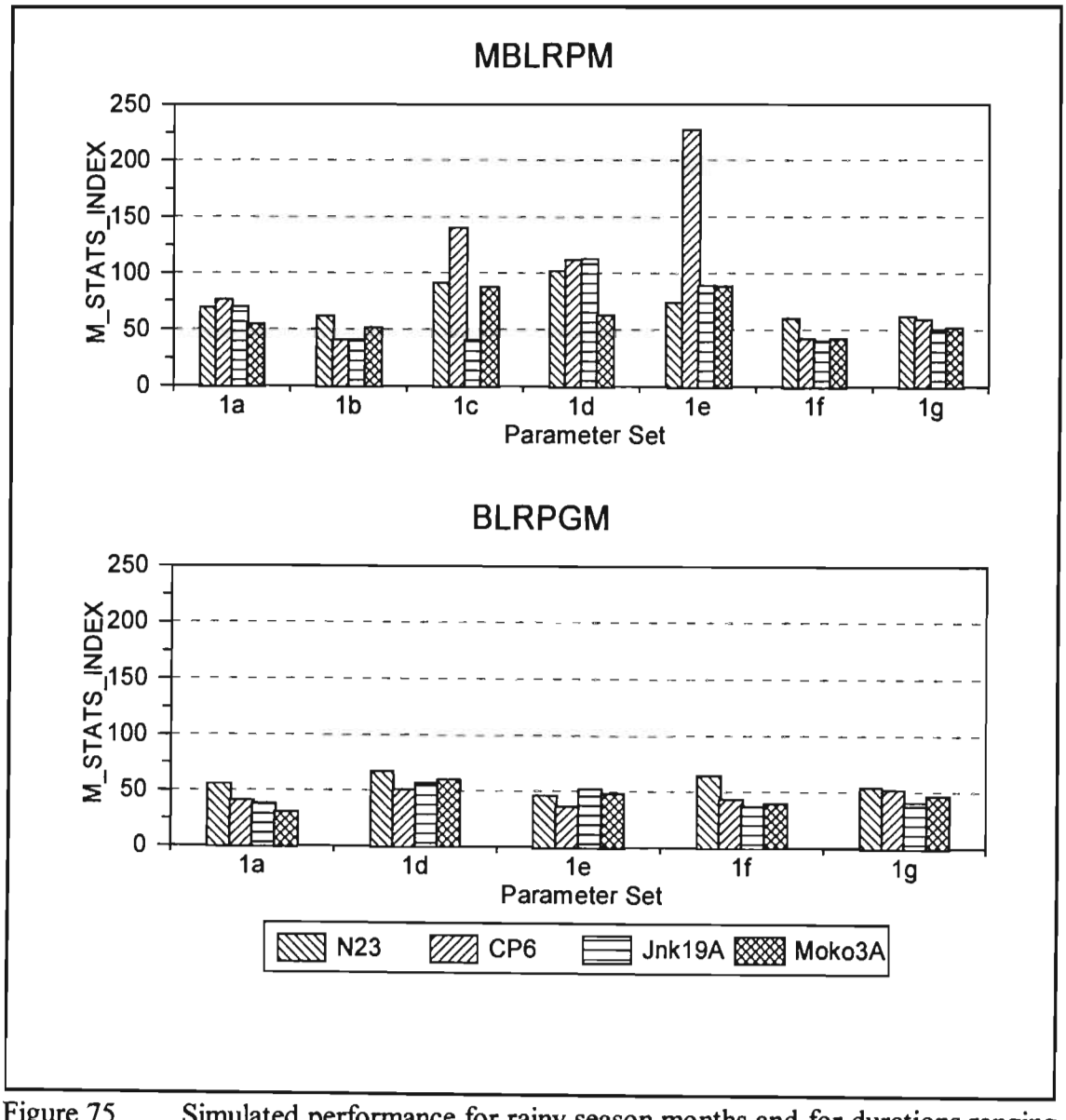


Figure 75 Simulated performance for rainy season months and for durations ranging from 2 h to 24 h of the MBLRPM and BLRPGM at selected stations using Set 1 parameters

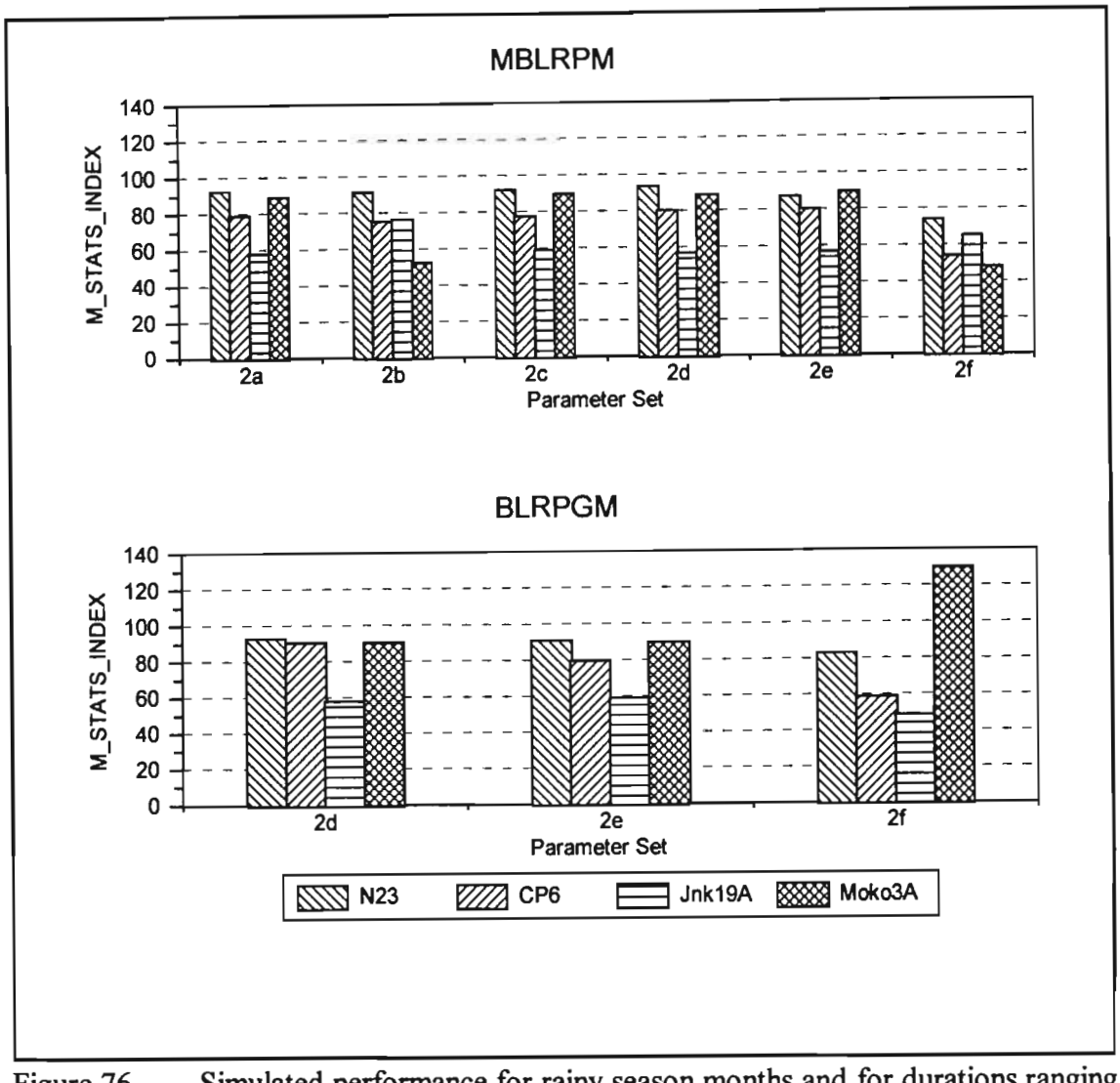


Figure 76 Simulated performance for rainy season months and for durations ranging from 2 h to 24 h of the MBLRPM and BLRPGM at selected stations using Set 2 parameters

7.8.2 Extreme Rainfall Events

For the observed data and for each of the 100 synthetic series generated by the model, design rainfall depths were calculated using the General Extreme Value (GEV) distribution fitted to the Annual Maximum Series (AMS) by L-moments. Design values for 2, 5, 10, 20, 50 and 100-year return periods were computed for rainfall durations of 0.25. 0.5, 1, 2, 3, 4, 5, 6, 9, 12, 15, 18 and 24 h. For each duration and return period, a frequency analysis was performed on the 100 values computed from the synthetic rainfall series generated by

the model. High -Low bar graphs depicting the observed design rainfall computed from the observed data and the 25-th and 75-th non-exceedance percentiles of the 100 synthetic data sets were used to evaluate the adequacy of the models. For example, the performance of the MBLRPM (Set 1b) for the best (January) and worst (December) rainy season month and annual totals is shown in Figure 77.

The estimation of design rainfall values at N23 from the synthetic rainfall series generated by the MBLRPM using Set 1b parameters compares well with the design values computed from observed data for January and annual totals shown, particularly for durations > 3 h and return periods < 50 years. The fit is not as good for December where the performance for durations ≤ 1 h and return periods ≤ 20 years is better than for durations > 1 h and return periods > 20 years. In order to objectively assess the performance of the two models and the various parameter sets, relative to the estimation of design rainfalls, the Mean Absolute Relative Error (*MARE*) was calculated to include rainy season months and annual totals and return periods ranging from 2 to 50 years, as shown in Equation 90.

$$MARE = \frac{100}{N_M \times N_L \times N_{RP}} \times \sum_{i=1}^{N_M} \sum_{j=1}^{N_L} \sum_{k=1}^{N_{RP}} \left(\frac{\left| S_{(i,j,k)} - O_{(i,j,k)} \right|}{O_{(i,j,k)}} \right) \qquad \ldots 90$$

where

$MARE$	=	mean absolute relative error of design rainfall (%),
$S_{(i,j,k)}$	=	mean k-th return period, j-th hour design rainfall computed for i-th period from model generated rainfall series,
$O_{(i,j)}$	=	k-th return period, j-th hour design rainfall computed for i-th period from observed data,
N_M	=	number of periods (7), 1 to 6 = rainy season months and 7 = annual period
N_L	=	number of aggregation levels (=10)
N_{RP}	=	number of return periods (=5 for 2, 5, 10, 20 and 50 year return periods)

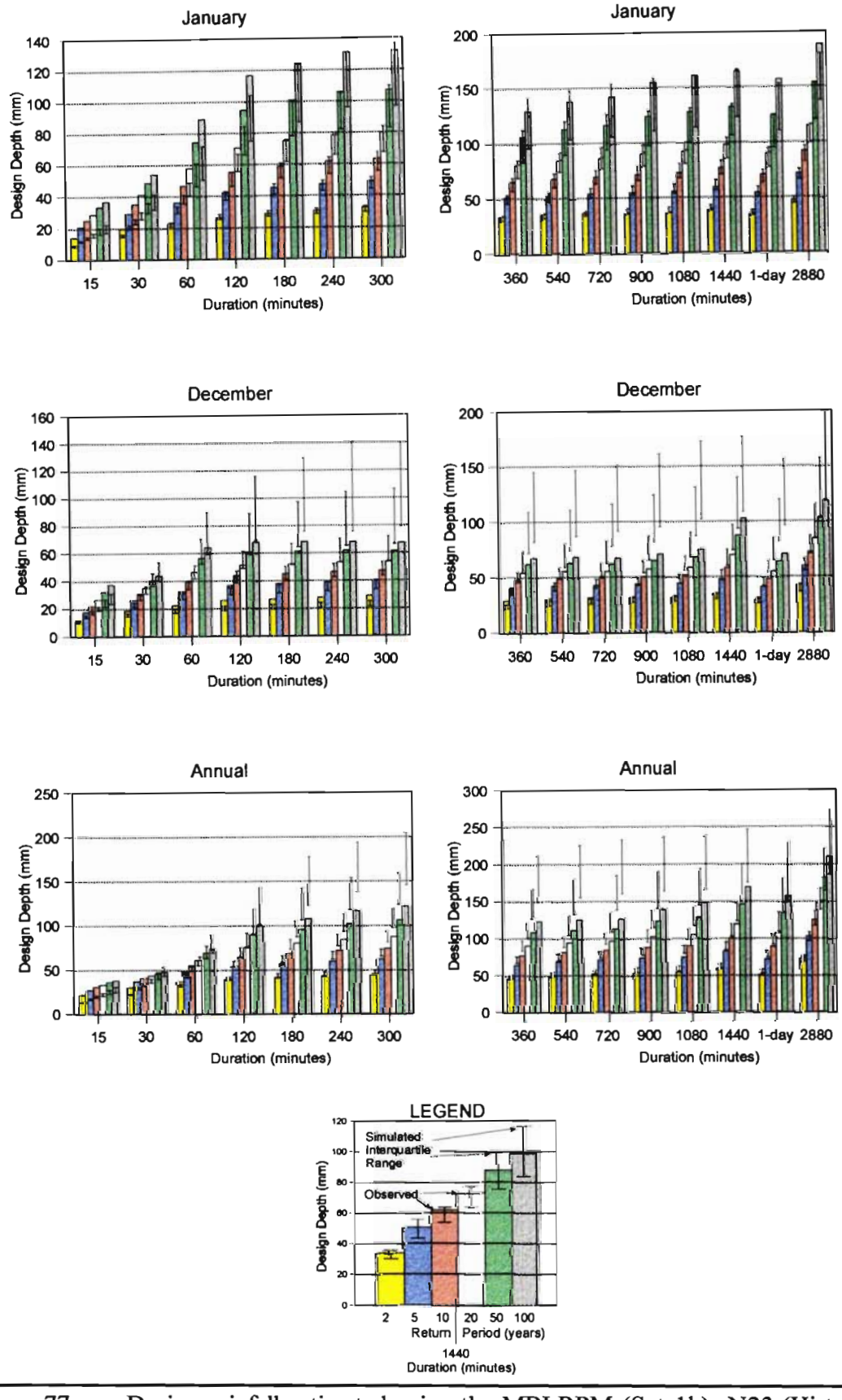


Figure 77 Design rainfall estimated using the MBLRPM (Set 1b): N23 (Historical values in histograms. Interquartile range of 100 simulations in I-beams)

The design rainfall *MARE* values, for rainy season months and annual periods and for 10 aggregation levels (2, 3, 4, 5, 6, 9, 12, 15, 18 and 24 h), computed at selected stations from rainfall series generated by both the MBLRPM and BLRPGM, are shown in Figure 78. For both models the *MARE* values for parameter Set 1 are better than those for Set 2, indicating that when short duration rainfall data are available at a site, better design rainfall values are computed using the models than when only daily rainfall data are available. Parameter Sets 1f and 1g resulted in the best performance of the MBLRPM for the Set 1 parameters, while similar performance at all stations was obtained for all Set 1 parameters for the BLRPGM. Parameter Set 2f, which uses variances estimated from the daily rainfall data for durations shorter than 24 h, resulted in the lowest *MARE* values for Set 2 parameters for both models and is thus recommended for use when only daily rainfall data are available for parameter determination. The *MARE* values from the BLRPGM are generally lower than those from the MBLRPM and hence the BLRPGM is recommended as the preferred model to use. Although the Set 2 parameters resulted in *MARE* values larger than those from the Set 1 parameters, the *MARE* values for Set 2 were generally less than 20 % for Set 2f at most stations. Thus the use of only daily rainfall data to determine the parameters for the models is considered to be feasible.

The above analysis has only considered *MARE* values durations > 1 h. The *MARE* values, computed using the BLRPGM, for durations ≤ 1 h (15, 30 and 60 min) as well as *MARE* values for durations > 1h are shown in Figure 79 for the test stations. Generally the Set 1 parameters result in better estimates of design rainfall values for longer duration values than for shorter (< 2h) durations. Clearly the use of the BLRPGM to estimate design rainfalls for short durations (< 2h), particularly when only daily data are used to determine the model parameters (Set 2), results in unacceptably large *MARE* values. The contrast in the *MARE* values when the digitised data are available (Set 1) and when only the daily data are available (Set 2) for parameter determination, particularly for the durations < 2h, is attributed to the poor estimates of the variances for shorter durations.

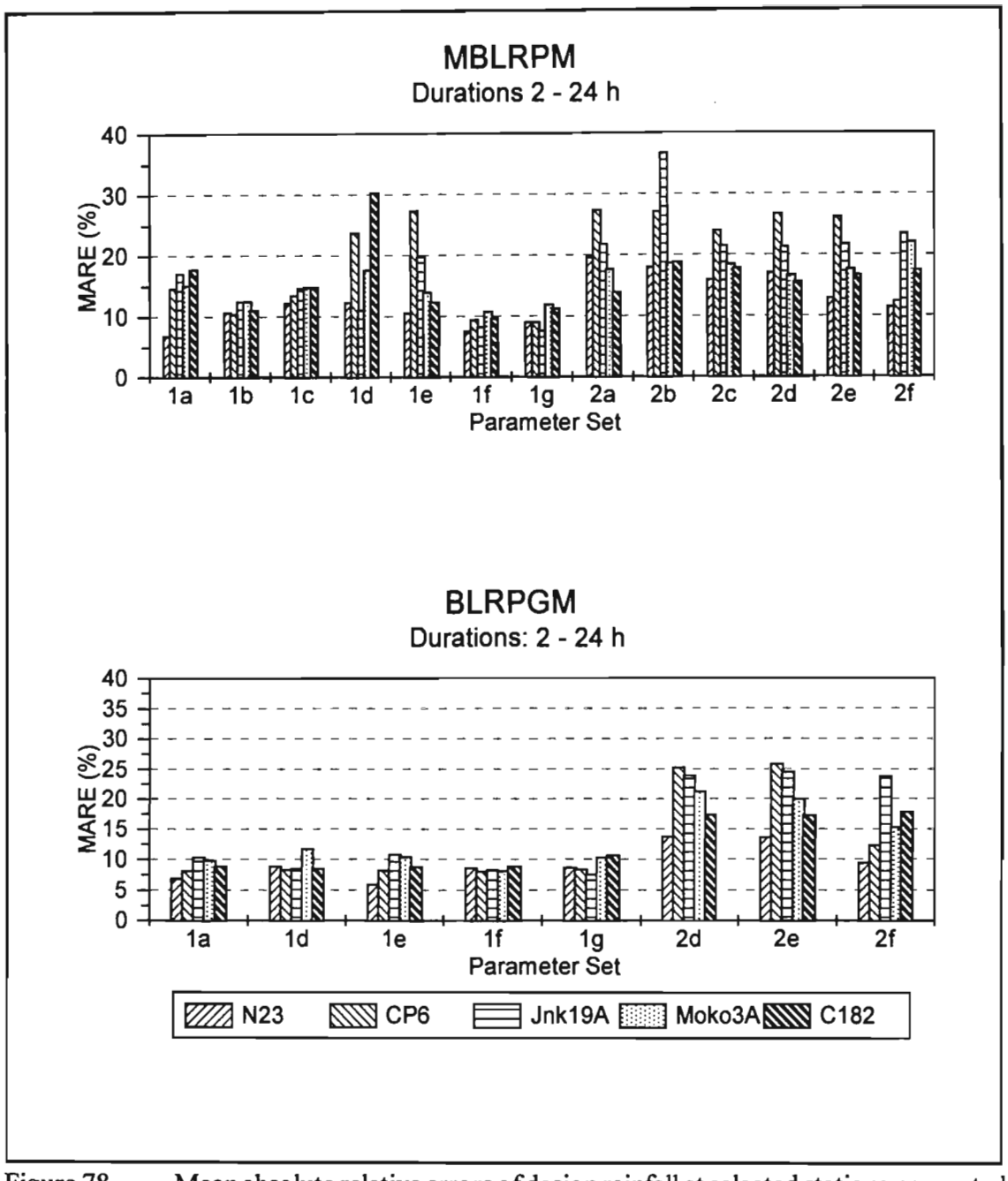


Figure 78 Mean absolute relative errors of design rainfall at selected stations computed from the synthetic rainfall series generated by the MBLRPM and BLRPGM, using various parameter sets

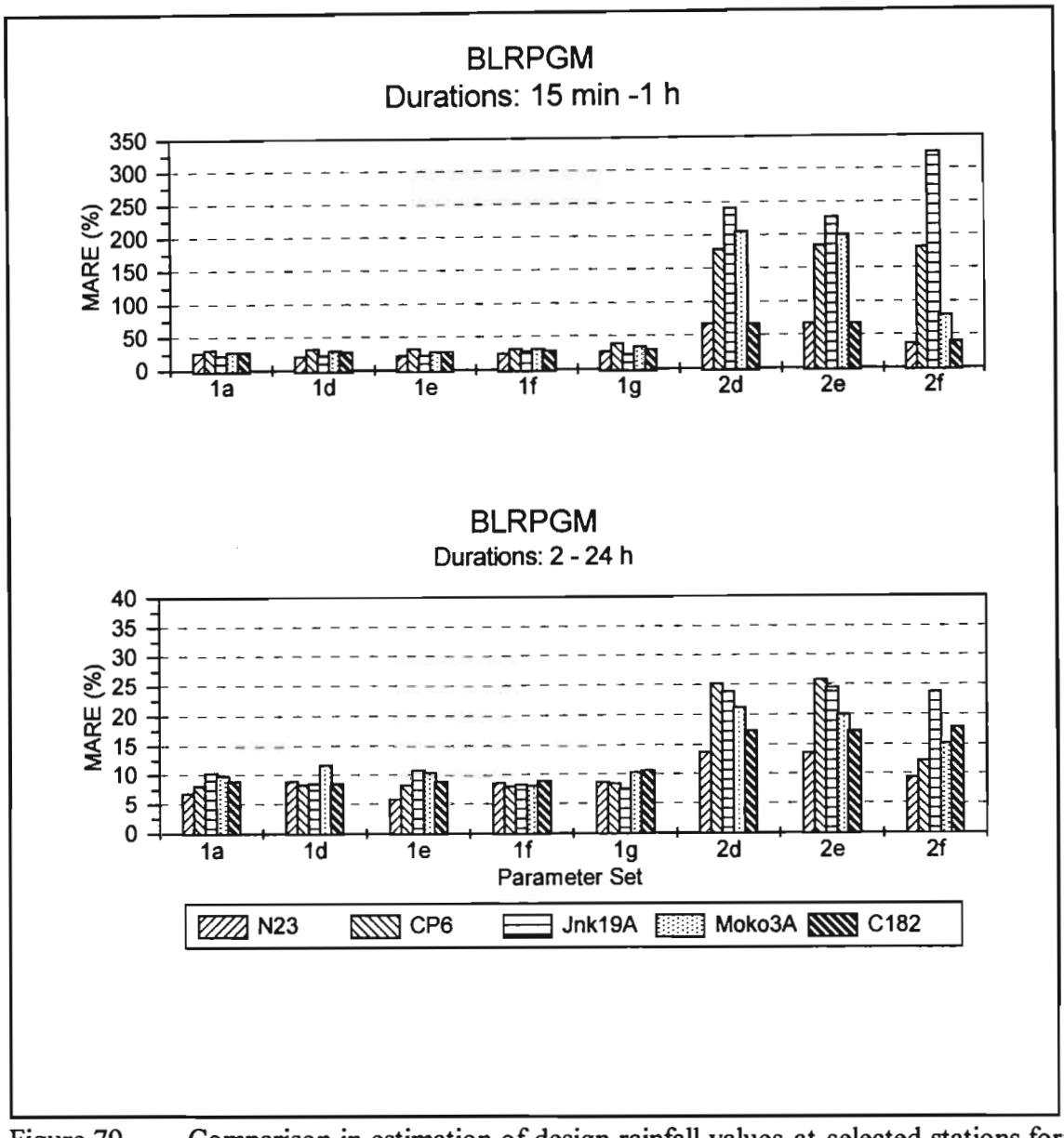


Figure 79 Comparison in estimation of design rainfall values at selected stations for shorter and longer durations using the BLRPGM

The trends in the estimation of design rainfall using the two models and various parameter sets are consistent with those found when evaluating the analytical performance of the models. The incremental search technique, developed to determine model parameters, improved the fit of the models to the observed moments and for all three measures of performance it was noted that:

- the BLRPGM generally performed better than the MBLRPM,
- the performance of the BLRPGM was generally less sensitive to the set of moments used to determine the model parameters than the MBLRPM,
- the performance of both models was best when over determined systems (more equations than parameters, e.g. Sets 1f, 1g and 2f) were used to determine model parameters,
- the use of variances for durations < 24 h estimated from the daily values successfully improved the model performance when only daily data are available to estimate model parameters, and
- the use of the BLRPGM with parameters determined using either Sets 1f or 2f moments, dependent on the availability of short duration rainfall data, is deemed to be a suitable technique to estimate design rainfall values in South Africa.

The above selection of the most appropriate model and parameter set and results are based on a selected number of non-SAWB stations where the data are considered to be reliable. The use of the BLRPGM to estimate design storms for these test stations and other stations, using parameter Sets 1f and 2f, is shown in Figure 80. The results contained in Figure 80 exclude outlier events in the observed data. For example, design storms estimated from the observed data at Cedara (SAWB 0239482) excluded outlier events from 26-29 September 1987. Similarly, outlier events which occurred on 20 January 1972 and 22 December 1978 at Johannesburg International Airport (SAWB 0476398) were excluded in the estimation of design storms from the observed data. Despite the exclusion of outlier events the performance of the models at some sites, even when digitised rainfall data are available (Set 1f), is not considered to be adequate. These anomalies are investigated in the following section.

7.8.3 Anomalies in the Estimation of Design Rainfalls

The relatively large differences in *MARE* values obtained using Set 2f parameters compared to values obtained using Set 1e parameters, as shown in Figure 80 at stations Jnk19A,

Moko3A and Newlands, is postulated to be the result of the poor estimation, from daily data, of the variance of short durations when parameters were determined using moment Set 2f.

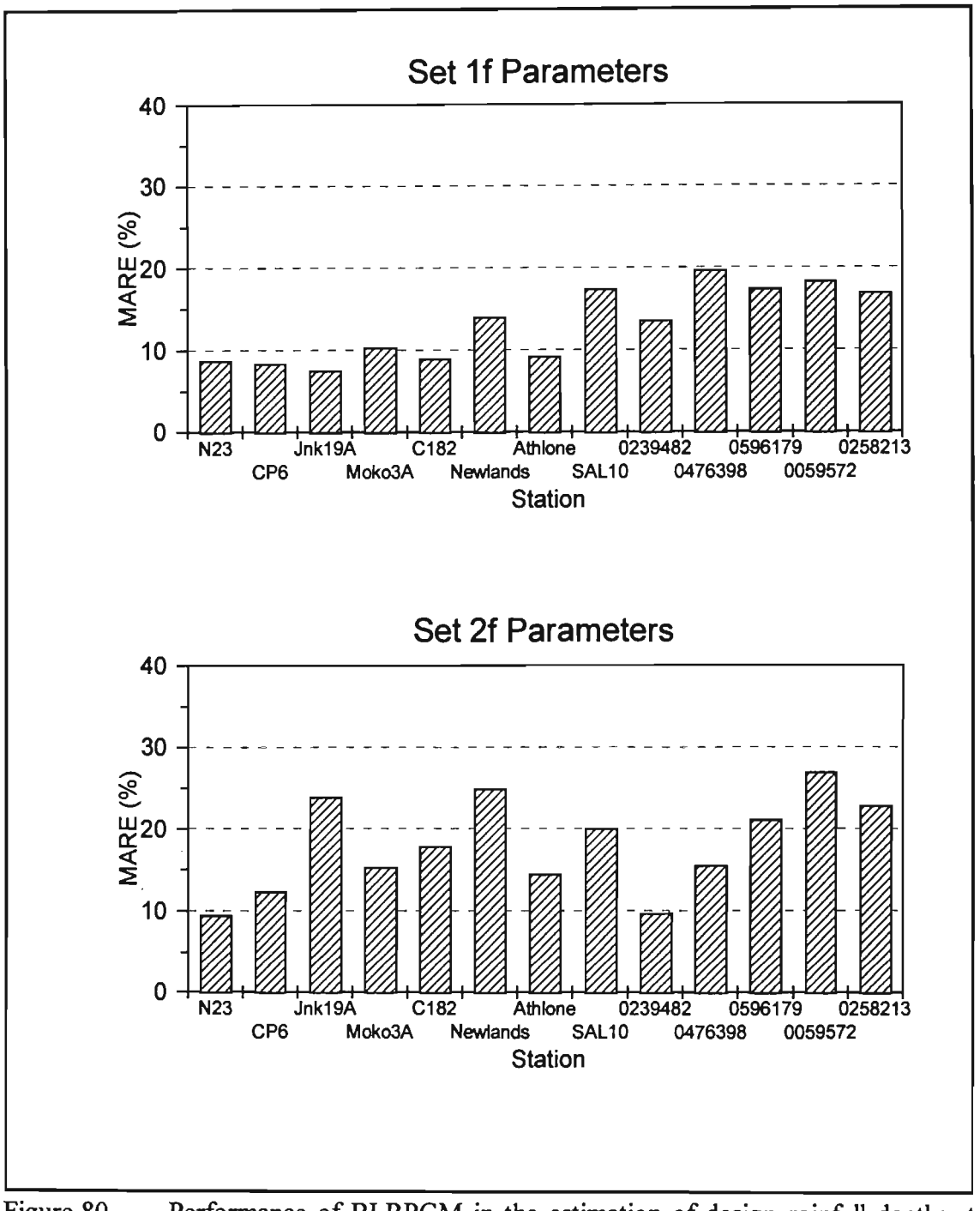


Figure 80 Performance of BLRPGM in the estimation of design rainfall depths at selected stations using parameter Sets 1f and 2f

As shown in Figure 80 for Set 1f parameters, design rainfall values estimated from the synthetic rainfall series generated by the BLRPGM are, without exception, better at non-SAWB stations than at SAWB stations. The reasons for this are attributed to the general unreliability and periods of missing data in the SAWB digitised database. These inconsistencies in the SAWB digitised database are illustrated in Figure 81 using data from SAWB 0258213 (Drieplotte). The results from the month which resulted in the smallest design rainfall *MARE* value (March) and the largest *MARE* value (November) and a month to illustrate the effect of periods of missing data (January) on design values are shown in Figure 81.

No high outliers were detected in the AMS extracted from either the digitised or daily rainfall data. However, an inconsistency between the 1 day and 24 h design storms is evident with the 1 day values exceeding the 24 h values for all months shown in Figure 81, thus indicating periods of missing digitised data during significant events. The effect of missing periods of digitised data on design values is also evident for January where the 100 year return period, 1440 min event is smaller than for shorter durations. Thus some larger events, which are extracted in the AMS for shorter durations, are not extracted for longer durations events, as periods of missing data appear within the longer duration and hence the entire event is excluded.

The problem of missing periods of data, particularly in the digitised data set, not only affects the design values computed from the data, but also affects the reliability of model parameters determined using the data. Twenty eight years of digitised rainfall records are available at Station 0258213. In the calculation of the moments from the observed data which are used to derive the parameters for the model, months are excluded if any missing data are encountered within the month. As shown in Table 56, more than 60 % of months are not used in parameter determination as a result of periods of missing data within the months and this consequently affects the reliability of the estimated model parameters.

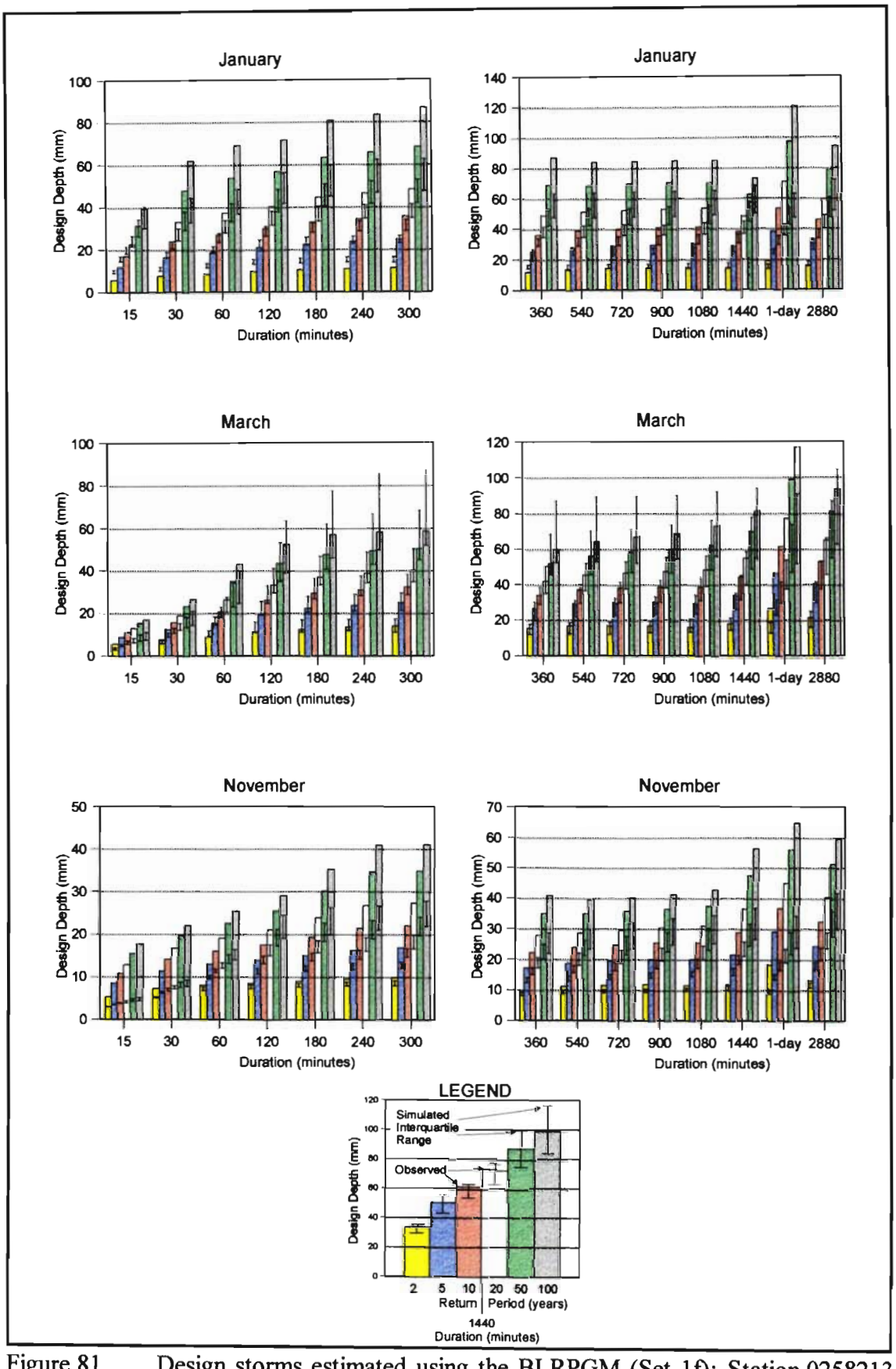


Figure 81 Design storms estimated using the BLRPGM (Set 1f): Station 0258213 (Drieplotte) (Historical values in histograms. Interquartile range of 100 simulations in I-beams)

Table 56 Percentage of months with no missing data: Drieplotte (SAWB 0258213)

Month	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Percentage	39	29	21	36	39	21	25	11	21	29	29	32

For the Set 2f parameters shown in Figure 80 the station with the largest *MARE* value was SAWB 0059572 (East London). The month at SAWB station 0059572 with the largest *MARE* value was November and although a number of large historical events occurred in November, these are statistically not outliers and hence are retained in the observed data. Two AMS, plotted using the Weibull plotting position, are shown for January and November in Figure 82. It is noticeable that the events in November appear to arise from two distinct meteorological conditions, as indicated by the sharp change in gradient at a return period of approximately 6 to 10 years. The design storms estimated from the observed data using the GEV distribution and those derived from the Weibull plotting formula agree reasonably well despite the possibility of the events arising from the different conditions. Hence it appears that the BLRPGM is unable to simulate extreme events arising from differing meteorological conditions. It is postulated that these relatively few larger events probably have little affect on the moments computed from the data which are used in the estimation of model parameters, but do have a large effect on the estimation of design storms from the synthetic rainfall series. These differing meteorological conditions resulting in an AMS with two distinct populations is typical of the East Coast of South Africa, where the use of the Two Component Extreme Value Distribution (TCEV) was used by Pegram and Adamson (1988).

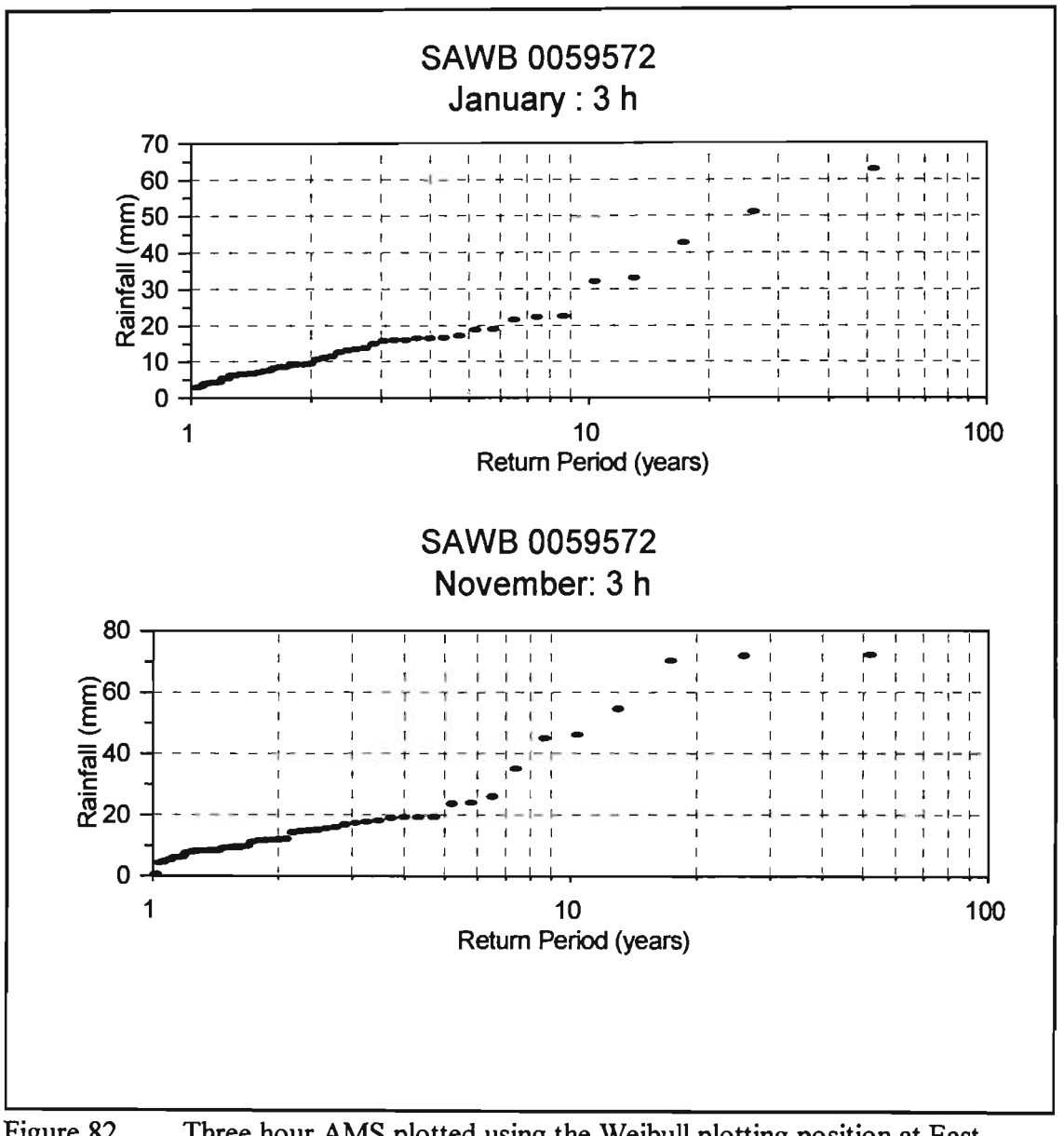


Figure 82 Three hour AMS plotted using the Weibull plotting position at East London

7.8.4 Concluding Remarks on Simulated Performance

The simulated performances of the MBLRPM and BLRPGM have been evaluated, at a number of sites in different climatic regions in South Africa, for different sets of moments used to determine model parameters. The estimation of model parameters proved to be an exacting task, particularly as similar performances were obtained from sets of parameters

which are very different. The use of constrained minimisation procedures, thereby ensuring reasonable mean analytical storm characteristics, aided in the estimation of parameters. In addition, the estimation of the reliability (CV) of the parameters and the correlation between the model parameters assisted in developing a strategy of fixing one or more of the parameters. Despite these measures, difficulties were still encountered in estimating "reasonable" parameters for some months at some locations. This can be only explained by either the total unsuitability of the BLRPGM to be applied at the location or the result of inconsistencies and errors in the data, some of which have been illustrated.

The three measures of performance used to evaluate the fit between observed and model values were analytical moments, simulated moments and the estimation of design values from the simulated rainfall series. It was noted that the performance of the BLRPGM, despite having one more parameter to estimate compared to the MBLRPM, was generally less sensitive than the MBLRPM to the set of moments used to estimate the parameters of the model. In addition it was found that the use of the BLRPGM generally resulted in better estimates of design rainfall values than those computed using the MBLRPM. Parameter Sets 1e and 1f resulted in the best performance of the models, assuming that the short duration digitised data were available, and parameter Set 2f gave the best performance when only daily rainfall data were available to estimate model parameters. Hence the use of the variances estimated from the daily data for durations < 24 h successfully assisted in the estimation of model parameters and improved the performance of the model.

Design storms were generally well estimated from the synthetic rainfall series generated by the BLRPGM for durations > 1 h when short duration data were available and for durations > 3 h when only daily recorded interval data were available. Thus, the BLRPGM with model parameters determined using moment Sets 1f or 2f, dependent on the availability of digitised data, is recommended as a feasible option for estimating short duration design rainfall values in South Africa. Thus, in cases where errors were apparent in the digitised data, it is postulated that the use of the BLRPGM would result in more reliable estimates of design storms than if the design storms were estimated directly from the observed data.

The model performed better for the durations of the moments which were used in the estimation of the model parameters, than for other durations. However, the BLRPGM did scale reasonably well particularly in an aggregation sense where, for example, the model performs better for longer durations when only shorter duration moments are used in the estimation of model parameters than for shorter durations when the parameters are estimated from longer durations (disaggregation).

Although limited by the amount of the data which was considered to be acceptably reliable, the use of the BLRPGM to estimate design storms was relatively successful in different climatic regions of South Africa. However, it appears that the model does not perform well at locations where there is a distinct difference between two sets of data in the AMS, probably as a result of different meteorological conditions. In the following section, the temporal distribution of synthetic hyetographs generated by the BLRPGM are investigated.

7.9 TEMPORAL DISTRIBUTION OF STORMS

Mass curves depicting, from the onset of a storm, the dimensionless cumulative storm duration vs the cumulative storm depth are important in certain hydrological design problems where it is necessary to estimate a design hyetograph. Thus it is important to assess how the stochastically generated storms compared to the historical storms.

The analysis performed was similar to that presented by Huff (1967) and Verhoest *et al.* (1997). Various periods of no rainfall, or Inter Event Times (IET), for identifying independent storms have been used in previous studies. For example, IETs that have been used are 1 h (Van den Berg, 1982), 3 h (Calles and Kulander, 1995), 6 h (Huff, 1967) and 24 h (Verhoest *et al.*, 1997). In this study a period of 12 h of no rainfall was used to identify independent storms.

The independent storms identified were classified into four groups or quartiles, depending on whether the heaviest rainfall fell in the first, second, third or fourth quarter of the

duration of the storm. A frequency analysis was then performed on the storms in all four quartiles. This analysis was performed both on the historical data and on periods of synthetic rainfall series, generated by the BLRPGM, which was equal in length to the historical data. In addition, the frequencies of occurrence of storms in the four quartiles computed from the historical data and synthetic series were compared.

The above analyses were performed at selected stations in South Africa. The results of the analyses are presented in the following sections.

7.9.1 Ntabamhlope (N23)

As shown in Figure 83 for storms identified having a 12 h IET, the temporal distribution of historical storms and synthetic storms generated by the BLRPGM using parameters Set 1e at Ntabamhlope (N23) are very similar, However, as shown in Figure 84, the frequency of Quartile 1 storms in the synthetic series is less than in the historical series and the frequency of Quartile 4 storms in the synthetic series is greater than in the historical data. Similar results were obtained for storms at N23 identified by 1, 6 and 24 h IETs. The frequency distribution of storm depths and durations computed from the historical data and synthetic series generated by the BLRPGM (Set 1e) are shown in Figure 85. The distribution of storm depths in the synthetic series is very similar to the historical distribution. However, the synthetic series contain fewer longer duration storms.

As shown in Figure 86, when the BLRPGM was used with parameter Set 2f, the temporal distribution of storms corresponded closely to those computed from the historical data and were similar to results obtained when parameter Set 1e was used. However, as shown in Figure 87, the duration of storms in the simulated series corresponded better to the durations of the observed storms when parameter Set 2f, which utilised longer duration moments in the estimation of parameters, than when parameter Set 1e was used.

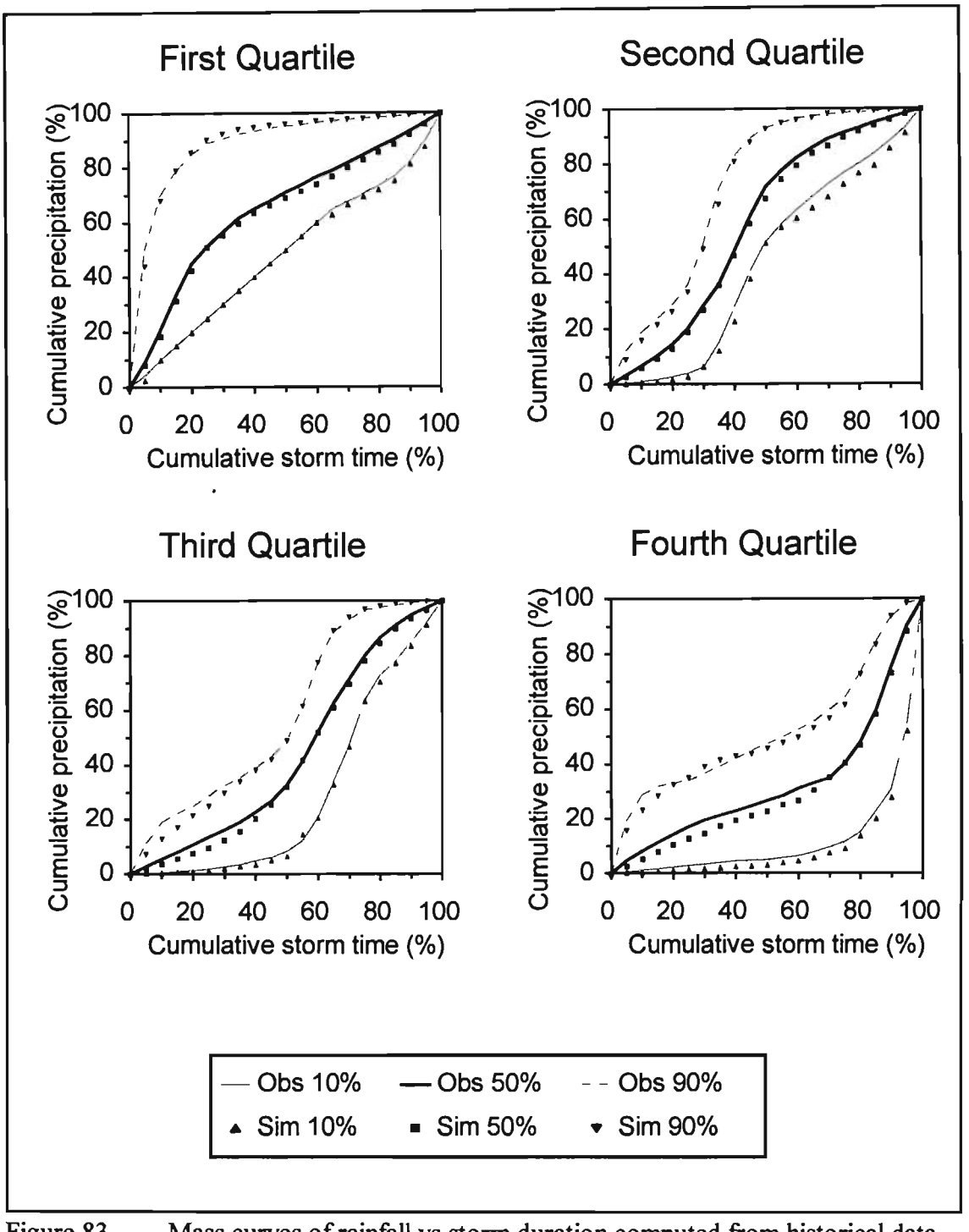


Figure 83 Mass curves of rainfall vs storm duration computed from historical data and from synthetic rainfall series generated by BLRPGM (parameter Set 1e) at N23

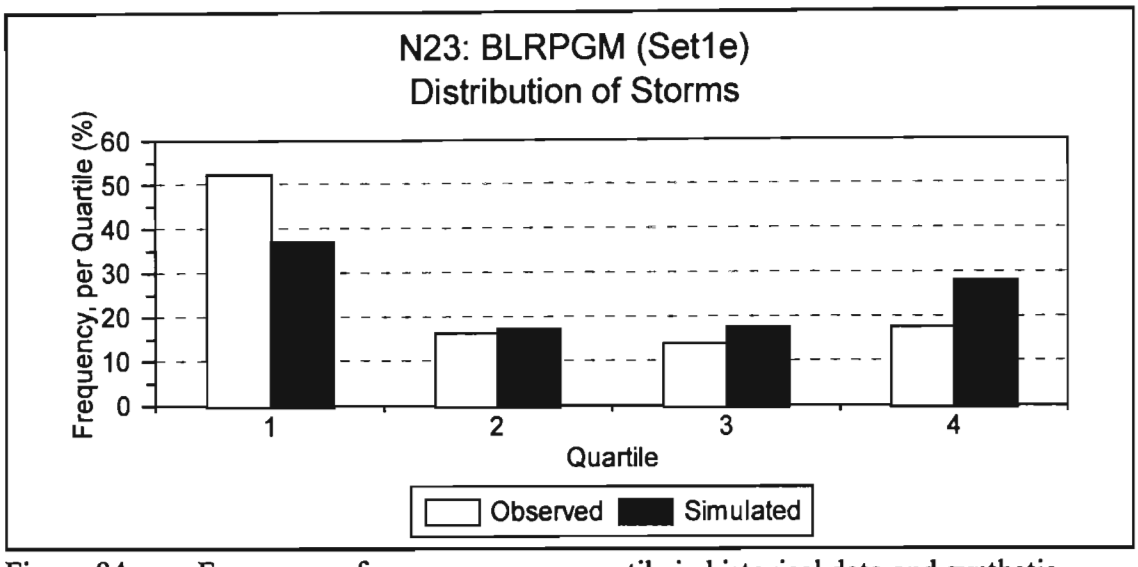


Figure 84 Frequency of occurrence per quartile in historical data and synthetic storm series generated by BLRPGM (Set 1e) at N23

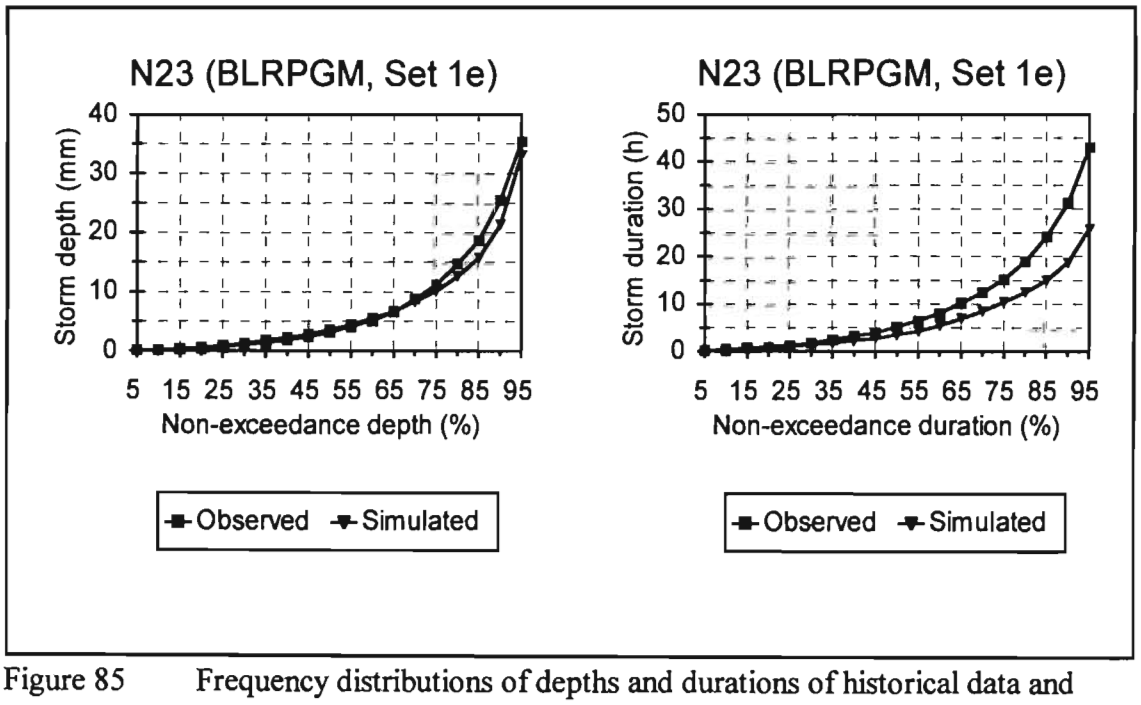


Figure 85 Frequency distributions of depths and durations of historical data and synthetic series generated by BLRPGM (Set 1e) at N23

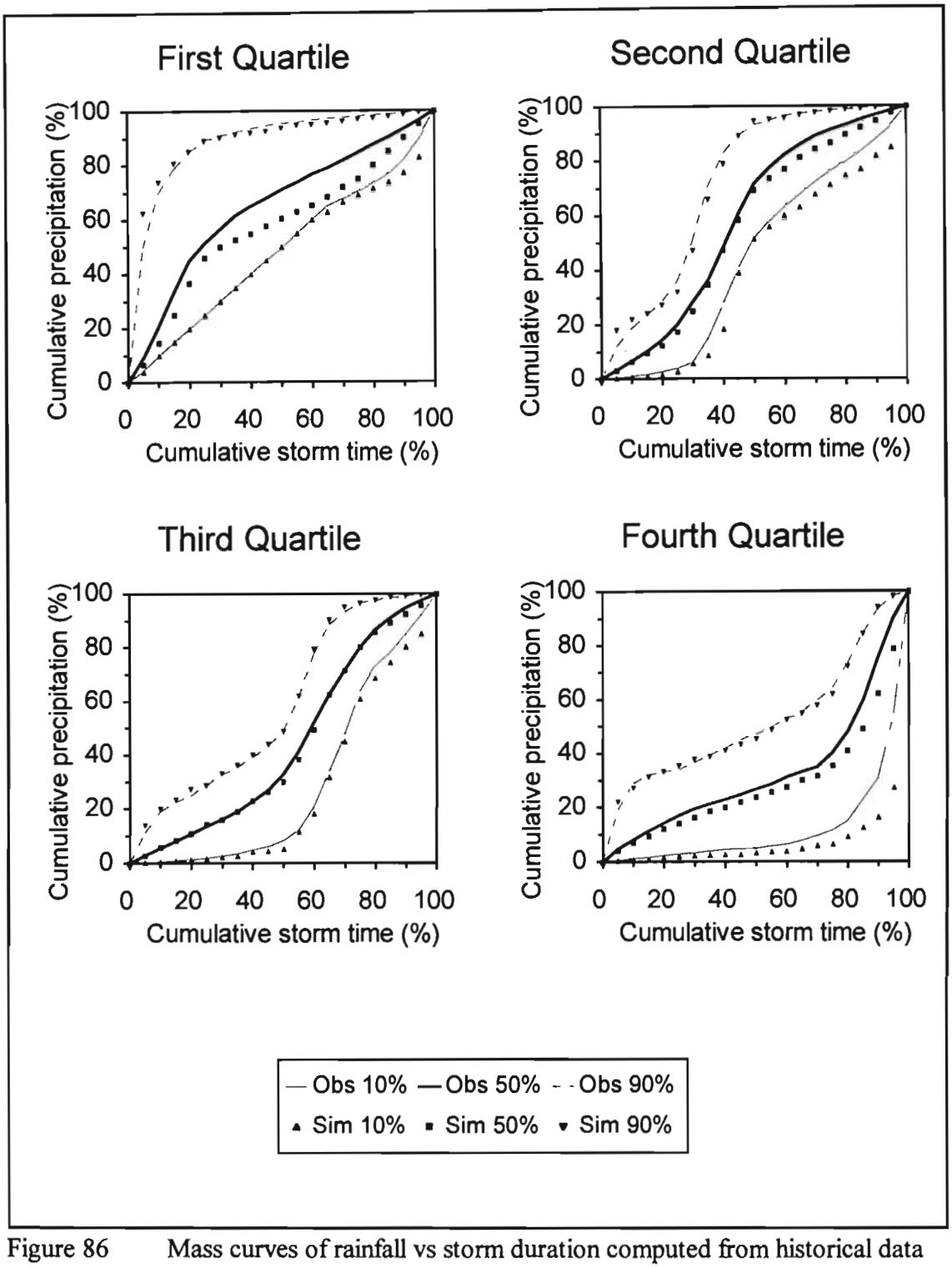


Figure 86 Mass curves of rainfall vs storm duration computed from historical data and from synthetic rainfall series generated by BLRPGM (parameter Set 2f) at N23

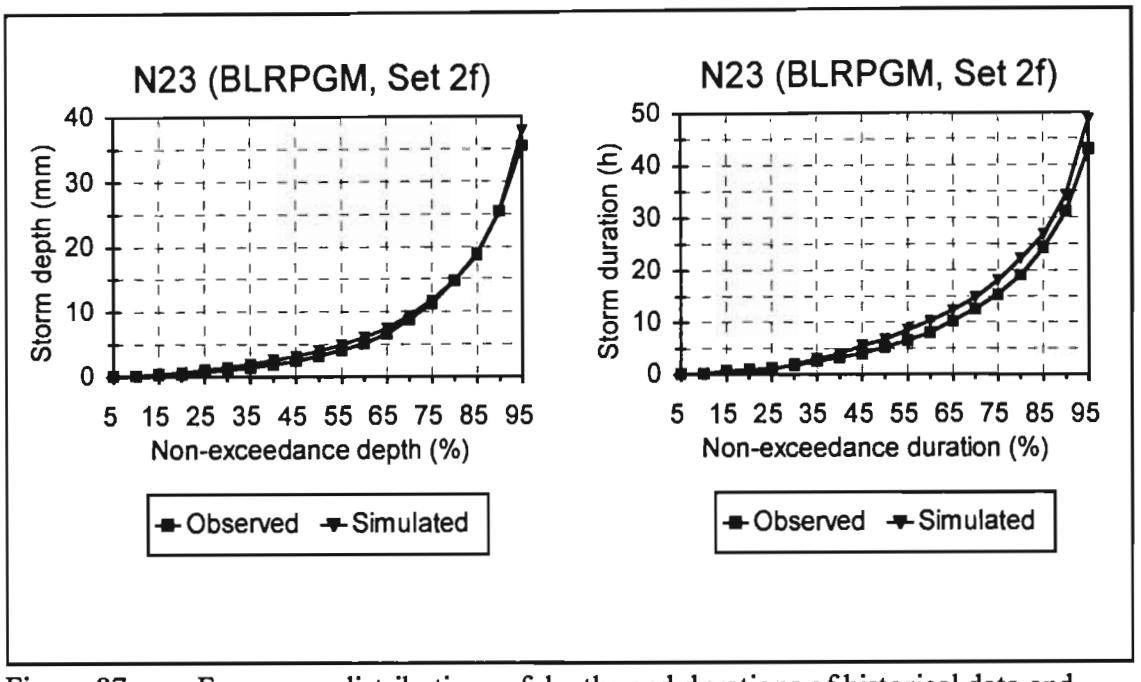


Figure 87 Frequency distributions of depths and durations of historical data and synthetic series generated by BLRPGM (parameter Set 2f) at N23

7.9.2 Jonkershoek (Jnk 19A)

As shown in Figure 88, the BLRPGM, with parameters derived using Set 2f, underestimated the frequency of Quartile 2 and 3 storms and overestimated the frequency of occurrence of Quartile 4 storms at Jnk19A. The distribution of storm depths was well simulated by the model, as shown in Figure 89. However, the longer duration storms in the synthetic series were generally shorter than the historical durations. The mass curves computed from the historical data and synthetic rainfall series at Jnk19A, shown in Figure 90, indicate that the synthetic rainfall storms generated by the BLRPGM have a similar distribution to the historical storms.

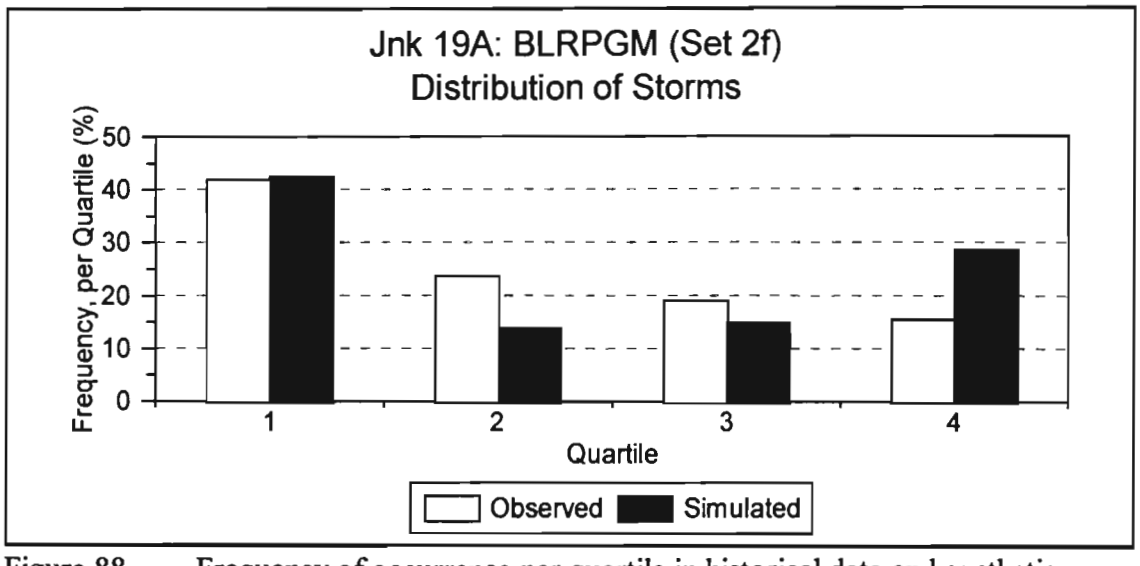


Figure 88 Frequency of occurrence per quartile in historical data and synthetic storms series generated by BLRPGM (parameter Set 2f) at Jnk19A

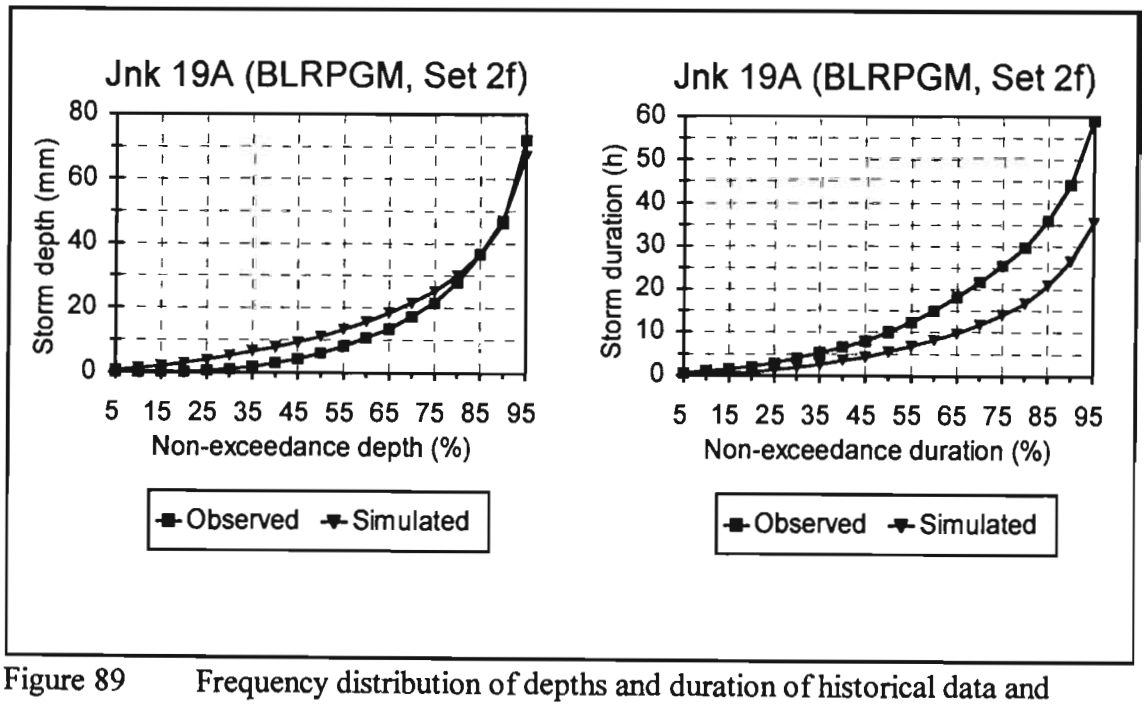


Figure 89 Frequency distribution of depths and duration of historical data and synthetic series generated by BLRPGM (parameter Set 2f) at Jnk19A

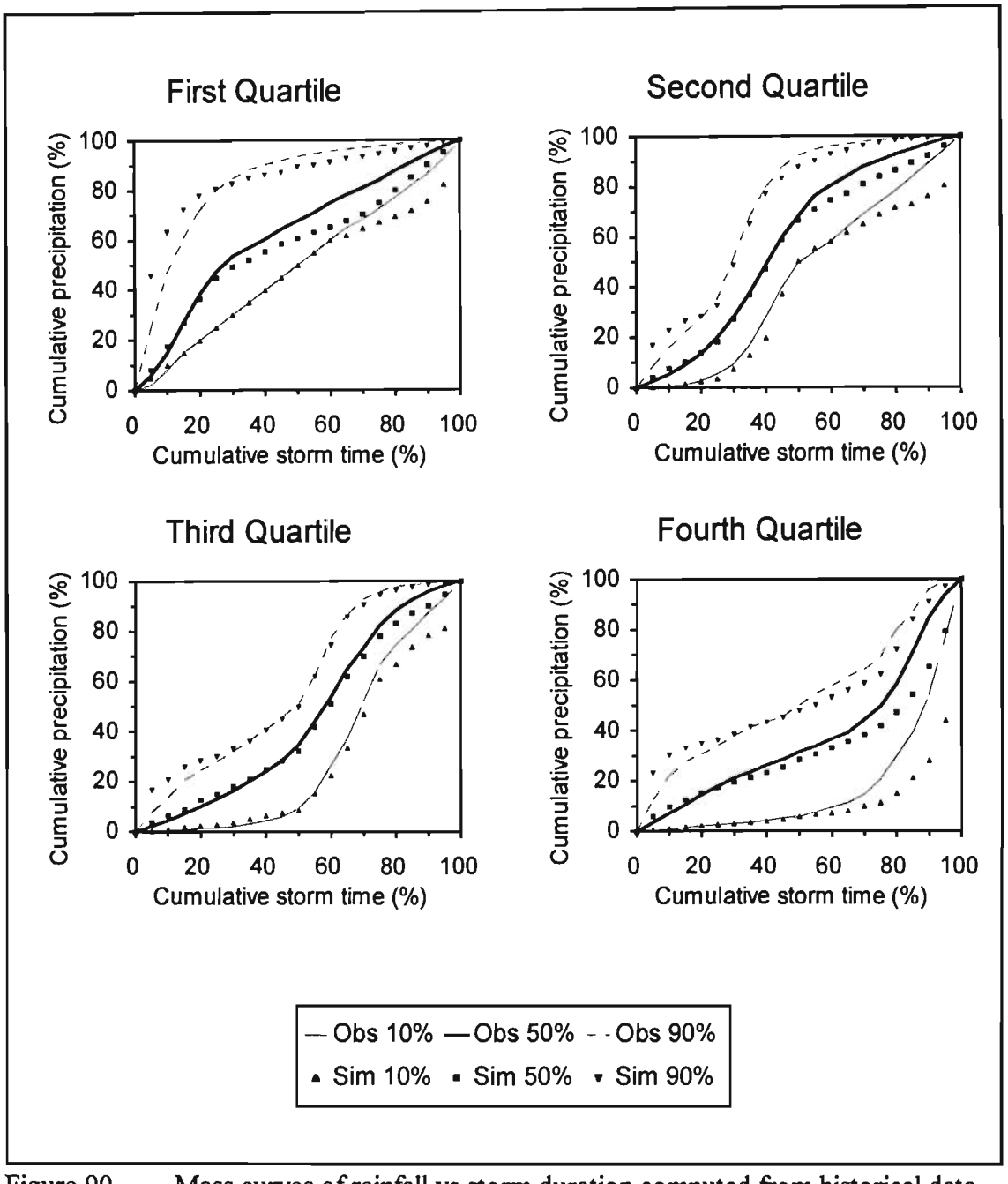


Figure 90 Mass curves of rainfall vs storm duration computed from historical data and from synthetic rainfall series generated by BLRPGM (parameter Set 2f) at Jnk19A

7.9.3 Mokobulaan (Moko3A)

As shown in Figure 91, the BLRPGM, with parameters derived using Set 2f underestimated the frequency of Quartiles 2 and 3 storms and overestimated the frequency of occurrence

of Quartiles 1 and 4 storms at Moko3A. The distribution and duration of storm depths was well simulated by the model, as shown in Figure 92. The mass curves computed from the historical data and synthetic rainfall series at Moko3A, shown in Figure 93, indicate that the synthetic rainfall storms generated by the BLRPGM have a similar distribution to the historical storms.

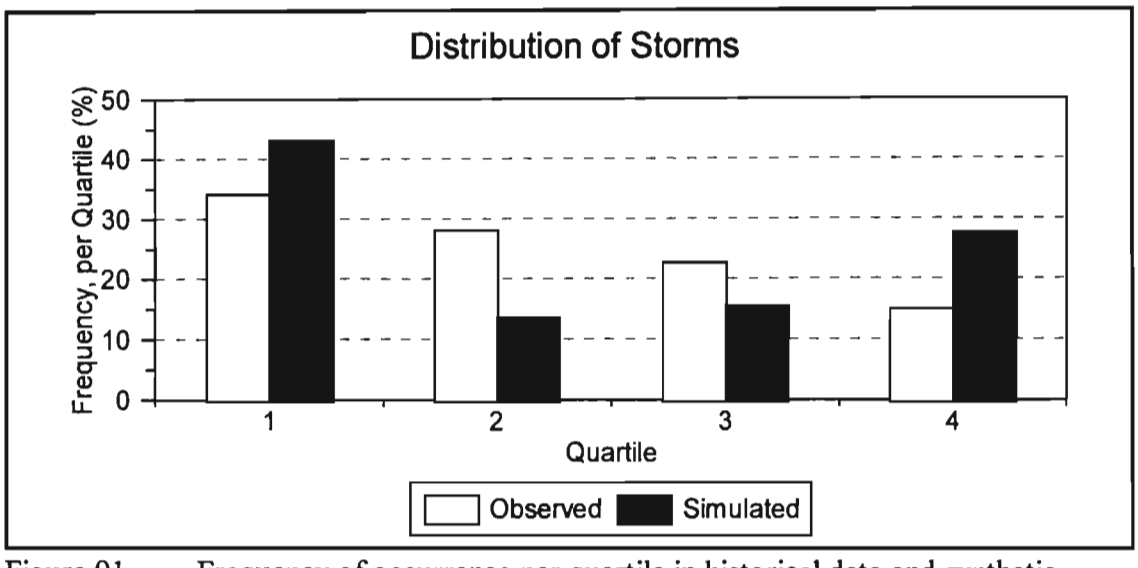


Figure 91 Frequency of occurrence per quartile in historical data and synthetic storms series generated by BLRPGM (parameter Set 2f) at Moko3A

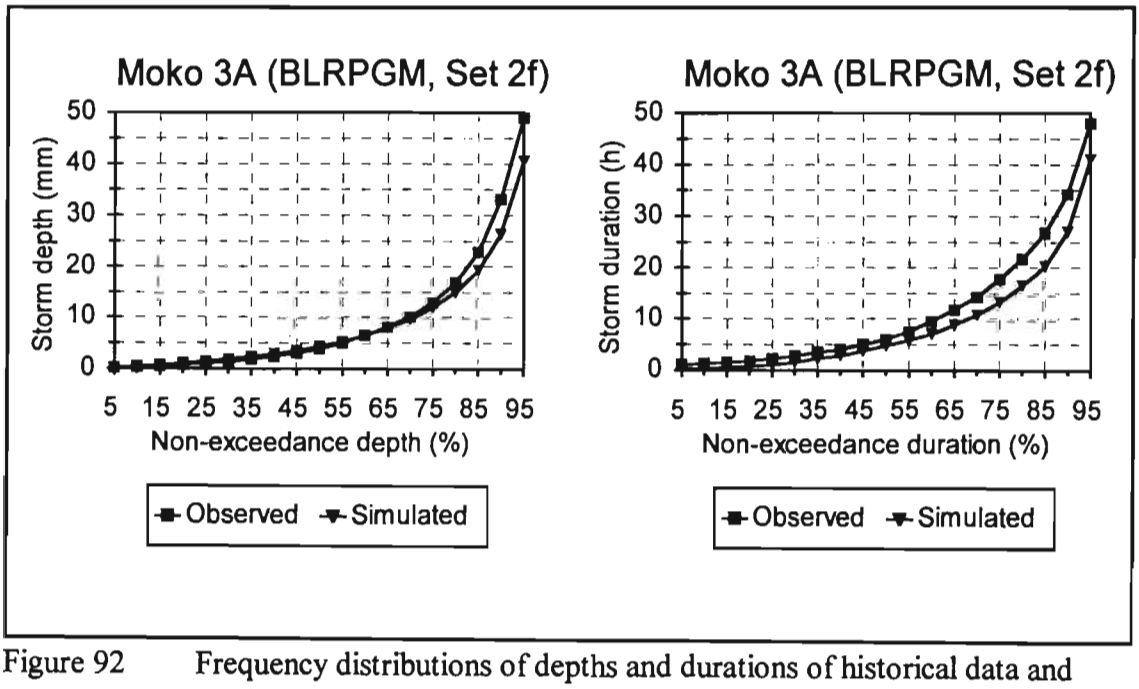


Figure 92 Frequency distributions of depths and durations of historical data and synthetic series generated by BLRPGM (parameter Set 2f) at Moko3A

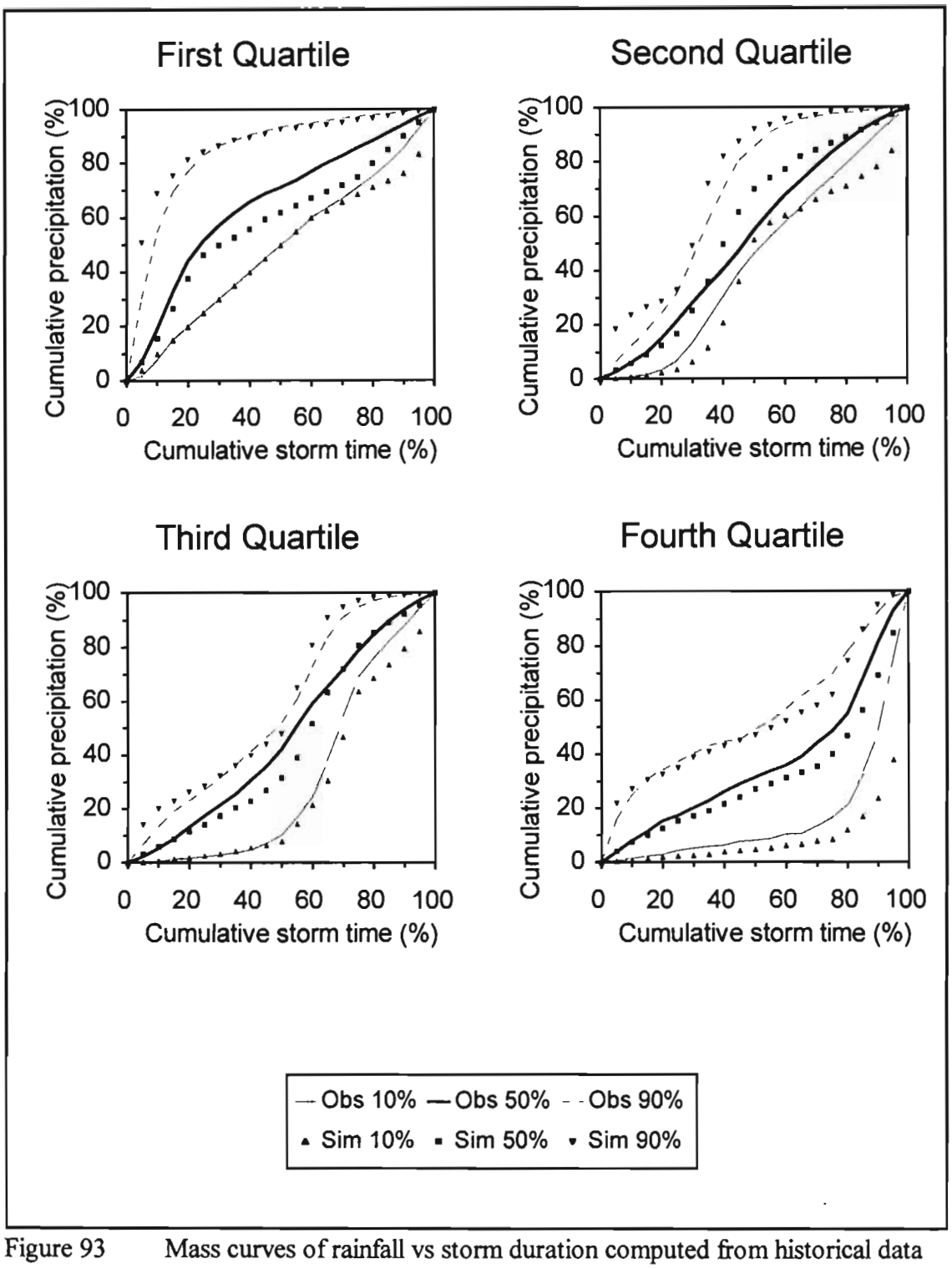


Figure 93 Mass curves of rainfall vs storm duration computed from historical data and from synthetic rainfall series generated by BLRPGM (parameter Set 2f) at Moko3A

7.9.4 Concluding Remarks on Temporal Distribution of Storms

The temporal distribution of historical and synthetic storms generated by the BLRPGM have been presented for three locations (Ntabamhlope, Jonkershoek and Mokobulaan) in very different climatic regions in South Africa. At all three locations the frequency of occurrence of storms in the different quartiles were different to those found in the historical data. However, at all three sites, the mass curves of the synthetic rainfall series and the frequency of rainfall depths and event durations matched the historical values very well for all quartiles. Hence it is concluded that temporal distribution of synthetic storms generated by the BLRPGM, with parameters determined from daily rainfall data, match the historical storms relatively well and can be used to estimate hyetographs..

It has been established that design rainfall depths for durations $\geq$ 1 h estimated from the synthetic rainfall series generated by the BLRPGM with parameter Set 1f, and in most cases Set 2f, correspond closely with those computed from the observed data. In the next section the optimisation of parameters to improve the estimation of design events from the synthetic rainfall series is investigated.

7.10 PARAMETER OPTIMISATION

In order to improve the simulations by BLRP models and to make the identification of parameters unique and better defined, three parameter optimisation strategies were evaluated at three selected stations. These were based on the moments of the AMS and on the characteristics of the events.

7.10.1 Annual Maximum Series

The magnitudes of design storms are a function of the statistical characteristics of the AMS and hence are a function of the mean, standard deviation and skewness of the AMS. Hence the parameters were optimised using a two-stage procedure. Initially the parameters were estimated as described in Section 7.6. Then one of the parameters associated either with cell intensity or duration was varied and the remaining parameters determined for discrete values of this parameter. For the BLRPGM the index of the gamma distributed cell intensity (δ) was kept constant. For each set of parameters determined for a single pre-determined parameter, a rainfall series was simulated with a record length equal to the historical data and the moments of the simulated and historical AMS were compared using the statistic Z defined in Equation 73 (Section 7.1). The first three moments (mean, variance and skewness) of the AMS of the observed data and simulated series for varying durations were used in the calculation of Z. Hence for each set of parameters, and for a constant value of the selected parameter, a value of Z was computed which reflected the difference in the moments of the historical and simulated AMS. The optimum parameter set selected was thus the set which resulted in the minimum value of Z. This optimisation procedure was termed Opt1.

7.10.2 Event Characteristics

Usually only four moments (mean, variance, autocorrelation and dry probability) for different levels of aggregation (duration) were used in the estimation of model parameters. Onof *et al.* (1994) presented analytical expressions of event duration, inter-event duration and mean number of events for the BL models. Onof and Wheater (1994a) and Onof and Wheater (1994b) adopted a two-stage procedure whereby, for incremental values of a fixed parameter, the remaining parameters were determined and the statistic Z in Equation 73 was computed for each solution using the event characteristics. A similar approach was adopted in this study and termed Opt2.

In an extension to this approach, instead of using a two-stage search approach, the event characteristics were used directly in the estimation of parameters in addition to the other moments. This procedure was termed Opt3. For example, if the model parameters were determined using moment Set 1e and optimised using the Opt3 procedure, the 1 h event duration and number of events would be used in addition to the moments in Set 1e in the determination of parameters. Similarly, if the parameters of the model were determined using moment Set 2f, which assumed that only daily rainfall data were available at the site, then the 24 h event duration and number of events would be used in addition to the moments in Set 2f in the determination of parameters. The three parameter optimisation techniques have been evaluated at a number of sites and the results are presented below.

7.10.3 Ntabamhlope (N23)

The effects of attempting to improve the simulations using the optimisation strategies outlined above were investigated at raingauge N23. The study was limited to the BLRPGM only and attempted to improve the estimation of parameters using moment Sets 1e and 2f. Owing to the vast amount of computing time required to implement Opt1, the procedure was limited to parameter Set 1e at N23. A comparison of the performance relative to the estimation of design rainfalls for the two sets of parameters and the effect of optimising the parameters is shown in Figure 94. Parameter optimisation had relatively little effect on the estimation of design events at raingauge N23, although Opt3 applied to Set 2f parameters performed slightly better than any of the other parameter estimation methods.

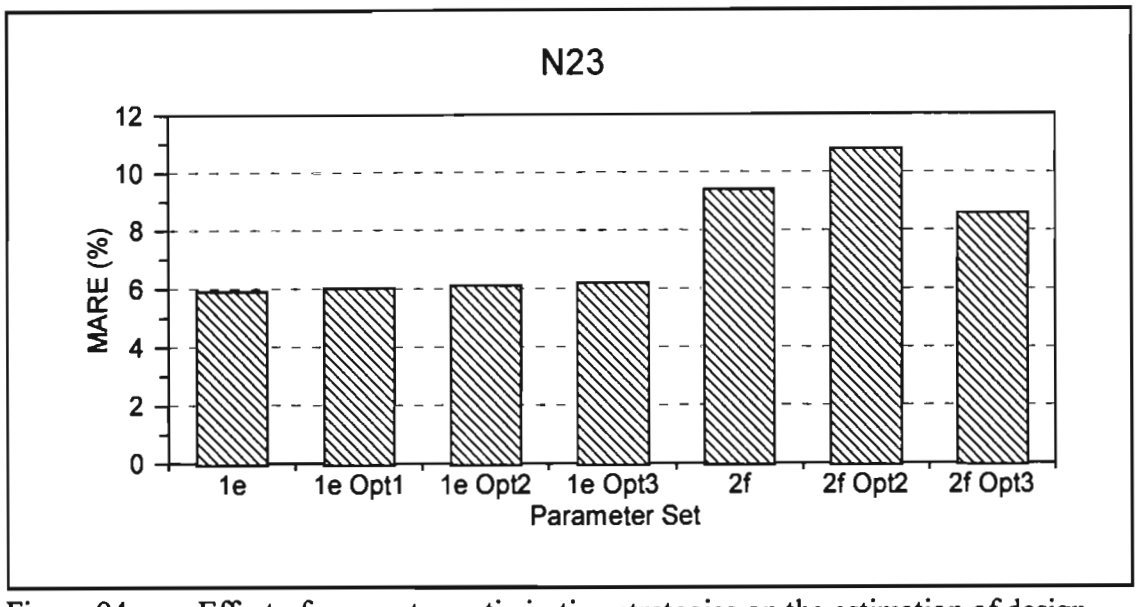


Figure 94 Effect of parameter optimisation strategies on the estimation of design rainfalls at N23

7.10.4 Cedara (C182)

The effect at C182 on design rainfall values, estimated using the BLRPGM with parameters determined using moments Sets 1e and 2f, of the Opt2 and Opt3 parameter optimisation strategies are shown in Figure 95. The parameter optimisation strategies had no effect on the estimation of design rainfall values for the Set 1 parameters, but did improve the *MARE* values for Set 2f parameters, with Opt3 giving the smallest *MARE* value.

7.10.5 Jonkershoek (Jnk 19A)

The effect at Jnk19A on design rainfall values of the Opt2 and Opt3 parameter optimisation strategies, estimated using the BLRPGM with parameters determined using moments Sets 1e and 2f, are shown in Figure 96. At Jnk19A the use of the Opt3 strategy improved the estimation of design rainfall values for both the Set 1e and Set 2f parameters.

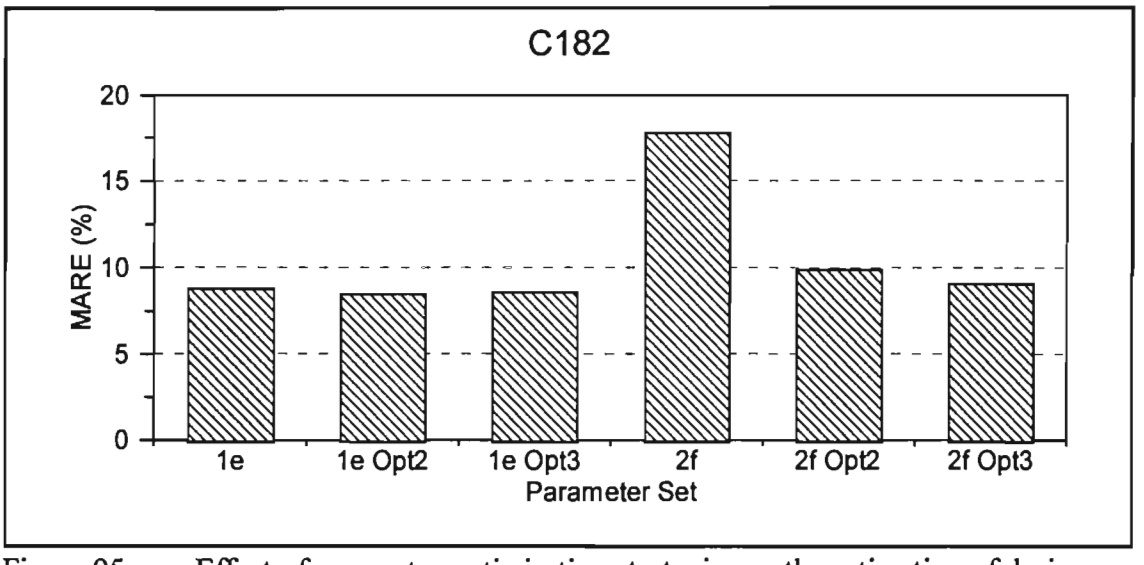


Figure 95 Effect of parameter optimisation strategies on the estimation of design rainfalls at C182

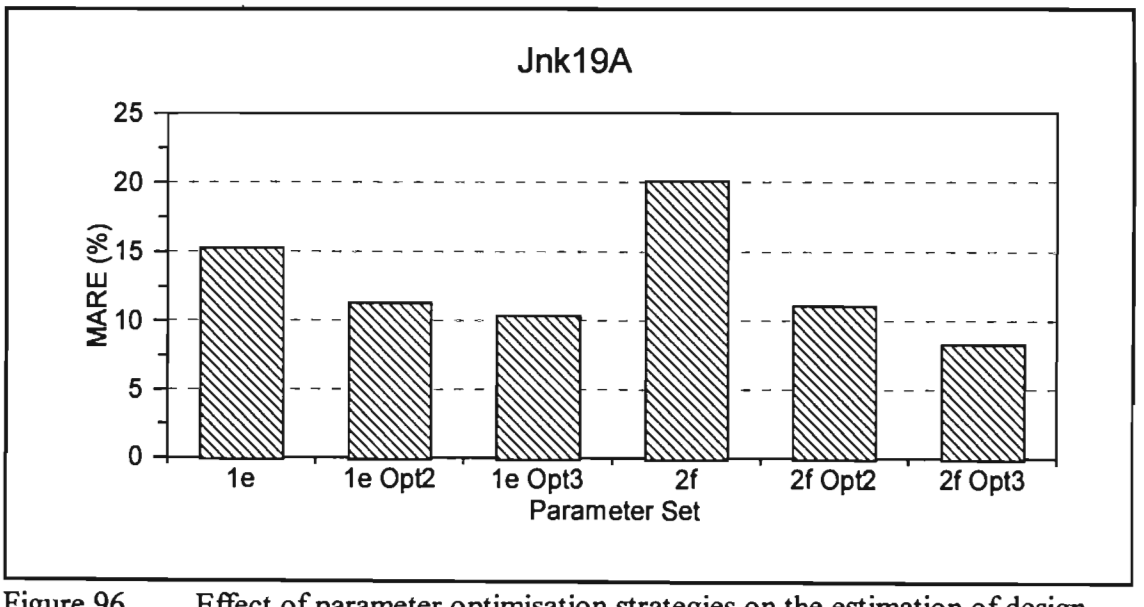


Figure 96 Effect of parameter optimisation strategies on the estimation of design rainfalls at Jnk19A

7.10.6 Concluding Remarks on Parameter Optimisation

Of the three parameter optimisation strategies evaluated, the Opt3 strategy, which includes the event duration and number of events directly in the parameter determination procedure,

resulted in the best estimation of design rainfall depths using the BLRPGM model. In cases where the BLRPGM with non-optimised parameters resulted in *MARE* values < 10% (e.g. Set 1e at N23 and C182), relatively little improvement was gained using the optimised parameters. However, in cases where the non-optimised parameters resulted in poorer estimation of design rainfalls (e.g Jnk19A), the Opt3 parameter estimation procedure improved the estimation of design rainfalls. Hence it is recommended that the Opt3 parameter determination procedure should be adopted in future use of the BLRPGM in South Africa

Frequently at a site where an estimate of design rainfall is required, only a short period of data is available. In the absence of regional schemes for estimating design events at the site, the design values are estimated using the short period of record, which may include the estimation of design values for return periods far in excess of the period of record. In the next section the use of the BLRPGM to estimate design storms from a short period of record vs the estimation of the design storms directly from the short record is investigated.

7.11 EXTENDING SHORT RECORD LENGTHS

The use of a short record length (e.g. ≤ 10 years) to estimate design events for return periods greater than twice the record length (e.g. ≥ 20 years) is generally not recommended. However, if only a short period of record is available at the site of interest and regional and other techniques of estimating the design event at the site are not available, then the design events would have to be estimated from the short period of available record.

This section investigates, by way of two case studies, whether a design event would be better estimated from the short record or if the design event would be better estimated by using the short record to estimate the parameters of the BLRPGM and then computing the design event from the synthetic rainfall series generated by the BLRPGM.

7.11.1 Ntabamhlope (N23)

The first case study utilised the 32 year rainfall record from raingauge N23 at Ntabamhlope, which is located in a summer rainfall region. Design storms for durations ranging from 15 min to 24 h were computed from both the entire record and using only the last 10 years of record. Similarly, parameters for the BLRPGM were derived using the full record and only the last 10 years of record. One hundred synthetic rainfall series were simulated for each set of the two sets of parameters, with the period simulated for each series equal to the record length used to derive the parameters (i.e. 32 and 10 years). The results of the study for the 50 year return period design storm for varying durations and for the 1 h design storm for varying return periods are shown in Figure 97. It is assumed that the best estimates of design rainfall are obtained from the full (32 year) period of record. From Figure 97, as shown by the 25-th and 75-th percentile range (high-low bars) of design values computed from the 100 synthetic series, it is evident that, at Ntabamhlope, the use of the BLRPGM, with parameters determined using only 10 years of data, to estimate the 50 year return period event would result in improved estimates of design storms, particularly for longer duration storms. Similarly for the relatively short 1 h duration event, the modelling approach would result in more reliable estimates of the design storms, particularly for larger return periods. Hence, based on the assumption that the design storms computed from the full record length are the best estimate of the true value, the use of the BLRPGM to estimate the design storms is recommended at Ntabamhlope.

7.11.2 Jonkershoek (Jnk 19A)

The second case study utilised the 54 year rainfall record from raingauge Jnk 19A at Jonkershoek, which is located in a winter rainfall region. The same analysis as described above was performed and the results of the study for the 50 year return period storm and 24 h design storms are shown in Figure 98. Again the estimation of design storms from the synthetic rainfall series simulated by the BLRPGM, with parameters determined using only

10 years of data, would result in more reliable estimates than direct estimation of design storms from the short period of data.

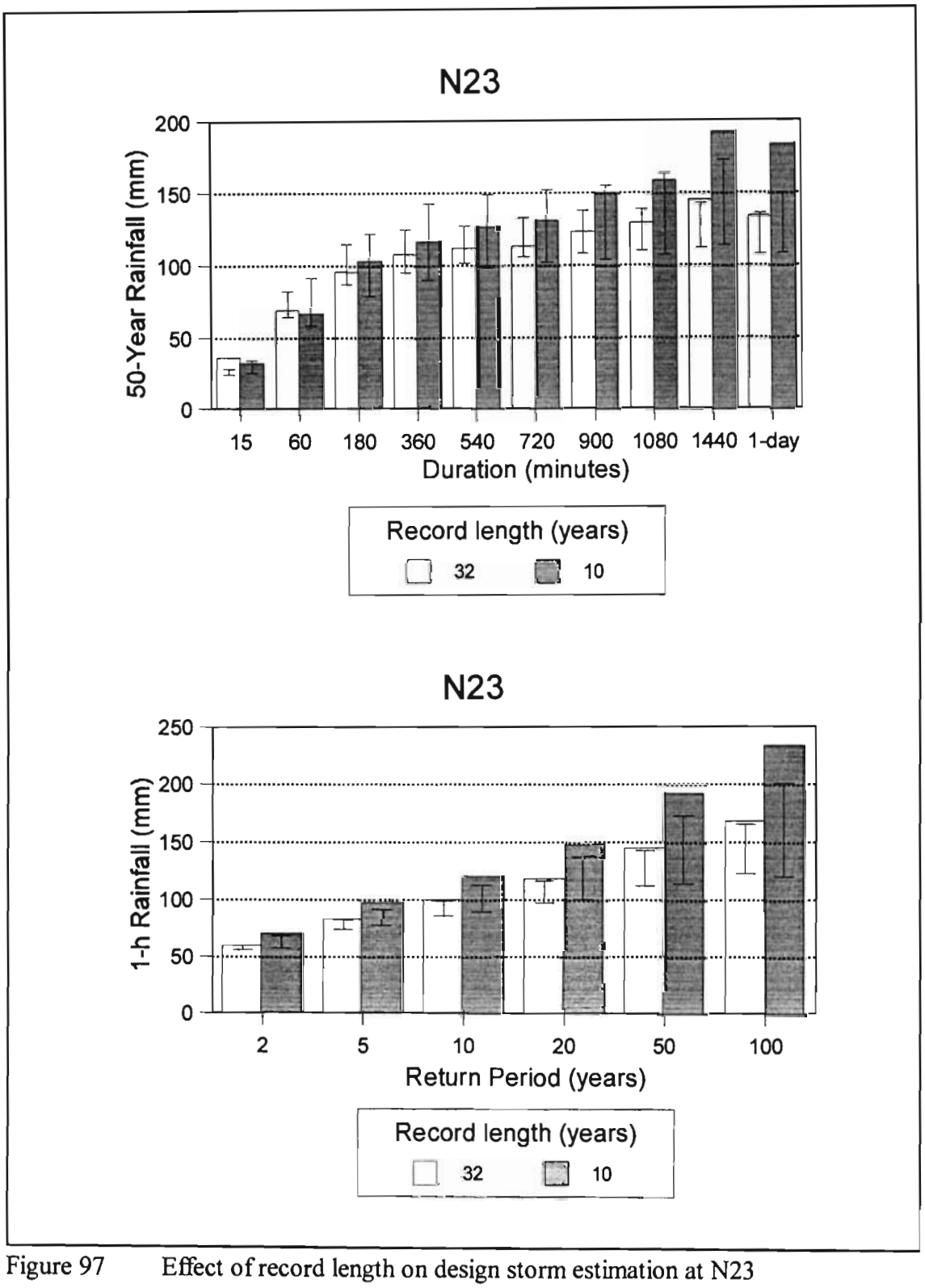


Figure 97 Effect of record length on design storm estimation at N23

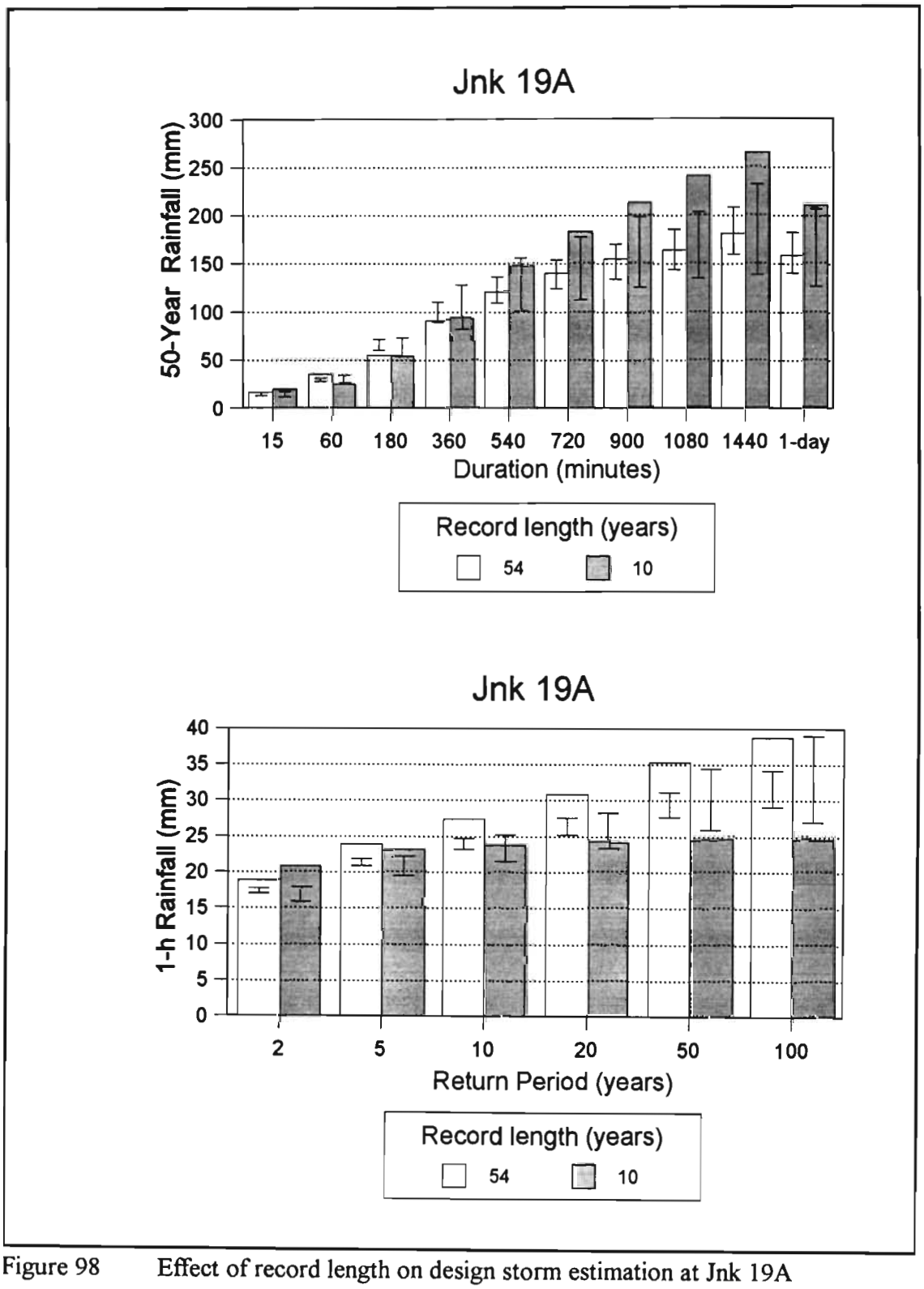


Figure 98 Effect of record length on design storm estimation at Jnk 19A

7.11.3 Concluding Remarks on Extending Short Record Lengths

In both case studies presented, the interquartile range of the design events estimated from the synthetic rainfall series, generated using parameters based on the shorter record length, resulted in better estimates of the "true" design values, estimated using the longer period of observed data, than had the design events been estimated directly from the shorter period of observed data. Thus it is concluded that, based on these two case studies, where only short periods of observed rainfall data are available, the design values should preferably be based on the synthetic rainfall series generated by the BLRPGM, with parameters estimated using the short period of data, than on estimating the design values directly from the short period of observed data.

7.12 CHAPTER CONCLUSIONS

The relationships between the parameters of the BL-models have been investigated and have revealed strong correlations between some parameters and hence some poorly defined parameters. Thus an incremental search strategy, with one of more parameters fixed, was successfully implemented to form a relatively robust technique to determine better defined parameters.

A comparison between the performances of the MBLRPM and BLRPGM was undertaken. The measures of performance used were analytical and simulated moments and the estimation of design rainfall events from the synthetic rainfall series generated by the models. It was noted that despite the BLRPGM requiring the estimation of an additional model parameter compared to the MBLRPM, the performance of the BLRPGM was generally less sensitive than the MBLRPM to the moments used to estimate the model parameters.

At a number of sites in different climatic regions in South Africa, the BLRPGM was shown to simulate synthetic rainfall series which fitted the statistics of the historical data better than

those computed from the series generated by the MBLRPM. Similarly, the design rainfall events estimated using the BLRPGM were better than those estimated using the MBLRPM. Generally the BLRPGM performed better when short duration digitised data were available to estimate the model parameters than when only daily rainfall data were available. However, the inclusion of variances for durations < 24 h, estimated from the daily data (Set 2f), generally resulted in adequate estimation of design rainfalls. The variances for short duration events were estimated using a linear relationship between the log of variance and log of duration. This generally resulted in poor estimates of variance for durations ≤ 1h. It is recommended that future research should consider adopting a curvilinear function, as proposed by Pegram (1998), and thus improve the estimates of variance for shorter durations.

Further improvements in the estimation of design rainfalls are possible by adopting the Opt3 parameter optimisation procedure, which includes event duration and number of events, in addition to other moments, directly in the determination of model parameters.

The temporal distribution of storms generated by the BLRPGM was found to closely match the observed data at three sites in different climatic regions in South Africa. However, the frequency of storms with particular profiles was not as well simulated as the temporal distribution. It is thus recommended that the use of the BLRPGM to estimate design rainfall values in South Africa, particularly for durations of 1h to 24 h, is a feasible option which can also be adopted at sites where only daily rainfall data are available.

The effect of record lengths on the estimation of design rainfall values was investigated at two sites in South Africa. In both cases, the design rainfall values estimated from the synthetic rainfall series generated by the BLRPGM, with parameters determined using a short period of record, better approximated the "true" design values, computed directly from the full period of observed record, than when the design values were computed directly from the short period of observed record. Thus it is recommended that, particular when only short periods of record are available and no other techniques of estimating short duration design rainfall values are available, design rainfall values should preferably be

computed using the synthetic rainfall series generated by the BLRPGM, with parameters estimated using the short period of data, than on estimating the design values directly from the short period of observed data.

CHAPTER 8

CONCLUSIONS AND RECOMMENDATIONS

The main objective of this project was to estimate short duration (≤ 24 h) design storms for South Africa. These were to be based on digitised rainfall data whereas previous studies conducted on a national scale in South Africa were based on data that were manually extracted from autographic charts. With the longer rainfall records currently available compared to the studies conducted in the early 1980s, it was expected that by utilising the longer, digitised rainfall data in conjunction with regional approaches, which have not previously been applied in South Africa, and new techniques such as L-moments, that more reliable short duration design rainfall values could be estimated. A short duration rainfall database was thus established for South Africa.

8.1 SHORT DURATION RAINFALL DATABASE

The short duration rainfall database currently consists of data from 412 stations and was constantly updated throughout the study as new data became available. The largest contribution to the database was from the South African Weather Bureau (SAWB). Some processing errors were found in the data from all the organisations which contributed data to the project. However, numerous errors in the digitisation of the autographically recorded rainfall, in addition to missing events in the SAWB data, resulted in a large portion of the database to be viewed as being of low reliability. This is particularly pertinent in the estimation of extreme events, as the autographic raingauges tend to malfunction during intense events. It is expected that the conversion of the recording rainfall network from autographic raingauges to data logger recorded rainfall systems will not only improve the reliability of the data, with a smaller probability of errors introduced into the data during the processing stage, but will also improve the temporal and depth resolution of the recorded rainfall data. It is estimated that the minimum temporal resolution of the autographically recorded and digitised rainfall data from charts changed on a daily basis may be as small

as 5 minutes and proportionately larger for charts changed once a week. In this study the minimum event duration analysed was 15 minutes. However, all the techniques evaluated in this study can generate design storms for durations shorter than 15 minutes, but the results should then be used with caution.

The majority of the errors identified in the SAWB data were negative and zero time steps (infinite intensities). Techniques were developed to identify the errors and make adjustments to the data points to enable smooth, automatic screening and processing of the data. The adjustments initially made an attempt to identify the probable cause of the error and, if successful, to make adjustments automatically in accordance with the nature of the probable cause of the error. If the probable cause of the error could not be identified a procedure was developed to make adjustments automatically such that a random selection of either the maximum, average or minimum intensity was introduced into the data as a result of adjusting the data points. The effect of making the adjustment on estimated design storms was shown not to be significant, but the exclusion of any event that had an error contained within it did result in a significant difference, thus indicating that the events should be retained and errors corrected.

A comparison at selected sites of manually extracted and digitised Annual Maximum Series (AMS) and the differences between rainfall totals recorded in the daily and digitised databases led to the conclusion that the digitised SAWB data were generally of low reliability and contained numerous periods of missing data. These periods of missing data were noted to extend over the whole range of events and were not confined to the smaller events. The effect of missing periods of data on the estimation of design storms was investigated at a selected site (East London) which had a long (> 50 years) period of record and which was judged to be in the top 5% of most reliable SAWB stations. In the analysis, a selected number of events for a selected number of years in the AMS were excluded and the differences in the estimated design values led to the disappointing conclusion, which is supported by other evidence throughout the document, that the digitised SAWB data were generally not adequate for estimating design storms for durations ≤24 h. This led to the

development of a three-pronged approach for estimating design storms from an inadequate database.

8.2 SHORT DURATION DESIGN RAINFALL ESTIMATION

The three approaches developed to estimate short duration design rainfall values were all based on the assumption that the daily manually recorded rainfall database was more reliable than the short duration rainfall database. An added advantage of using the daily rainfall database to estimate short duration design storms is the relatively dense network of daily raingauges available in South Africa which generally have much longer records than the short duration rainfall database.

8.2.1 Regional Approach

The first approach used an index-storm based regional L-moment algorithm developed by Hosking and Wallis (1993; 1997) to estimate design storms for various durations and results for South Africa were presented in Chapter 5. The use of a regional approach has many claimed benefits, including robustness and improving the reliability of at-site design values. The underlying assumption when using an index-storm type approach is that homogeneous regions can be identified where the distribution of extreme events is the same, except for a local scaling factor. Thus 15 relatively homogeneous regions were identified in South Africa and the General Extreme Value (GEV) distribution was determined to be the most appropriate common distribution to use in all 15 regions. The homogeneous regions were successfully identified by an appropriately scaled cluster analysis of site characteristics which included indices of location, MAP, altitude, seasonality of rainfall, distance from the sea and concentration of rainfall. The advantage of using only site characteristics in the cluster analysis is that the clusters identified can be tested independently for homogeneity using data from the site. The 24 h duration rainfall data from the short duration rainfall database were used to establish the homogeneity of the clusters. It has been shown that the short

duration data from the SAWB is generally of low reliability and hence there may be some doubt as to the validity of the homogeneity tests which may have been based on unreliable data. It is intended that a future project will refine and extend the relatively homogeneous clusters identified in this study by performing a cluster analysis, similar to the regionalisation performed in this study, but based on the site characteristics of the locations of the daily rainfall gauges and the subsequent testing of the clusters identified for homogeneity using the daily rainfall data.

Quantile growth curves were developed for each of the 15 homogeneous regions for 16 durations ranging from 15 min to 24 h. The index used to scale the relationships was the mean of the AMS (*L_1*) for each duration. Thus, information from the entire region can be used to estimate design storms at a particular site by utilising the regional growth curve and the at-site *L_1* value. This approach lends itself to design storm estimation at ungauged sites if the index used to scale the relationship can be estimated at the site of interest. As an example, regression analyses were performed between the 24 h *L_1* values and rainfall related site characteristics which are readily available as 1'x1' images for South Africa (Schulze, 1997). The results of the regression analyses in 13 of the 15 clusters enabled the 24 h *L_1* values to be estimated reasonably confidently. It is recommended only *L_1* values determined from gauged data be used in Clusters 10 and 11, where the regression analyses were not successful.

The accuracy of the regional design storm estimates were assessed for one site (N23) in Cluster 3 which was not used in the regional analysis. It was found that at N23 the regional and at-site estimated design storms corresponded extremely well for all durations and return periods. This "hidden station" approach to testing the method was not used in the other clusters owing to the limited number of available stations, but this analysis is a qualified validation of the methodology.

The accuracies of the quantile Regional Growth Curves (RGC) were successfully established using a Monte Carlo type simulation of a hypothetical region which has the same number of stations and record lengths as the cluster under evaluation. In this manner 90 %

confidence intervals were established for both the regional growth curves and the estimated at-site design storms. The simulation of more than 100 hypothetical regions for each cluster may increase the reliability of the confidence intervals at the expense of more computing time.

8.2.2 Scaling of L-moments

The second approach to estimating design storms with an inadequate database was to investigate the scaling relationships between the moments of the AMS and rainfall event duration and results using this approach were reported in Chapter 6. Previous studies have used this approach to interpolate design values from published durations to other durations and have used conventional product moments in deriving the relationships. It was noted at selected sites from different climatic regions in South Africa that the log-transformed relationship between L-moments and duration was more linear over a wider range of durations than when conventional moments were used. Thus, the use of L-moments was adopted for this application in the study.

Six hypotheses were proposed and evaluated at selected sites in each of the relatively homogeneous clusters. Hypothesis 1 proposed that the L-moments for durations < 24 h could be derived directly from the 24 h and 48 h L-moments, which can be computed from the daily rainfall data. It was found that the slope of the relationship for durations from 1h to 24 h was frequently different to the slope computed for durations ≥24 h and hence the L-moments for durations < 24 h could not be reliably estimated directly from the 24 h and 48 h values at all sites.

It was noted that the slopes of the log transformed L-moment:duration relationship at different sites within a cluster tended to be similar. Multiple linear regression relationships were thus developed for each cluster to estimate the regression slope of the log-transformed *L_1* and *L_2*:duration relationships as a function of site characteristics. The slopes at site *i* estimated as a function of the site characteristics were termed the Regional Slopes, $RS_{(1,i)}$ and $RS_{(2,i)}$ for the *L_1* and *L_2* relationships respectively. Reasonably good relationships

were obtained for 13 of the 15 clusters. However, Clusters 1 and 11 had coefficient of determination values < 0.5 for both the L_1 and L_2 regressions. Case studies using the *RS* at selected sites in Cluster 11 yielded acceptable results despite the poor estimation of the *RS* in Cluster 11. Hence the use of the *RS* could be used with caution to estimate L-moments in Clusters 1 and 11.

In Hypothesis 2 the *RS* and 24 h L_1 and L_2 values, computed from the observed digitised data, were used to estimate the first two L-moments for durations < 24 h.

Hypotheses 4, 5 and 6 all utilise the regional average L-moments, which are record length weighted averages of the L-moments computed for the AMS, scaled by the mean of the AMS (L_1), for each duration at each site. Thus the first regional average L-moment (L_1^R), being the regional average of the first at-site L-moments, which are scaled by L_1, is equal to 1. These hypotheses differ in the manner in which the regional average L-moments (L_x^R) are re-scaled at each site.

Hypothesis 3 assumed that the observed $L\_1_{(i,D)}$ values for each duration (D) were available at each site (i) in order to re-scale $L\_1^R_{(D)}$ and $L\_2^R_{(D)}$ and thus estimate the first two L-moments at each site. Hypothesis 4 estimated the at-site $L\_1_{(i,24)}$ value using regional regression relationships and site characteristics and $L\_1_{(i,D)}$ values for durations < 24 h were then computed using the estimated $L\_1_{(i,24)}$ value and the $RS_{(1,i)}$. The $L\_1_{(i,D)}$ value for each duration estimated in this manner was then used to re-scale the relevant $L\_1^R_{(D)}$. Instead of estimating $L\_1_{(i,24)}$ from site characteristics, Hypothesis 5 estimated this value directly from the daily data and then used the same procedure as Hypothesis 4 to estimate $L\_1_{(i,D)}$ for shorter durations, which were then used to re-scale $L\_x^R_{(D)}$, where $x \leq 2$. Similarly, Hypothesis 6 used the 1 day L_1 value computed from the daily data and adjusted this value into $L\_1_{(i,24)}$ using regionalised 24 h : 1 day L_1 ratios, which compensate for the differences between the AMS extracted from rainfall recorded continuously (24 h) and at fixed intervals (1 day). Thus Hypotheses 4 - 6 utilised different techniques to estimate the $L\_1_{(i,D)}$ values for durations ≤24 h in order to re-scale the $L\_x^R_{(D)}$ at sites where only daily rainfall data are available. In addition Hypothesis 4 can be applied to a site that has no gauged data. In order

to fit distributions with more than two parameters, Hypothesis 4, 5 and 6 assume that third and higher order L-moments can be estimated using the regional, average, record length weighted L-moment ratios at all sites.

Hypothesis 1 is intuitively the most attractive as it is the simplest of the hypotheses evaluated. Although this hypothesis was found to be adequate at a number of sites in different climatic regions (e.g. Cathedral Peak, Newlands, Mokobulaan), breaks in linear scaling for durations < 24 h and ≥ 24 h at a number of stations (e.g. Ntabamhlope, Cedara, Mount Edgecombe) resulted in the rejection of the hypothesis for general use in South Africa.

The estimation of the *RS* for *L_1* and *L_2* from regionalised regressions and site characteristics, as used in Hypotheses 2, 4, 5 and 6, did not appear to adversely influence the estimation of design storms even in regions where weak relationships were obtained.

Hypothesis 4 is the only method evaluated that can be applied at an ungauged site within a cluster and would be expected to yield reasonable estimates of the at-site L-moments and hence design storms within a homogeneous region. Generally, at sites where the data were deemed to be reliable, the method performed well. However, at most SAWB stations where the method was evaluated, the hypothesis did not perform well as the L-moments computed from the 1 day data were larger than the L-moments computed from the digitised data. This anomaly is attributed to periods of missing digitised data for those stations. The errors in the digitised data from numerous SAWB stations also resulted in Hypotheses 2, 3 and 4 generally not performing well at these sites when compared to the L-moments and design storms estimated from the 1 day rainfall data.

All the hypotheses evaluated assume that the L-moment:duration relationship is linear when plotted as log-transformed values. This power law relation appears to hold true for most clusters over the range from 4 to 24 h. However, a change in the linear relationship at durations ranging from 1 to 4 h was noted at most summer rainfall sites (e.g. Ntabamhlope, Cedara, Kokstad, Mokobulaan and Drieplotte), where thunderstorms are the predominant

rainfall generating mechanism. In the winter rainfall region (e.g. Jonkershoek, Cape Town and Vredendal), where frontal rainfall systems predominate, the deviation in linear scaling at a particular duration is not as marked. Although deficiencies in the temporal resolution of the rainfall measurement and digitisation processes cannot entirely be discounted as the cause of the change in linear scaling, it is postulated that the phenomenon is mainly the result of the predominant rainfall generating system. The durations at which the breaks occur at a particular site are hypothesised to be related to the typical duration of thunderstorm activity. Thus it is recommended that Hypotheses 4 to 6 should not be used to estimate design rainfall values for durations < 2 h, particularly in clusters where thunderstorms are the predominant rainfall generating mechanism.

Hypothesis 6 requires that the 24 h L_1 value computed from the daily rainfall data be converted into a continuous 24 h value, as would be estimated from the digitised data. Although different conversion factors for each cluster were used in this study, it is recommended that a value of 1.20 could be used to convert 1 day to 24 h L_1 values in South Africa

It is postulated that the method outlined in Hypothesis 6, which performed well in all clusters and attempts to compensate for errors and periods of missing digitised rainfall data, will yield the most accurate estimates for design storms of the hypotheses evaluated and should be adopted in the estimation of design storms. Although Hypothesis 6 requires daily rainfall data and cannot be applied at sites which have no rainfall data, as is the case with Hypothesis 4, the dense network of daily rainfall stations with relatively long records used in conjunction with Hypothesis 6, enables the estimation of short duration design storms at a large number of locations in South Africa. The estimation of regional regression relationships to estimate the 1 day L_1 value, computed from the daily rainfall data, as a function of site characteristics would enable Hypothesis 6 to be applied at any location in South Africa.

An option not pursued in this study, but which warrants further investigation, is the use of stochastic daily rainfall models, as have been developed for South Africa by Zucchini *et al.*

(1992), to simulate daily rainfall series. The stochastically generated daily rainfall model would thus enable Hypothesis 6 to be applied at any ungauged location in South Africa.

As discussed in Chapter 6, it is assumed that the regional average L-moments and *RS* estimated from the digitised data are sufficiently reliable to be used despite the numerous deficiencies illustrated in the digitised SAWB rainfall database. It was shown in Chapter 2 that the errors in the daily totals of rainfall computed from the digitised database occurred over a wide range of values. It is probable that the wide range of event totals where errors occurred is associated with a wide range of event durations. Thus it is postulated that RGC and RS are probably reasonable estimates of their "true" values as events over all durations are affected by the periods of missing data. It is noted in Chapter 6 that it is probable that design storms estimated directly from the SAWB digitised data would, on average over durations ranging from 2 h - 24 h at most stations considered, have underestimated short duration design storms by up to 65 %.

8.2.3 Stochastic Rainfall Modelling

In the third approach to short duration design rainfall estimation, with results reported in Chapter 7, two variations of Bartlett-Lewis type of intra-daily stochastic models were used to generate synthetic series of rainfall. The series were accumulated at 1 minute intervals within the models and output at 15 minute incremental totals in order to conserve disk space and subsequent processing time.

The estimation of the parameters of the models proved to be an exacting task with similar performance possible with very different sets of parameters. The constrained parameter search technique developed in this study ensured that the mean storm characteristics computed from the derived parameters were reasonable and aided in the determination of parameters. The parameters estimated by function minimisation were found to be relatively sensitive to the initial estimates of parameters at the start of the minimisation procedure and the parameter search technique adopted assisted in overcoming this sensitivity. It became

clear that the unconstrained minimisation procedures frequently used in the literature are reliant on the careful selection of initial conditions. The explicit presentation of the relationships between the model parameters and the methods used to estimate the parameter correlation matrix are not evident in the literature reviewed. The correlation matrix assisted in the determination of model parameters by identifying parameters that were highly correlated and which could thus be fixed.

Despite the utilisation of these parameter determination procedures, the parameters for some months at some stations were difficult to estimate. This can only be attributed to the unsuitability of the model to the data which, in the range of locations and months where the parameters were relatively easily determined, is improbable, or to errors and missing periods of the data which alter the moments used in the estimation of parameters. Another problem encountered, particularly with the SAWB data, is that frequently a long period of record only contains relatively few individual months with no missing data and hence the reliability of the moments computed for the months is low, which in turn may affect the performance of the model.

The confidence intervals estimated by computing the 25-th and 75-th percentiles and thus explicitly showing the stochastic variation in the output from the models was not evident in the literature reviewed pertaining to stochastic rainfall models. Generally, other studies have only generated a single long synthetic rainfall series, frequently only for a single month of good data with a long record. In such cases, when the moments of the historical data have been reported in the literature, the determination of reasonable parameters similar to, or better than, those reported, were relatively easily obtained.

In this study, a means of assessing the fit and appropriateness of models to different data sets of varying reliabilities and from varying climates had to be devised and applied in a routine way. This was an ambitious task and was not achieved without difficulties. For example, the cost of estimating the stochastic confidence intervals in terms of computing time was enormous and the mainframe computing facilities provided by the Computing Centre for Water Research (CCWR) proved to be inadequate with most runs for a single

station generally taking longer than 24 h. Hence the super-parallel computing facilities provided at the University of Potchefstroom were utilised successfully.

A comparison between the performances of the Modified Bartlett-Lewis Rectangular Pulse Model (MBLRPM) and Bartlett-Lewis Rectangular Pulse Gamma Model (BLRPGM) was performed at selected sites in South Africa. The performance of the models and the ease of parameter determination were found to be sensitive to the composition of the moments used to determine the parameters of the model. It was noted that despite the BLRPGM requiring the estimation of an additional model parameter compared to the MBLRPM, the performance of the BLRPGM was generally less sensitive than the MBLRPM to the moments used to estimate the model parameters.

At a number of sites in different climatic regions in South Africa, the BLRPGM was shown to simulate synthetic rainfall series which fitted the statistics of the historical data better than those computed from the series generated by the MBLRPM. Similarly, the design rainfall events estimated using the BLRPGM were better than those estimated using the MBLRPM. Generally the BLRPGM performed better when short duration digitised data were available to estimate the model parameters than when only daily rainfall data were available. It was shown that the variances for durations < 24 h could be estimated directly from the 1 and 2 day values and were reasonably accurate at most locations tested for durations as short as 1 h. The use of only the daily rainfall, with the inclusion of variances for durations < 24 h estimated from the daily data (Set 2f), generally resulted in adequate estimation of design rainfalls. Further improvements in the estimation of design rainfalls are possible by adopting the Opt3 parameter optimisation procedure, which includes event duration and number of events, in addition to other moments, directly in the determination of model parameters.

The performance of both the MBLRPM and BLRPGM was generally better for durations close to those defining the moments used to determine the model parameters than for other durations, but did scale reasonably well to other durations.

Design storms were well estimated from the synthetic series generated from the BLRPGM at a range of sites in different climatic regions in the country. However, it is recommend that design storms for durations shorter than 1 h should not be estimated from the synthetic series generated by the BLRPGM, even when short duration rainfall data are available to estimate model parameters. In cases where only daily rainfall data are available to estimate the parameters of the model, it is recommended that design storms should not be estimated for durations shorter than 2 h and should be used with caution for durations from 2 to 6 h. It was evident from the results obtained that any anomalies in the historical data, as was often the case with the SAWB data, are highlighted by comparisons to the synthetic rainfall series. Thus it was shown in some cases that design storms estimated using the BLRPGM were more reliable than the design storms estimated using historical short duration data.

Design storms are only estimated well using the BLRPGM when the historical AMS contain no high outliers and hence the BLRPGM does not appear to work well at locations where a mixture of meteorological conditions cause extreme events. Thus the model performance does not appear to be adequate in areas where the variation in range of values in the AMS for a particular month is smaller for longer duration events than for shorter duration events.

The temporal distribution of storms generated by the BLRPGM was found to closely match the observed data at three sites in different climatic regions in South Africa. However, the frequency of storms with particular profiles was not as well simulated as the temporal distribution. It is thus recommended that the use of the BLRPGM to estimate design rainfall values in South Africa, particularly for durations of 1h to 24 h, is a feasible option which can also be adopted at sites where only daily rainfall data are available.

The effect of record lengths on the estimation of design rainfall values was investigated at two sites in South Africa. In both cases, the design rainfall values estimated from the synthetic rainfall series generated by the BLRPGM, with parameters determined using a short period of record, better approximated the "true" design values, computed directly from the full period of observed record, than when the design values were computed directly from the short period of observed record. Thus it is highly recommended that,

particularly when only short periods of record are available and no other techniques of estimating short duration design rainfall values are available, design rainfall values should preferably be computed using the synthetic rainfall series generated by the BLRPGM, with parameters estimated using the short period of data, than when estimating the design values directly from the short period of observed data.

An option not considered in this study, but one which would allow the BLRPGM to be applied at any location in South Africa, would be to generate daily rainfall series using stochastic models such as developed by Zucchini *et al.* (1992) and then to use the synthetic daily rainfall series to estimate the parameters of the BLRPGM.

In the following section design storms estimated using Hypothesis 6, which estimates design storms using a combination of the regional and scaling approaches, are compared to the design storms estimated from the synthetic rainfall series generated by the BLRPGM.

8.3 COMPARISON OF TECHNIQUES

The Mean Absolute Relative Error (*MARE*) between design rainfall values estimated using both Hypothesis 6 and the synthetic rainfall series generated by the BLRPGM, with parameters determined using Set 2f and optimised using the Opt 3 option as described in Section 7.10.2, and design values estimated from the historical data are shown in Figure 99 for selected stations where the data were deemed to be reliable. In the calculation of the *MARE*s, the 2, 10, 20 and 50 year return period values for durations of 2, 4, 6, 12 and 24 h were considered. It is evident from Figure 99 that design rainfall values computed using either Hypothesis 6 or from the synthetic rainfall series generated by the BLRPGM, with parameters estimated from daily rainfall data, are similar. Hence it is concluded that both methods are acceptable for estimating design storms in South Africa for durations > 1 h.

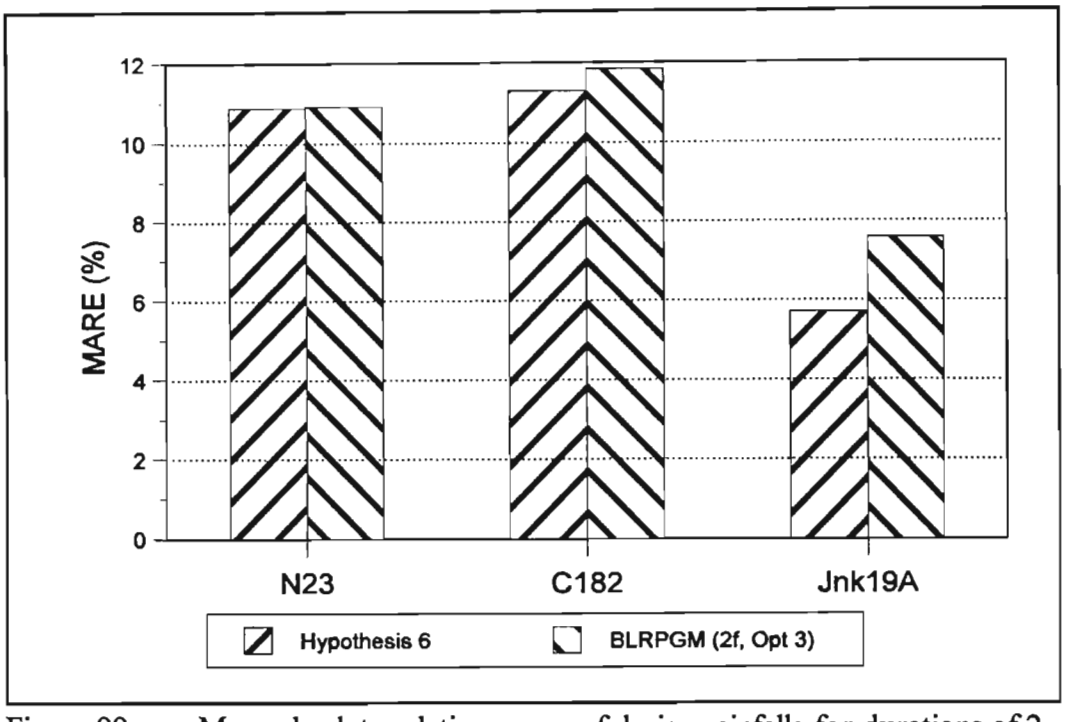


Figure 99 Mean absolute relative errors of design rainfalls for durations of 2 - 24 h and return periods of 2 - 50 years estimated at selected stations using Hypothesis 6 and the BLRPGM

8.4 RECOMMENDATIONS

All three approaches which were evaluated to estimate short duration design storms with an inadequate database performed well, considering the limitations of the data. However, the combined method of regional average L-moments and *RS*, scaled using an adjusted *L_1* value computed from the daily rainfall data (Hypothesis 6), is recommended for general use as it combines the strengths of the regional approach, which may compensate to some extent for stations with poor data, with the explicit attempt to compensate for the inadequate digitised data by using the *L_1* value computed from the daily data. It is also recommended that the BLRPGM be used at selected sites, in addition to the method detailed in Hypothesis 6, in order to ensure reasonable design estimates are obtained.

The results of the regional regression analyses used to estimate *L_1* and *RS* as a function of site characteristics may be affected by correlation between the independent variables and, in some clusters, by the limited number of stations and hence insufficient degrees of freedom

in the analyses. It is recommended that the selection of independent variables should be reviewed and simpler approaches to the regression analyses should be sought.

Hypothesis 6 can only be applied at sites which have daily rainfall data. It is recommended that regional relationships be developed to estimate the at-site 1 day *L_1* value, computed from the daily rainfall data, as a function of site characteristics, as reported in Section 5.4 for the 24 h *L_1* values, which were computed from the digitised rainfall data. This relationship, in conjunction with the regionalised 24 h : 1 day *L_1* ratios and *RS*, would enable reliable estimation of design storms for durations ≤ 24 h at any site in South Africa.

Design rainfalls estimated using the recommended approaches generally did not compare well to design values for durations shorter than 1 h. This suggests either that the data are more unreliable for shorter durations or that the techniques developed do not capture the characteristics of the extreme events for shorter time scales. It is therefore recommended that the techniques should be evaluated on more reliable, high resolution rainfall data such as recorded by data loggers, which may have to be obtained from sites not in South Africa.

The breaks in scaling at approximately 1 h and 24 h durations noted at many of the sites in South Africa, should be further investigated. Reliable, high resolution rainfall data should be obtained to further investigate the nature of these inconsistencies.

The variances for short duration events, used for determining parameters of the BLRPMs from the daily rainfall data, were estimated in this study using a linear relationship between the log of variance and log of duration. This generally resulted in poor estimates of variance for durations ≤1h. It is recommended that future research should consider adopting a curvilinear function, which may improve the estimates of variance for short durations and result in better model parameters and improved model performance.

It is further recommended that the stochastic daily rainfall models, as developed by Zucchini *et al.* (1992) for South Africa, should be evaluated as a technique to estimate the mean of the AMS at ungauged site, as is required for Hypothesis 6. If successful, this would provide

an alternative method for Hypothesis 4 in order to estimate design rainfall values for durations shorter than 24 h at ungauged sites. The stochastic rainfall series should also be evaluated as a means to determine the parameters of the BLRPGM, which in turn could be used to estimate short duration design rainfall values.

The 15 relatively homogeneous rainfall regions should be further verified and refined using the daily rainfall database for South Africa. The results in this study should then be adjusted to make use of a single set of homogeneous clusters for all durations.

CHAPTER 9

REFERENCES

Adamson, P.T., 1978. Extreme values and return periods for rainfall in South Africa. Technical Note No. 78, Department of Water Affairs, Pretoria, RSA.

Adamson, P.T., 1981. Southern African storm rainfall. Technical Report No. TR 102, Department of Water Affairs, Pretoria, RSA.

Alexander, W.J.R., 1978. Depth-area-duration-frequency properties of storm precipitation in South Africa. Technical Report No. TR 83, Department of Water Affairs, Pretoria, RSA.

Alexander, W.J.R., 1990. Flood Hydrology for Southern Africa. SANCOLD, Pretoria, RSA.

Arnell, N. and Beran, M., 1987. Testing the suitability of the TCEV distribution for regional flood estimation. Regional Flood Frequency Analysis, 159-175. D. Reidel, Amsterdam.

Aron, G., Wall, D.J., White, E.L. and Dunn, C.N., 1987. Design rainfall for Pennsylvania. In: A.D. Feldman (Editor), Engineering Hydrology. Hydraulics Division, ASCE, New York, USA.

Bell, F.C., 1969. Generalised rainfall-duration-frequency relationship. Journal of Hydraulic Division , ASCE, 95(HY1): 311-327.

Benson, M.A., 1968. Uniform flood-frequency estimating methods for federal agencies. Water Resources Research, 4(5): 891-908.

Bergman, N.M.C., 1974. A study of generalised rainfall-intensity relationships in the winter rainfall area. Report 33/74, Division of Agricultural Engineering Services, Stellenbosch, RSA.

Bergman, N.M.C. and Smith, A.R., 1973. A study of rainfall intensities in the winter rainfall area. Report 15/73, Division of Agricultural Engineering Services, Stellenbosch, RSA.

Blodgett, J.C. and Nasseri, I., 1995. Precipitation depth-duration characteristics, Antelope Valley, California. In: W.H. Espey and P.G. Combs (Editors), Water Resources Engineering, 1. ASCE, New York, USA, pp. 274-278.

Bo, Z., Islam, S. and Eltahir, E.A.B., 1994. Aggregation-disaggregation properties of a stochastic rainfall model. Water Resources Research, 30(12): 3423-3435.

Bobbee, B. and Robitaille, R., 1977. The use of the Pearson Type 3 Distribution revisited. Water Resources Research, 13(2): 427-443.

Buishand, T.A., 1991. Extreme rainfall estimation by combining data from several sites. Hydrological Sciences Journal, 36(4): 345-365.

Burlando, P. and Rosso, R., 1996. Scaling and multiscaling models of depth-duration-frequency curves for storm precipitation. Journal of Hydrology, 187: 45-65.

Burn, D.H., 1990a. An appraisal of the "region of influence" approach to flood frequency analysis. Hydrological Sciences Journal, 35(2): 149-165.

Burn, D.H., 1990b. Evaluation of regional flood frequency analysis with a region of influence approach. Water Resources Research, 26(10): 2257-2265.

Calles, B. and Kulander, L., 1995. Storm rainfall characteristics at Roma, Lesotho. South African Geographical Journal, 77(1): 6-11.

Cannarozzo, M., D'Asaro, F. and Ferro, V., 1995. Regional rainfall and flood frequency analysis for Sicily using the two component extreme value distribution. Hydrological Sciences Journal, 40(1): 19-42.

Canterford, R.P., Pescod, N.R., Pearce, H.J. and Turner, L.H., 1987a. Design Intensity-frequency-duration rainfall (Chapter 2). In: D.H. Pilgrim (Editor), Australian Rainfall and Runoff : A Guide to Flood Estimation. The Institution of Engineers, Barton, Australia.

Canterford, R.P., Pescod, N.R., Pearce, H.J., Turner, L.H. and Atkinson, R.J., 1987b. Frequency analysis of Australian rainfall data as used for flood analysis and design. In: V.P. Singh (Editor), Hydrologic Frequency Modeling. D. Reidel, Dordrecht, Germany.

Canterford, R.P. and Pierrehumbert, C.L., 1977. Frequency distributions for heavy rainfalls in tropical Australia, Hydrology Symposium, National Conference Publication No. 77/5. The Institution of Engineers, Barton, Australia, pp. 119-124.

Chandler, R., 1995. A spectral method for estimating parameters in rainfall models. Report No. 142, Department of Statistical Science, University College, London, UK.

Chandler, R., KaKou, A., Isham, V. and Northrop, P., 1995. Spatial-temporal rainfall processes: stochastic models and data analysis. Report No. 148, Department of Statistical Science, University College, London, UK.

Chen, C.L., 1983. Rainfall intensity-duration-frequency formulas. Journal of Hydraulic Engineering, 109(12): 1603-1621.

Chow, K.C., Ander and Watt, W.E., 1990. Knowledge-based expert system for flood frequency analysis. Canadian Journal of Civil Engineering, 17(4): 597-609.

Cowpertwait, P.S.P., 1991. Further developments of the Neyman-Scott clustered point process for modeling rainfall. Water Resources Research, 27(7): 1431-1438.

Cowpertwait, P.S.P., O'Connell, P.E., Metcalfe, A.V. and Mawdsley, J.A., 1996a. Stochastic point process modelling of rainfall. 1. Single-site fitting and validation. Journal of Hydrology, 175: 17-46.

Cowpertwait, P.S.P., O'Connell, P.E., Metcalfe, A.V. and Mawdsley, J.A., 1996b. Stochastic point process modelling of rainfall. 2. Regionalisation and disaggregation. Journal of Hydrology, 175: 47-65.

Cunnane, C., 1989. Statistical distributions for flood frequency analysis. WMO Report No. 718, World Meteorological Organization, Geneva, Switzerland.

Dent, M.C. and Schulze, R.E., 1987. The hydrological data manager and digitization in 1985: points to ponder in the development of a new digitizing system. Technical Note, Water SA, 13(1): 49-52.

Dwyer, I.J. and Reed, D.W., 1995. Allowance for discretization on hydrological and environmental risk estimation. Report No. 123, Institute of Hydrology, Wallingford, Oxfordshire, UK.

Entekhabi, D., Rodriguez-Iturbe, I. and Eagleson, P.S., 1989. Probabilistic representation of the temporal rainfall process by a modified Neyman-Scott rectangular pulses model: Parameter estimation and validation. Water Resources Research, 25(2): 295-302.

Ferrari, E., Gabriele, S. and Villani, P., 1993. Combined regional frequency analysis of extreme rainfalls and floods. In: Z.W. Kundzewicz, D. Rosbjerg, S.P. Somonovic and K. Takeuchi (Editors), Extreme Hydrological Events: Precipitation, Floods and Droughts. IAHS Press, Institute of Hydrology, Wallingford, UK, pp. 333-346.

Ferreri, G.B. and Ferro, V., 1990. Short-duration rainfalls in Sicily. Journal of Hydraulic Engineering, 116(3): 430-435.

Fletcher, R., 1987. Chapter 6: Sums of Squares and Nonlinear Equations. Practical methods of optimisation. John Wiley & Sons, New York, USA, 436 pp.

Foufoula-Georgiou, E. and Krajewski, W., 1995. Recent advances in rainfall modeling, estimation, and forecasting. Reviews of Geophysics, Supplement, July: 1125-1137.

Froehlich, D.C., 1995. Intermediate-duration-rainfall intensity equations. Journal of Hydraulic Engineering, 121(10): 751-756.

Froehlich, D.C. and Tufail, M., 1995. Rainfall intensity equations for durations from one to 24 hours. In: W.H. Espey and P.G. Combs (Editors), Water Resources Engineering, 1. ASCE, New York, USA, pp. 109-113.

Gabriele, S. and Arnell, N., 1991. A hierarchical approach to regional flood frequency analysis. Water Resources Research, 27(6): 1281-1289.

Gingras, D. and Adamowski, K., 1992. Coupling of nonparametric frequency and L-moment analyses for mixed distribution identification. Water Resources Bulletin, 28(2): 263-272.

Gingras, D. and Adamowski, K., 1994. Performance of L-moments and nonparametric flood frequency analysis. Canadian Journal of Civil Engineering, 21(5): 856-862.

Greenwood, J.A., Landwehr, J.M., Matalas, N.C. and Wallis, J.R., 1979. Probability weighted moments: Definition and relation to parameters of several distributions expressible in inverse form. Water Resources Research, 15(5): 1049-1064.

Griffiths, G.A. and Pearson, C.P., 1993. Distribution of high intensity rainfalls in metropolitan Christchurch, New Zealand. Journal of Hydrology (N.Z.), 31(1): 5-22.

Gupta, V.K. and Waymire, E.C., 1990. Multiscaling properties of spatial rainfall and river flow distribution. Journal of Geophysics Research, 95(D3): 1999-2009.

Guttman, N.B., 1992. Regional precipitation quantile values for the continental U.S. computed from L-moments. RC 18258, IBM Research Division, T.J. Watson Research Division, New York, USA.

Guttman, N.B., 1993. The Use of L-moments in the Determination of Regional Precipitation Climates. National Climate Centre, Asheville, North Carolina, USA.

Guttman, N.B., Hosking, J.R.M. and Wallis, J.R., 1993. Regional precipitation quantile values for the continental US computed from L-moments. Journal of Climate, 6(December): 2326-2340.

Hargreaves, G.H., 1988. Extreme rainfall for Africa and other developing areas. Journal of Irrigation and Drainage, 114(2): 324-333.

Henderson-Sellers, A., 1980. The spatial and temporal variation of rainfall intensity in South Africa. SA Geographer, 8(2): 109-112.

Hershfield, D.M., 1962. Extreme rainfall relationships. Proceedings of the American Society of Civil Engineers, HY6(11): 73-92.

Hershfield, D.M., 1982. 2-minute rainfall extremes, International Symposium on Hydrometoerology. American Water Resources Association, pp. 585-588.

Hershfield, D.M., 1984. Some statistical properties of short duration rainfall. In: P. Harremoes (Editor), Rainfall as a Basis for Urban Runoff Design and Analysis. Pergamon Press, Copenhagen, Denmark.

Hirsch, R.M., Helsel, D.R., Cohn, T.A. and Gilroy, E.J., 1993. Statistical analysis of hydrological data. In: D.R. Maidment (Editor), Handbook of Hydrology. McGraw-Hill, New York, USA.

Hogg, W.D., 1991. Time series of daily rainfall extremes, Fifth Conference on Climate Variations. American Meteorological Society, Boston, USA.

Hogg, W.D., 1992. Inhomogeneities in time series of extreme rainfall, Fifth International Meeting on Statistical Climatology. The Steering Committee for International Meetings on Statistical Climatology, Toronto, Canada, pp. 481-484.

Hosking, J.R.M., 1990. L-moments: analysis and estimation of distribution using linear combinations of order statistics. Journal of Royal Statistics Society, 52(1): 105-124.

Hosking, J.R.M., 1991a. Approximations for use in constructing L-moment ratio diagrams. RC-16635, IBM Research Division, T.J. Watson Research Centre, New York, USA.

Hosking, J.R.M., 1991b. Fortran routines for use with method of L- Moments Version 2. RC-17097, IBM Research Division, T.J. Watson Research Center, New York, USA.

Hosking, J.R.M., 1995. The use of L-moments in the analysis of censored data. In: N. Balakrishnan (Editor), Recent Advances in Life-Testing and Reliability. CRC Press, Boca Raton, Fl, USA, pp. 545-564.

Hosking, J.R.M., 1996. Fortran routines for use with method of L- Moments Version 3. RC-20525, IBM Research Division, T.J. Watson Research Center, New York, USA.

Hosking, J.R.M. and Wallis, J.R., 1987. An index flood procedure for regional rainfall frequency analysis. EOS, Transactions, American Geophysical Union, 68: 312.

Hosking, J.R.M. and Wallis, J.R., 1988. The effect of intersite dependence on regional flood frequency analysis. Water Resources Research, 24(4): 588-600.

Hosking, J.R.M. and Wallis, J.R., 1993. Some statistics useful in a regional frequency analysis. Water Resources Research, 29(2): 271-281.

Hosking, J.R.M. and Wallis, J.R., 1995. A comparison of unbiased and plotting-position estimators of L moments. Water Resources Research, 31(8): 2019-2025.

Hosking, J.R.M. and Wallis, J.R., 1997. Regional Frequency Analysis: An Approach Based on L-Moments. Cambridge University Press, Cambridge, UK, 224 pp.

Huff, F.H., 1967. Time distribution of rainfall in heavy storms. Water Resources Research, 3(4): 1007-1019.

James, E.J., Saseendran, S.A., Chandrasekharan, M.E. and Anitha, A.B., 1987. Rainfall frequency studies for Kerala region. Journal of the Institution of Engineers (India), 68, September: 74-81.

Karim, A. and Chowdhury, J.U., 1995. A comparison of four distributions used in flood frequency analysis in Bangladesh. Hydrological Sciences Journal, 40(1): 55-66.

Khaliq, M.N. and Cunnane, C., 1996. Modelling point rainfall occurrences with the modified Bartlett-Lewis rectangular pulses model. Journal of Hydrology, 180: 109-138.

Kite, G.W., 1988. Frequency and Risk Analysis in Hydrology. Water Resources Publications, Littleton, USA.

Kothyari, U.C. and Garde, R.J., 1992. Rainfall intensity-duration-frequency formula for India. Journal of Hydraulic Engineering, 118(2): 323-336.

Lettenmaier, D.P., 1985. Regionalisation in flood frequency analysis - Is it the answer ?, US-China Bilateral Symposium on the Analysis of Extraordinary Flood Events, Nanjing, China.

Lettenmaier, D.P. and Potter, K.W., 1985. Testing flood frequency estimation methods using a regional flood generation model. Water Resources Research, 21: 1903-1914.

Lettenmaier, D.P., Wallis, J.R. and Wood, E.F., 1987. Effect of regional heterogeneity on flood frequency estimation. Water Resources Research, 23(2): 313-323.

Lin, B. and Vogel, J.L., 1993. A Comparison of L-moments with method of moments. In: C.Y. Kuo (Editor), Engineering Hydrology. ASCE, New York, USA, pp. 443-448.

Markham, C.G., 1970. Seasonality of precipitation in the United States. Annals of Association of American Geographers, 60: 593-597.

McConachy, F., 1995. Estimation of extreme rainfalls for Victoria- .Application of Schaefer's method. Working Document 95/6, Cooperative Research Centre for Catchment Hydrology, Monash University, Clayton, Victoria, Australia.

McKerchar, A.I. and Pearson, C.P., 1990. Maps of flood statistics for regional flood frequency analysis in New Zealand. Hydrological Sciences Journal, 35(6): 609.

Menabde, M., Seed, A. and Pegram, G.G.S., 1998. A simple scaling model for extreme rainfall. in press.

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 Witwatersrand, Johannesburg, RSA.

Naghavi, B., Yu, F.X. and Singh, V.P., 1993. Comparative evaluation of frequency distributions for Louisiana extreme rainfall. Water Resources Bulletin, 29(2): 211-219.

Nandakumar, N., 1995. Estimation of extreme rainfalls for Victoria - Application of the Forge method. Working Document 95/7, Cooperative Research Centre for Catchment Hydrology, Monash University, Clayton, Victoria, Australia.

Nathan, R.J. and Weinmann, P.E., 1991. Application of at-site and regional flood frequency analyses, Challenges for Sustainable Development National Conference Publication. The Institute of Engineers, Barton, Australia., pp. 769-774.

NERC, 1975. Flood Studies Report, Natural Environment Research Council, London, UK.

Onof, C., Faulkner, D. and Wheater, H.S., 1996. Design rainfall modelling in the Thames catchment. Hydrological Sciences Journal, 41(5): 715-733.

Onof, C. and Wheater, H.S., 1993. Modelling of British rainfall using a random parameter Bartlett-Lewis rectangular pulse model. Journal of Hydrology, 149(1-4): 67-95.

Onof, C. and Wheater, H.S., 1994a. Improved fitting of the Bartlett-Lewis rectangular pulse model for hourly rainfall. Hydrological Sciences Journal, 39(6): 663-680.

Onof, C. and Wheater, H.S., 1994b. Improvements to the modelling of British rainfall using a modified random parameter Bartlett-Lewis rectangular pulse model. Journal of Hydrology, 157: 177-195.

Onof, C., Wheater, H.S. and Isham, V., 1994. Note on the analytical expression of the inter-event time characteristics for Bartlett-Lewis type rainfall models. Journal of Hydrology, 157: 197-210.

Op Ten Noort, T.H., 1983. Flood peak estimation in South Africa. The Civil Engineer in South Africa, October: 557-563.

Oyebande, L., 1982., 1982. Deriving rainfall intensity-duration-frequency relationships and estimates for regions with inadequate data. Hydrological Sciences Journal, 27(3): 353-367.

Pearson, C.P., McKerchar, A.I. and Woods, R.A., 1991. Regional flood frequency analysis of Western Australian data using L-moments, National Conference Publication: Challenges for Sustainable Development. The Institute of Engineers, Barton, Australia.

Pegram, G.G.S., 1998. Personal Communication. Department of Civil Engineering, University of Natal, Durban, RSA.

Pegram, G.G.S. and Adamson, P., 1988. Revised risk analysis for extreme storms and floods in Natal/Kwazulu. The Civil Engineer in South Africa: January: 15-20, and discussion July: 331-336.

Pescod, N.R. and Canterford, R.P., 1985. Expected probability as applied to rainfall statistics in Australia, Hydrology and Water Resources Symposuim. The Institution of Engineers, Barton, Australia, pp. 106-109.

Pilgrim, D.H. and Doran, D.H., 1987. Flood Frequency Analysis (Chapter 10). In: D.H. Pilgrim (Editor), Australian Rainfall and Runoff: A Guide to Flood Estimation. The Institution of Engineers, Barton, Australia, pp. 197-236.

Pilon, P.J. and Adamowski, K., 1992. The value of regional information to flood frequency analysis using the method of L-moments. Canadian Journal of Civil Engineering, 19(1): 137-147.

Potter, K.W., 1987. Research on flood frequency analysis: 1983-1986. Review of Geophysics, 25(2): 113-118.

Press, W.H., Teukolsky, S.A., Vetterling, W.T. and Flannery, B.P., 1992. Chapter 15: Modelling of Data. Numerical Recipes in Fortran: The Art of Scientific Computing. Cambridge University Press, Cambridge, 962 pp.

Raynal-Villasenor, J.A. and Acosta, A., 1995. A drought bivariate extreme value model. In: W.H. Espey and P.G. Combs (Editors), Water Resources Engineering, 1. ASCE, New York, USA, pp. 234-238.

Reich, B.M., 1961. Short duration rainfall intensity in South Africa. South African Journal of Agricultural Science, 4(4): 589-614.

Reich, B.M., 1963. Short-duration rainfall-intensity estimates and other design aids for regions of sparse data. Journal of Hydrology, 1: 3-28.

Richards, F. and Wescott, R.G., 1987. Very low probability precipitation frequency estimates-A perspective. In: V.P. Singh (Editor), Hydrologic Frequency Modeling. D. Reidel, Dordrecht, Germany, pp. 243-252.

Rodriguez-Iturbe, I., Cox, D.R. and Isham, V., 1987a. Some models for rainfall based on stochastic point processes. Proceedings of the Royal Society, London, A , 410: 269-288.

Rodriguez-Iturbe, I., Cox, D.R. and Isham, V., 1988. A point process model for rainfall: further developments. Proceedings of the Royal Society, London, A , 417: 283-289.

Rodriguez-Iturbe, I., Febres de Power, B. and Valdes, J.B., 1987b. Rectangular pulses point process models for rainfall: Analysis of empirical data. Journal of Geophysics Research, 92(D8): 9654-9656.

Rossi, F., Fiorentino, M. and Versace, P., 1984. Two component extreme value distribution for flood frequency analysis. Water Resources Research, 20(7): 847-856.

SAS, 1989. SAS/STAT Users Guide. SAS Institute Inc., Cary, NC, USA.

SAWB, 1956. Climate of South Africa. Part 3: Maximum 24-hour rainfall. SAWB Publication WB 21, Pretoria, RSA.

SAWB, 1974. Climate of South Africa. Part 11: Extreme values of rainfall, temperature and wind for selected return periods. SAWB Publication WB 36, Pretoria, RSA.

Schaefer, M.G., 1990. Regional analyses of precipitation annual maxima in Washington State. Water Resources Research, 26(1): 119-131.

Schulze, R.E., 1980. Potential flood producing rainfall for medium and long duration in southern Africa, Report to Water Research Commission, Pretoria, RSA.

Schulze, R.E., 1984. Depth-duration-frequency studies in Natal based on digitised data. South African National Hydrology Symposium, Technical Report TR119. Department of Environment Affairs, Pretoria, RSA.

Schulze, R.E., 1997. South African Atlas of Agrohydrology and -Climatology. TT82/96, Water Research Commission, Pretoria, RSA.

Schulze, R.E., 1998. Personal communication. Department of Agricultural Engineering, University of Natal, Pietermaritzburg, RSA.

Sendil, U. and Salih, A.M., 1987. Rainfall frequency studies for Central Saudi Arabia. Hydrologic Frequency Modeling, 315-326. D. Reidel, Dordrecht, Germany.

Sevruk, B. and Geiger, H., 1981. Selection of distribution types for extremes of precipitation, World Meteorological Organization, Geneva, Switzerland.

Shuy, E.B., 1990. Microcomputer program for derivation of rainfall intensity-duration-frequency curves. In: W.R. Blain and D. Ouazar (Editors), Hydraulic Engineering: Software Applications. Computational Mechanics Publications, Singapore.

Sinske, B.H., 1982. Bepaling van uiterste neërslag vir intermediëre reënvalduurtes in Suidelike Afrika. Water SA, 8(3): 149-154.

Smithers, J.C., 1993. The effect on design rainfall estimates of errors in the digitised rainfall database. In: S.A. Lorentz, S.W. Kienzle and M.C. Dent (Editors), Proceedings of the Sixth South African National Hydrological Symposium. SANCHIAS, Pretoria, RSA, pp. 95-102.

Smithers, J.C., 1994. Short duration rainfall frequency model selection in southern Africa, Fifty Years of Water Engineering in South Africa. A Tribute to Prof Des Midgley. South African Institution of Civil Engineers, Division of Water Engineering, Johannesburg, RSA, pp. 377-388.

Smithers, J.C., 1996. Short-duration rainfall frequency model selection in Southern Africa. Water SA, 22(3): 211-217.

Stedinger, J.R., Vogel, R.M. and Foufoula-Georgiou, E., 1993. Frequency analysis of extreme events. Handbook of Hydrology, New York, USA.

Steel, R.G.D. and Torrie, J.H., 1980. Principles and Procedures of Statistics: A Biometrical Approach, McGraw-Hill, London, UK.

Stuart, A. and Ord, J.K., 1987. Kendall's Advanced Theory of Statistics: Distribution Theory, 1. Oxford University Press, New York, USA, 604 pp.

Tomlinson, A.I., 1980. The frequency of high intensity rainfalls in New Zeland - Part I. Water and Soil Division, Ministry of Works and Development, Christchurch, New Zealand, Water & Soil Technical Publication No. 19, Wellington, 36 pp.

Van den Berg, H.A., 1982. 'n Ondersoek na die tydverspreiding van reën in Suid Africa, Department of Civil Engineering, University of Pretoria, Pretoria, RSA, 17 pp.

Van Heerden, W.M., 1978. Standaard intensiteitkrommes vir reënval van kort duurtes. The Civil Engineer in South Africa, October: 261-268.

Van Wyk, W. and Midgley, D.C., 1966. Storm studies in South Africa: small area high intensity rainfall. The Civil Engineer in South Africa, June: 188-197.

Velghe, T., Troch, P.A., De Troch, F.P., Van de Velde, J., De Troch, F.P. and Van De Velde, J., 1994. Evaluation of cluster-based rectangular pulses point process models for rainfall. Water Resources Research, 30(10): 2847-2857.

Verhoest, N., Troch, P.E. and Troch, F.P., 1997. On the applicability of Bartlett-Lewis rectangular pulses models in the modeling of design storms at a point. Journal of Hydrology, 202: 108-120.

Versace, M.F. and Rossi, F., 1985. Regional flood frequency etimation using the TCEV distribution. Hydrological Sciences Journal, 30(3): 51-64.

Viessman, W., Lewis, G.L. and Knapp, J.W., 1989. Introduction to Hydrology, New York, USA.

Vogel, R.M. and Fennessy, N.M., 1993. L-Moments diagrams should replace product moment diagrams. Water Resources Research, 29(6): 1746-1752.

Vogel, R.M., McMahon, T.A. and Chiew, F.H.S., 1993a. Floodflow frequency model selection in Australia. Journal of Hydrology, 146(1-4): 421-449.

Vogel, R.M., Thomas, W.O. and McMahon, T.A., 1993b. Flood-flow frequency model selection in Southwestern United States. Journal of Water Resources Planning and Management, 119(3): 353-366.

Vorster, J.A., 1945. Ingenieursprobleme by gronderosiebestryding. Bulletin No. 259, Department of Agriculture, Pretoria, RSA.

Wallis, J.R., 1989. Regional frequency studies using L-moments. RC 14597, IBM Research Division, T.J. Watson Research Center, New York, USA.

Wallis, J.R., 1993. Regional frequency studies using L-moments. Concise Encylopedia of environmental Systems. Pergamon Press, Oxford, 468-476 pp.

Wallis, J.R., 1997. Personal communication. IBM Research Division, T.J. Watson Research Centre, New York, USA.

Wallis, J.R. and Wood, E.F., 1985. Relative accuracy of log-Pearson III procedures. Journal of Hydraulic Engineering, 111(7): 1043 - 1056.

Wang, J., 1987. Study of design storms in China. Journal of Hydrology, 96: 279-291.

Weddepohl, J.P., 1988. Design rainfall distributions for Southern Africa. Unpublished M.Sc. Dissertation Thesis, University of Natal, Pietermaritzburg, RSA.

Weddepohl, J.P., Schulze, R.E. and Schmidt, E.J., 1987. Design rainfall time distributions based on digitized data. In: D.A. Hughes and A. Stone (Editors), Proceedings of the Third South African National Hydrological Symposium. SANCHIAS, Pretoria, RSA, pp. 577-597.

Werick, W.J., Willeke, G.E., Guttman, N.B., Hosking, J.R.M. and Wallis, J.R., 1993. National drought atlas developed. Geophysics News: 8-10.

Woolhiser, D.A. and Pegram, G.G.S., 1979. Maximum likelihood estimation of Fourier coefficients to describe seasonal variations of parameters in stochastic daily precipitation models. Journal of Applied Meteorology, 18(1): 34-41.

Woolley, L.C., 1947. Rainfall intensity duration curves and their application to South Africa, Minutes of Proceedings, Third Meeting. SAICE, Cape Town, RSA, pp. 115-153.

Zrinji, Z. and Burn, D.H., 1994. Flood frequency analysis for ungauged sites using a region of influence approach. Journal of Hydrology, 153: 1-21.

Zucchini, W., Adamson, P. and McNeill, L., 1992. A model of southern African rainfall. South African journal of science, 88(2): 103-109.

APPENDIX A

SITE CHARACTERISTICS OF STATIONS USED IN CLUSTER ANALYSIS AND SCALING

Organisation	Location	Station No.	Years Record	Cluster No.	Latitude			Longitude			MAP	Altitude	Seasonality	Precipitation Concentration	Distance to sea
					°	'	"	°	'	"	(mm)	(m)		(%)	(m)
DAEUN	CEDARA	C161	15	3	29	35	13	30	13	38	974	1340	4	50	83758
DAEUN	CEDARA	C162	20	3	29	34	40	30	13	53	913	1207	4	50	84339
DAEUN	CEDARA	C163	14	3	29	33	50	30	15	10	866	1170	4	50	81924
DAEUN	CEDARA	C164	20	3	29	34	0	30	14	22	891	1158	4	50	82810
DAEUN	CEDARA	C165	20	3	29	33	0	30	14	45	848	1130	4	50	83439
DAEUN	CEDARA	C172	20	3	29	34	10	30	15	50	883	1175	4	50	81284
DAEUN	CEDARA	C173	20	3	29	33	50	30	15	0	866	1143	4	50	81924
DAEUN	CEDARA	C182	20	3	29	35	18	30	14	50	957	1261	4	50	82217
DAEUN	CEDARA	C191	20	3	29	32	37	30	16	34	873	1058	4	50	81103
DAEUN	CEDARA	C201	20	3	29	32	40	30	16	57	873	1121	4	50	81103
DAEUN	CEDARA	C181	11	3	29	35	43	30	15	43	906	1445	4	50	80680
DAEUN	DEHOEK	D1	11	3	29	0	7	29	39	55	925	1201	4	56	183460
DAEUN	NTABAMHLOPE	N11	19	3	29	0	44	29	37	38	851	1529	4	56	185238
DAEUN	NTABAMHLOPE	N18	20	3	29	2	26	29	39	43	1103	1448	4	56	180620
DAEUN	NTABAMHLOPE	N20	11	3	29	1	10	29	40	21	859	1473	4	56	181165
CSIR	BIESIEVLEI	Jnk19a	52	6	33	58	21	18	56	56	1095	282	2	44	18215
CSIR	CATHEDRAL PEAK	Cp6	32	3	28	59	15	29	15	7	1046	1920	4	55	196426
SASEX	MTUNZINI	Samtz	14	7	28	56	0	31	42	0	1338	36	3	29	6368
SASEX	MT_EDGECOMBE	Samte	19	8	29	42	0	31	2	0	951	96	4	40	5624
SASEX	UMHLANGA	Sacfs	20	8	29	43	0	31	3	0	915	76	4	38	3267
SASEX	LAMERCY	Sal10	20	8	29	36	0	31	1	0	937	81	4	42	12051
SAWB	RIVERSDALE	0010425	12	9	34	5	0	21	15	0	377	137	1	7	35104
SAWB	RIVERSDALE	0010456	27	9	34	6	0	21	16	0	416	116	1	7	33938
SAWB	CAPE TOWN:WINGFIELD	0021054	19	6	33	54	0	18	32	0	440	17	2	50	4184
SAWB	CAPE TOWN:DFMALAN	0021178	28	6	33	58	0	18	36	0	535	46	2	50	12678
SAWB	CAPE TOWN:DFMALAN	0021179	11	6	33	59	0	18	36	0	556	42	2	50	10824
SAWB	ELSENBURG	0021591	32	6	33	51	0	18	50	0	658	181	2	47	26317
SAWB	ROBERTSON	0023708	11	6	33	48	0	19	54	0	345	209	2	28	88086

Organisation	Location	Station No.	Years Record	Cluster No.	Latitude			Longitude			MAP	Altitude	Seasonality	Precipitation Concentration	Distance to sea
					°	'	"	°	'	"	(mm)	(m)		(%)	(m)
SAWB	ROBERTSON	0023710	25	6	33	50	0	19	54	0	272	159	2	30	85199
SAWB	ROOIHEUWEL	0028428	12	9	33	38	0	22	15	0	348	301	1	5	46292
SAWB	GEORGE	0028690	15	9	34	0	0	22	23	0	581	193	1	9	6352
SAWB	UITENHAGE	0034767	40	9	33	47	0	25	26	0	400	32	1	6	20589
SAWB	PORT ELIZABETH	0035179	55	9	33	59	0	25	36	0	611	60	1	10	2748
SAWB	BATHURST	0037541	31	9	33	31	0	26	49	0	669	259	1	12	12391
SAWB	MATROOSBURG	0043566	37	15	33	26	0	19	49	0	263	967	2	46	118849
SAWB	TOUWSRIVIER	0044081	14	15	33	21	0	20	3	0	256	778	2	44	132521
SAWB	WILLOWMORE	0050887	37	10	33	17	0	23	30	0	233	840	6	24	77518
SAWB	EAST LONDON	0059572	51	13	33	2	0	27	50	0	874	125	3	24	3868
SAWB	LANGEBAANWEG	0061298	20	5	32	58	0	18	10	0	263	31	2	58	13057
SAWB	JANSENVILLE	0074296	26	10	32	56	0	24	40	0	268	417	6	35	120409
SAWB	SOMERSET EAST	0076134	35	10	32	44	0	25	35	0	580	717	6	29	112642
SAWB	KING WILLIAMS TOWN	0079712	17	13	32	52	0	27	24	0	594	400	3	33	41811
SAWB	DOHNE	0079811	33	13	32	31	0	27	28	0	752	899	5	41	69556
SAWB	SUTHERLAND	0088293	38	15	32	23	0	20	40	0	339	1459	2	32	219373
SAWB	BEAUFORT WEST	0092141	16	10	32	21	0	22	35	0	238	857	6	33	183219
SAWB	BEAUFORT WEST	0092229	11	10	32	19	0	22	38	0	190	869	6	35	186914
SAWB	BEAUFORT WEST	0092288	23	10	32	18	0	22	40	0	188	893	6	36	188777
SAWB	GRAAFF-REINET	0096045	25	10	32	15	0	24	32	0	326	741	6	35	196982
SAWB	CRADOCK-MUN	0098190	12	10	32	10	0	25	37	0	312	927	6	44	173885
SAWB	VREDENDAL	0106880	35	15	31	40	0	18	30	0	141	37	2	59	27941
SAWB	QUEENSTOWN	0123654	22	13	31	54	0	26	52	0	520	1066	5	47	156714
SAWB	NCORA	0125409	19	13	31	49	0	27	44	0	648	990	5	47	112175
SAWB	UMTATA	0127272	21	13	31	32	0	28	40	0	608	742	5	47	67362
SAWB	UMTATA	0127485	17	13	31	35	0	28	47	0	595	685	5	45	55452
SAWB	CALVINIA	0134478	26	15	31	28	0	19	46	0	210	980	2	43	149386
SAWB	GROOTFONTEIN	0145059	34	12	31	29	0	25	2	0	354	1263	5	47	260128
SAWB	CARNARVON	0165898	24	12	30	58	0	22	0	0	204	1280	6	51	340975
SAWB	DE AAR	0170009	32	12	30	39	0	24	1	0	303	1243	6	51	370883
SAWB	ALIWAL NORTH	0175371	14	12	30	41	0	26	43	0	524	1310	5	47	269401
SAWB	ALIWAL NORTH	0175373	16	12	30	43	0	26	43	0	511	1348	5	46	266867
SAWB	SHEEPRUN	0178689	22	3	30	59	0	28	23	0	813	1213	5	52	126727
SAWB	KOKSTAD	0180722	21	3	30	32	0	29	25	0	756	1304	4	52	95959

Organisation	Location	Station No.	Years Record	Cluster No.	Latitude			Longitude			MAP	Altitude	Seasonality	Precipitation Concentration	Distance to sea
					°	'	"	°	'	"	(mm)	(m)		(%)	(m)
SAWB	VANWYKSVLEI	0193561	35	4	30	21	0	21	49	0	175	962	6	63	377867
SAWB	MATATIELE	0207531	11	3	30	21	0	28	48	0	838	1490	4	56	152896
SAWB	O'KIEP	0214636	26	15	29	36	0	17	52	0	173	921	2	55	79412
SAWB	PRIESKA	0224430	31	4	29	40	0	22	45	0	228	932	6	62	481168
SAWB	FAURESMITH	0229556	32	12	29	46	0	25	19	0	422	1363	5	51	431424
SAWB	WEPENER	0233044	36	12	29	44	0	27	2	0	503	1438	5	50	316023
SAWB	WATERFORD	0237591	20	3	29	51	0	29	20	0	975	1643	4	58	144406
SAWB	SHALEBURN	0237618	16	3	29	48	0	29	21	0	977	1614	4	57	145719
SAWB	CEDARA	0239482	46	3	29	32	0	30	17	0	876	1076	4	50	79609
SAWB	PIETERMARITZBURG-PUR	0239577	14	3	29	37	0	30	20	0	949	765	4	50	71842
SAWB	PIETERMARITZBURG-PUR	0239756	19	3	29	36	0	30	26	0	817	613	4	49	63319
SAWB	LOUIS BOTHA AIRPORT	240808	36	8	29	58	0	30	57	0	986	8	5	37	2433
SAWB	POFADDER	0247668	34	4	29	8	0	19	23	0	130	989	6	63	235432
SAWB	DOUGLAS	0256424	14	4	29	4	0	23	45	0	316	994	6	62	545129
SAWB	RIETRIVIER	0258157	15	14	29	7	0	24	36	0	385	1140	6	54	524499
SAWB	DRIEPLOTTE	0258213	29	14	29	3	0	24	38	0	404	1120	5	53	530990
SAWB	BLOEMFONTEIN	0261516	31	12	29	6	0	26	18	0	514	1351	5	53	415989
SAWB	ESTCOURT	0268631	15	3	29	1	0	29	52	0	700	1181	4	56	141406
SAWB	ALEXANDER BAY	0274034	38	15	28	34	0	16	32	0	43	21	2	54	9563
SAWB	KIMBERLEY	0290468	43	14	28	48	0	24	46	0	414	1198	6	56	549313
SAWB	UINTJIESHOEK	0296005	11	11	28	35	0	27	31	0	639	1584	4	53	366341
SAWB	GLENMORGAN	0296583	11	11	28	43	0	27	50	0	685	1676	4	52	332048
SAWB	LADYSMITH	0300423	13	3	28	33	0	29	45	0	768	1034	4	59	179743
SAWB	LADYSMITH	0300454	21	3	28	34	0	29	46	0	734	1079	4	59	177332
SAWB	ESTCOURT	0300690	24	3	29	0	0	29	53	0	731	1148	4	56	140971
SAWB	RICHARDS BAY	0305168	13	7	28	47	30	32	6	0	1226	47	5	22	500
SAWB	UPINGTON	0317474	25	4	28	24	0	21	16	0	176	836	6	65	436190
SAWB	UPINGTON	0317476	18	4	28	26	0	21	16	0	180	814	6	65	434974
SAWB	KOOPMANSFONTEIN	323102	39	14	28	12	0	24	4	0	419	1341	5	63	635824
SAWB	ROODEPOORT	0330421	11	11	28	1	0	27	45	0	672	1569	4	56	376479
SAWB	CHICAGO	0330843	11	11	28	3	0	27	59	0	616	1615	4	57	354813
SAWB	LOCH LOMOND	0331520	27	11	28	10	0	28	18	0	662	1631	4	55	321496
SAWB	BETHLEHEM	0331585	13	11	28	15	0	28	20	0	670	1680	4	54	313910
SAWB	BABANANGO	0337143	15	3	28	23	0	31	5	0	883	1288	3	54	92293
SAWB	TAUNG	0360453	11	14	27	33	0	24	46	0	453	1124	5	64	643931

Organisation	Location	Station No.	Years Record	Cluster No.	Latitude			Longitude			MAP	Altitude	Seasonality	Precipitation Concentration	Distance to sea
					°	'	"	°	'	"	(mm)	(m)		(%)	(m)
SAWB	HOOPSTAD	0362710	13	1	27	50	0	25	54	0	446	1239	4	59	545512
SAWB	PLESSISDRAAI	0363239	19	1	27	59	0	26	8	0	479	1249	4	59	517400
SAWB	KROONSTAD	0365430	26	1	27	40	0	27	15	0	593	1348	4	55	438626
SAWB	NEWCASTLE	0370734	11	3	27	44	0	29	55	0	846	1235	4	59	225976
SAWB	NEWCASTLE	0370765	13	3	27	45	0	29	56	0	818	1197	4	59	223515
SAWB	KURUMAN	0393778	26	14	27	28	0	23	26	0	480	1312	5	64	672242
SAWB	CILLIERSRUS	0403537	11	11	27	27	0	28	18	0	617	1630	4	55	367183
SAWB	FRANKFORT	0403886	37	11	27	16	0	28	30	0	647	1500	3	56	364509
SAWB	MAKATINI	0411323	15	7	27	23	0	32	11	0	558	63	3	49	51267
SAWB	MAKATINI	0411324	16	7	27	24	0	32	11	0	571	73	3	48	50842
SAWB	ARMOEDSVLAKTE	0432237	36	14	26	57	0	24	38	0	437	1234	5	64	702258
SAWB	OTTOSDAL	0435019	20	1	26	49	0	26	1	0	559	1498	4	64	592202
SAWB	DOORNLAAGTE	0435157	16	1	26	37	0	26	6	0	574	1473	4	63	597721
SAWB	POTCHEFSTROOM	0437104	15	1	26	44	0	27	4	0	618	1350	4	60	512613
SAWB	POTCHEFSTROOM	0437134	31	1	26	44	0	27	5	0	618	1345	4	60	511304
SAWB	VANDERBYLPARK	0438553	12	11	26	43	0	27	49	0	674	1496	3	57	455738
SAWB	STANDERTON	0441416	15	11	26	56	0	29	14	0	610	1570	3	58	335867
SAWB	NOOITGEDACHT	0442811	28	11	26	31	0	29	58	0	722	1694	3	57	266978
SAWB	PIET RETIEF	0444540	21	3	27	0	0	30	48	0	887	1235	3	54	195096
SAWB	CARLETONVILLE	0474680	19	11	26	20	0	27	23	0	660	1500	3	59	516176
SAWB	KRUGERSDORP	0475456	40	11	26	6	0	27	46	0	798	1699	3	60	481914
SAWB	JOHANNESBURG	0476042	16	11	26	12	0	28	2	0	701	1719	3	59	455411
SAWB	JOHANNESBURG	0476131	17	11	26	11	0	28	5	0	784	1700	3	59	450362
SAWB	JAN SMUTS	0476398	33	11	26	8	0	28	14	0	696	1692	3	58	435247
SAWB	CAROLINA	0480184	32	11	26	4	0	30	7	0	749	1696	3	59	246510
SAWB	MMABATHO	0508047	13	1	25	47	0	25	32	0	503	1281	4	64	698045
SAWB	MAFIKENG	0508261	11	1	25	51	0	25	39	0	585	1278	4	64	684285
SAWB	RUSTENBURG	0511523	45	1	25	43	0	27	18	0	639	1157	4	63	530374
SAWB	PRETORIA	0513314	29	1	25	44	0	28	11	0	674	1330	3	62	441852
SAWB	IRENE	0513385	19	11	25	55	0	28	13	0	667	1524	3	60	437225
SAWB	PRETORIA	0513405	37	1	25	45	0	28	14	0	765	1372	4	62	436694
SAWB	PRETORIA	0513465	31	1	25	45	0	28	16	0	687	1372	3	62	433359
SAWB	RIETVLEI	0513531	20	1	25	51	0	28	18	0	743	1524	4	61	429241
SAWB	ROODEPLAAT	0513605	25	1	25	35	0	28	21	0	653	1164	3	61	426975
SAWB	PILANESBERG	0548290	12	1	25	20	0	27	10	0	611	1043	4	63	548773

Organisation	Location	Station No.	Years Record	Cluster No.	Latitude			Longitude			MAP	Altitude	Seasonality	Precipitation Concentration	Distance to sea
					°	'	"	°	'	"	(mm)	(m)		(%)	(m)
SAWB	OUDESTAD	0552581	18	2	25	11	0	29	20	0	609	953	3	61	338860
SAWB	LYDENBURG	0554816	31	11	25	6	0	30	28	0	670	1439	3	59	234622
SAWB	NELSPRUIT	0555837	14	2	25	27	0	30	58	0	750	660	3	58	173136
SAWB	NELSPRUIT-FRIEDENHEIM	0555866	20	2	25	26	0	30	59	0	752	671	3	58	172215
SAWB	WARMBAD	0589594	51	1	24	54	0	28	20	0	629	1143	3	64	444649
SAWB	TSWELOPELE	0593489	11	2	24	39	0	30	17	0	566	700	3	62	274905
SAWB	SKUKUZA	0596179	38	2	24	59	0	31	36	0	526	263	4	59	141097
SAWB	GROENDRAAI	0631791	11	1	24	11	0	27	57	0	546	1025	4	65	509049
SAWB	POTGIETERSRUS-TABAK	0634011	33	1	24	11	0	29	1	0	624	1116	4	65	411952
SAWB	ELLISRAS	0674311	11	5	23	41	0	27	41	0	471	849	4	68	557727
SAWB	PIETERSBURG	0677802	39	5	23	52	0	29	27	0	458	1230	3	65	392429
SAWB	PIETERSBURG	0677866	14	5	23	56	0	29	29	0	446	1294	3	64	385834
SAWB	TZANEEN	0679260	13	2	23	50	0	30	9	0	972	716	4	61	334401
SAWB	PUSELLA	0679289	14	2	23	49	0	30	10	0	1015	749	4	62	334127
SAWB	PHALABORWA	0681266	24	2	23	56	0	31	9	0	531	407	4	65	250416
SAWB	MARNITZ	0719369	14	5	23	9	0	28	13	0	388	946	4	67	541394
SAWB	MARNITZ	0719370	27	5	23	10	0	28	13	0	391	932	4	67	540448
SAWB	MARA	0722099	36	5	23	9	0	29	34	0	438	897	3	66	428169
SAWB	LEVUBU	0723485	32	2	23	5	0	30	17	0	882	706	4	60	379990
SAWB	THOHOYANDOU	0766898	15	2	22	58	0	30	30	0	812	600	4	62	374884
SAWB	TSHANDAMA	0767046	12	5	22	46	0	30	32	0	555	600	4	66	390011
SAWB	MESSINA	0809706	32	5	22	16	0	29	54	0	345	525	4	70	474148
SAWB	GEORGE	0028748	18	9	33	58	0	22	25	0	606	221	1	11	10086
SAWB	FRASERBURG	0113025	40	12	31	55	0	21	31	0	181	1264	6	46	246825
SAWB	BLOEMFONTEIN	0261307	24	12	29	7	0	26	11	0	537	1422	5	52	422895
SAWB	BETHAL	0478867	25	11	26	27	0	29	29	0	689	1663	3	58	313261
CTCE	ATHLONE	Athlone	40	6	33	57	11	18	30	55	638	14	2	50	5617
CTCE	NEWLANDS	Newlands	20	6	33	58	1	18	27	3	973	140	2	47	5208
UZ	KWA-DLANGZWA	0304320	12	7	28	50	0	31	41	0	1201	378	4	34	15967
UZ	KWA-DLANGZWA	0304353	12	7	28	53	0	31	42	0	1325	173	4	30	10356
UZ	KWA-DLANGZWA	0304410	12	7	28	50	0	31	44	0	1269	331	4	32	14028
UZ	KWA-DLANGZWA	0304412	12	7	28	52	0	31	44	0	1310	142	4	30	10707
UZ	KWA-DLANGZWA	0304470	11	7	28	50	0	31	46	0	1314	252	4	30	12604
UZ	KWA-DLANGZWA	0304473	12	7	28	53	0	31	46	0	1310	63	5	27	7578

Organisation	Location	Station No.	Years Record	Cluster No.	Latitude			Longitude			MAP	Altitude	Seasonality	Precipitation Concentration	Distance to sea
					°	'	"	°	'	"	(mm)	(m)		(%)	(m)
UZ	KWA-DLANGZWA	0304474	12	7	28	54	0	31	46	0	1292	32	5	26	5965
UZ	KWA-DLANGZWA	0304501	12	7	28	51	0	31	47	0	1320	142	4	28	10349
UZ	KWA-DLANGZWA	0304530	12	7	28	50	0	31	48	0	1243	142	4	28	11819
UZ	KWA-DLANGZWA	0304562	12	7	28	52	0	31	49	0	1384	95	5	25	7767
UZ	KWA-DLANGZWA	0304593	12	7	28	53	0	31	50	0	1476	95	4	24	5443
UZ	KWA-DLANGZWA	0304622	12	7	28	52	0	31	51	0	1390	95	5	24	6622

APPENDIX B

PROBABILITY DISTRIBUTIONS

A number of probability distribution were evaluated in Chapter 5 as candidate distributions for estimating short design rainfalls in South Africa. These were the log-normal LN2, 3 parameter log-normal (LN3), Pearson type 3 (PE3), log-Pearson type 3 (LP3), Gumbel (EV1), log-EV1 (L-EV1), General Extreme Value (GEV), generalised Pareto (GPA), generalised logistic (GLO) and Wakeby (WAK) probability distributions. Where possible, the probability density function, $f(x)$, and cumulative density function, $F(x)$, inverse of the cumulative density function $x(F)$, L-moments and parameters as reported by Hosking and Wallis (1997), are presented in this Appendix. These distributions were implemented in the study using routines developed by Hosking (1996).

B.1 GUMBEL (EXTREME-VALUE TYPE I) DISTRIBUTION

B.1.1 Definition

Parameters (2) : ξ (location), α (scale)
Range of x : $-\infty < x < \infty$

$$f(x) = \alpha^{-1}\exp\{-(x-\xi)/\alpha\}\exp\left[-\exp\{-(x-\xi)/\alpha\}\right] \quad ...91$$

$$F(x) = \exp\left[-\exp\{-(x-\xi)/\alpha\}\right] \quad ...92$$

$$x(F) = \xi - \alpha \log(-\log F) \quad ...93$$

B.1.2 L-moments

$$\lambda_1 = \xi + \alpha\gamma \quad ...94$$

$$\lambda_2 = \alpha \log 2 \quad ...95$$

$$\tau_3 = 0.1699 = \log(9/8)/\log 2 \qquad ...96$$

$$\tau_4 = 0.1504 = (16\log 2 - 10\log 3)/\log 2 \qquad ...97$$

where γ Euler's constant (0.5772).

B.1.3 Parameters

$$\alpha = \lambda_2 / \log 2, \quad \xi = \lambda_1 - \gamma\alpha \qquad ...98$$

B.2 NORMAL DISTRIBUTION

B.2.1 Definition

Parameters (2) : μ (location), σ (scale).
Range of x : $-\infty < x < \infty$

$$f(x) = \sigma^{-1}\phi\left(\frac{x-\mu}{\sigma}\right) \qquad ...99$$

$$F(x) = \Phi\left(\frac{x-\mu}{\sigma}\right) \qquad ...100$$

$x(F)$ has no explicit analytical form

where

$$\phi(x) = (2\pi)^{-1/2}\exp\left(-\tfrac{1}{2}x^2\right), \qquad \Phi(x) = \int_{-\infty}^{x}\phi(t)dt\,. \qquad ...101$$

B.2.2 L-moments

$$\lambda_1 = \mu \qquad ...102$$

$$\lambda_2 = 0.5642\sigma = \pi^{-1/2}\sigma \qquad ...103$$

$$\tau_3 = 0 \qquad ...104$$

$$\tau_4 = 0.1226 = 30\pi^{-1}\arctan\sqrt{2} - 9 \qquad ...105$$

B.2.3 Parameters

$$\mu = \lambda_1, \quad \sigma = \pi^{1/2}\lambda_2 \qquad ...106$$

B.3 GENERALISED PARETO DISTRIBUTION

B.3.1 Definition

Parameters (3) : ξ (location), α (scale), k (shape).

Range of x : $\xi \le x \le \xi + \alpha/k$ if $k > 0$; $\xi \le x < \infty$ if $k \le 0$.

$$f(x) = \alpha^{-1} e^{-(1-k)y}, \quad y = \begin{cases} -k^{-1}\log\{1-k(x-\xi)/\alpha\}, & k \neq 0 \\ (x-\xi)/\alpha, & k = 0 \end{cases} \qquad ...107$$

$$F(x) = 1 - e^{-y} \qquad ...108$$

$$x(F) = \begin{cases} \xi + \alpha\{1-(1-F)^k\}/k, & k \neq 0 \\ \xi - \alpha\log(1-F), & k = 0 \end{cases} \qquad ...109$$

When $k = 0$, $f(x)$ is the exponential distribution and for $k = 1$ $f(x)$ is the uniform distribution on the interval $\xi \le x \le \xi + \alpha$.

B.3.2 L-moments

L-moments are defined for $k > -1$.

$$\lambda_1 = \xi + \alpha/(1+k) \qquad ...110$$

$$\lambda_2 = \alpha/\{(1+k)(2+k)\} \qquad ...111$$

$$\tau_3 = (1-k)/(3+k) \qquad ...112$$

$$\tau_4 = (1-k)(2-k)/\{(3+k)(4+k)\} \qquad ...113$$

The relation between τ_3 and τ_4 is defined as

$$\tau_4 = \frac{\tau_3(1+5\tau_3)}{5+\tau_3}. \qquad ...114$$

B.3.3 Parameters

If ξ is known, the two parameters α and k are given by

$$k = (\lambda_1 - \xi)/\lambda_2 - 2, \qquad \alpha = (1+k)(\lambda_1 - \xi). \qquad \text{...115}$$

If ξ is unknown, the three parameters are given by

$$k = (1-3\tau_3)/(1+\tau_3), \quad \alpha = (1+k)(2+k)\lambda_2, \quad \xi = \lambda_1 - (2+k)\lambda_2. \qquad \text{...116}$$

B.4 GENERALIZED EXTREME-VALUE DISTRIBUTION

B.4.1 Definition

Parameters (3) : ξ (location), α (scale), k (shape).

Range of x : $-\infty < x \le \xi + \alpha/k$ if $k > 0$;

$-\infty < x < \infty$ if $k = 0$;

$\xi + \alpha/k \le x < \infty$ if $k < 0$.

$$f(x) = \alpha^{-1} e^{-(1-k)y - e^{-y}}, \quad y = \begin{cases} -k^{-1}\log\{1 - k(x-\xi)/\alpha\}, & k \neq 0 \\ (x-\xi)/\alpha, & k = 0 \end{cases} \qquad \text{...117}$$

$$F(x) = e^{-e^{-y}} \qquad \text{...118}$$

$$x(F) = \begin{cases} \xi + \alpha\{1 - (-\log F)^k\}/k, & k \neq 0 \\ \xi - \alpha\log(-\log F), & k = 0 \end{cases} \qquad 119$$

When $k = 0$ $f(x)$ is the Gumbel distribution and when $k = 1$ $f(x)$ is a reverse exponential distribution i.e. 1 - $F(-x)$ is the cumulative distribution function of an exponential distribution. Three types of extreme-value distributions are often classified with cumulative distribution functions as follows:

Type I : $F(x) = \exp(e^{-x}), \qquad -\infty < x < \infty,$ ...120

Type II : $F(x) = \exp(-x^{-\delta}), \qquad 0 \le x < \infty,$ ...121

Type III : $F(x) = \exp(-|x|^{\delta}), \qquad -\infty < x \le 0.$ ...122

B.4.2 L-moments

L-moments are defined for $k > -1$.

$$\lambda_1 = \xi + \alpha\{1 - \Gamma(1+k)\} / k \qquad ...124$$

$$\lambda_2 = \alpha(1 - 2^{-k})\Gamma(1+k) / k \qquad ...125$$

$$\tau_3 = 2(1 - 3^{-k}) / (1 - 2^{-k}) - 3 \qquad ...126$$

$$\tau_4 = \left\{5(1 - 4^{-k}) - 10(1 - 3^{-k}) + 6(1 - 2^{-k})\right\} / (1 - 2^{-k}) \qquad ...127$$

where Γ (.) denotes the gamma function

$$\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt. \qquad ...128$$

B.4.3 Parameters

$$k \approx 7.8590c + 2.9554c^2, \qquad c = \frac{2}{3 + t_3} - \frac{\log 2}{\log 3}. \qquad ...129$$

$$\alpha = \frac{\lambda_2 k}{(1 - 2^{-k})\Gamma(1+k)}, \qquad \xi = \lambda_1 - \alpha\left\{1 - \Gamma(1+k)\right\} / k. \qquad ...130$$

B.5 GENERALIZED LOGISTIC DISTRIBUTION

B.5.1 Definition

Parameters (3) : ξ (location), α (scale), k (shape).

Range of x : $-\infty < x \le \xi + \alpha/k$ if $k > 0$;

$-\infty < x < \infty$ if $k = 0$;

$\xi + \alpha/k \le x < \infty$ if $k < 0$.

$$f(x) = \frac{\alpha^{-1} e^{-(1-k)y}}{(1 + e^{-y})^2}, \quad y = \begin{cases} -k^{-1} \log\{1 - k(x - \xi)/\alpha\}, & k \neq 0 \\ (x - \xi)/\alpha, & k = 0 \end{cases} \qquad ...131$$

$$F(x) = 1 / (1 + e^{-y}) \qquad ...132$$

$$x(F) = \begin{cases} \xi + \alpha\left[1 - \{(1-F)/F\}^{k}\right]/k, & k \neq 0 \\ \xi - \alpha \log\{(1-F)/F\}, & k = 0 \end{cases} \quad \ldots 133$$

When $k = 0$ $f(x)$ is the logistic distribution.

B.5.2 L-moments

L-moments are defined for $-1 < k < 1$.

$$\lambda_1 = \xi + \alpha\left(1/k - \pi/\sin(k\pi)\right) \quad \ldots 134$$

$$\lambda_2 = \alpha k\pi / \sin(k\pi) \quad \ldots 135$$

$$\tau_3 = -k \quad \ldots 136$$

$$\tau_4 = (1 + 5k^2)/6 \quad \ldots 137$$

B.5.3 Parameters

$$k = -\tau_3, \qquad \alpha = \frac{\lambda_2 \sin(k\pi)}{k\pi}, \qquad \xi = \lambda_1 - \alpha\left(\frac{1}{k} - \frac{\pi}{\sin(k\pi)}\right) \quad \ldots 138$$

B.6 LOG-NORMAL DISTRIBUTION

B.6.1 Definition

Parameters (3) : ξ (location), α (scale), k (shape).

Range of x : $-\infty < x \leq \xi + \alpha/k$ if $k > 0$;

$-\infty < x < \infty$ if $k = 0$;

$\xi + \alpha/k \leq x < \infty$ if $k < 0$.

$$f(x) = \frac{e^{ky - y^2/2}}{\alpha\sqrt{2\pi}}, \qquad y = \begin{cases} -k^{-1}\log\{1 - k(x-\xi)/\alpha\}, & k \neq 0 \\ (x-\xi)/\alpha, & k = 0 \end{cases} \quad \ldots 139$$

$$F(x) = \Phi(y) \quad \ldots 140$$

$x(F)$ has no explicit analytical form

Here Φ is the cumulative distribution function of the standard Normal distribution, defined in Equation 101.

The lognormal distribution is usually defined by

$$F(x) = \Phi\left[\{\log(x-\zeta)-\mu\}/\sigma\right], \qquad \zeta \le x < \infty. \qquad ...141$$

B.6.2 L-moments

L-moments are defined for all values of k.

$$\lambda_1 = \xi + \alpha(1-e^{k^2/2})/k \qquad ...142$$

$$\lambda_2 = \frac{\alpha}{k}e^{k^2/2}\left\{1-2\Phi(-k/\sqrt{2})\right\} \qquad ...143$$

There are no simple expressions for the L-moment ratios $\tau_r, r \ge 3$. They are functions of k alone and can be computed by numerical integration, as in Hosking (1996).

B.6.3 Parameters

The approximation

$$k \approx \tau_3 \frac{E_0 + E_1\tau_3^2 + E_2\tau_3^4 + E_3\tau_3^6}{1 + F_1\tau_3^2 + F_2\tau_3^4 + F_3\tau_3^6}. \qquad ...144$$

is valid for $|\tau_3| \le 0.94$, corresponding to $|k| \le 3$, with $E_0...E_3$ and $F_1....F_3$ defined by Hosking and Wallis (1997, page 199)

$$\alpha = \frac{\lambda_2 k e^{-k^2/2}}{1-2\Phi(-k/\sqrt{2})}, \qquad \xi = \lambda_1 - \frac{\alpha}{k}(1-e^{k^2/2}). \qquad ...145$$

B.7 PEARSON TYPE III DISTRIBUTION

B.7.1 Definition

Parameters (3) : μ (location), σ (scale), γ (shape).

Range : $\xi \le 0 < \infty$ for $\gamma > 0$

$\infty < 0 < \infty$ for $\gamma = 0$

$-\infty < 0 \le \xi$ for $\gamma < 0$

If $\gamma \neq 0$, let $\alpha = 4/\gamma^2$, $\beta = \frac{1}{2}\sigma|\gamma|$, and $\xi = \mu - 2\sigma/\gamma$, then

$$f(x) = \frac{(x-\xi)^{\alpha-1} e^{-(|x-\xi|)/\beta}}{\beta^{\alpha}\Gamma(\alpha)}, \qquad \text{...146}$$

$$F(x) = G\left(\alpha \frac{x-\xi}{\beta}\right) / \Gamma(\alpha). \qquad \text{...147}$$

$x(F)$ has no explicit analytical form

where

$$G(\alpha, x) = \int_0^x t^{\alpha-1} e^{-t} dt$$

is the incomplete gamma function.

B.7.2 L-moments

L-moments are defined for all values of α, $0 < \alpha < \infty$.

$$\lambda_1 = \xi + \alpha\beta \qquad \text{...148}$$

$$\lambda_2 = \pi^{-1/2}\beta\Gamma(\alpha + \tfrac{1}{2}) / \Gamma(\alpha) \qquad \text{...149}$$

$$\tau_3 = 6I_{1/3}(\alpha, 2\alpha) - 3 \qquad \text{...150}$$

where $I_x(p,q)$ is the incomplete beta function ratio

$$I_x(p,q) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} \int_0^x t^{p-1}(1-t)^{q-1} dt. \qquad \text{...151}$$

If $\alpha \geq 1$, the following approximations are accurate to 10^{-6} :

$$\tau_3 \approx \alpha^{-1/2} \frac{A_0 + A_1\alpha^{-1} + A_2\alpha^{-2} + A_3\alpha^{-3}}{1 + B_1\alpha^{-1} + B_2\alpha^{-2}}, \qquad ...152$$

$$\tau_4 \approx \frac{C_0 + C_1\alpha^{-1} + C_2\alpha^{-2} + C_3\alpha^{-3}}{1 + D_1\alpha^{-1} + D_2\alpha^{-2}}, \text{ and} \qquad ...153$$

if $\alpha < 1$,

$$\tau_3 \approx \frac{1 + E_1\alpha + E_2\alpha^2 + E_3\alpha^3}{1 + F_1\alpha + F_2\alpha^2 + F_3\alpha^3}, \qquad ...154$$

$$\tau_3 \approx \frac{1 + G_1\alpha + G_2\alpha^2 + G_3\alpha^3}{1 + H_1\alpha + H_2\alpha^2 + H_3\alpha^3}, \qquad ...155$$

with coefficients as defined by defined by Hosking and Wallis (1997, page 201).

B.7.3 Parameters

The following approximations have relative accuracy better than 5 x 10^{-5} for all values of α. If $0 < |\tau_3| < \frac{1}{3}$, let $z = 3\pi t_3^2$ and use

$$\alpha \approx \frac{1 + 0.2906z}{z + 0.1882z^2 + 0.0442z^3}; \qquad ...156$$

if $\frac{1}{3} \leq |\tau_3| < 1$, let $z = 1 - |\tau_3|$ and use

$$\alpha \approx \frac{0.36067z - 0.59567z^2 + 0.25361z^3}{1 - 2.78861z + 2.56096z^2 - 0.77045z^3}. \qquad ...157$$

$$\gamma = 2\alpha^{-1/2}\text{sign}(\tau_3), \qquad \sigma = \lambda_2\pi^{1/2}\alpha^{1/2}\Gamma(\alpha)/\Gamma(\alpha + \tfrac{1}{2}), \qquad \mu = \lambda_1. \qquad ...158$$

B.8 KAPPA DISTRIBUTION

B.8.1 Definition

Parameters (4) : ξ (location), α (scale); k, h.

Range of x : upper bound is $\xi + \alpha/k$ if $k > 0$, ∞ if $k \leq 0$;
lower bound is $\xi + \alpha(1-h^{-k})/k$ if $h > -0$, $\xi + \alpha/k$ if $h \leq 0$ and $k < 0$,
and $-\infty$ if $h \leq 0$ and $k \geq 0$.

$$f(x) = \alpha^{-1}\left\{1-k(x-\xi)/\alpha\right\}^{1/k-1}\left\{F(x)\right\}^{1-h} \qquad ...159$$

$$F(x) = \left[1-h\left\{1-k(x-\xi)/\alpha\right\}^{1/k}\right]^{1/h} \qquad ...160$$

$$x(F) = \xi + \frac{\alpha}{k}\left\{1-\left(\frac{1-F^h}{h}\right)^k\right\} \qquad ...161$$

B.8.2 L-moments

L-moments are defined if $h \geq 0$ and $k > -1$, or if $h < 0$ and $-1 < k < -1/h$.

$$\lambda_1 = \xi + \alpha(1-g_1)/k \qquad ...162$$

$$\lambda_2 = \alpha(g_1-g_2)/k \qquad ...163$$

$$\tau_3 = (-g_1+3g_2-2g_3)/(g_1-g_2) \qquad ...164$$

$$\tau_4 = (-g_1+6g_2-10g_3+5g_4)/(g_1-g_2) \qquad ...165$$

where

$$g_r = \begin{cases} \dfrac{r\Gamma(1+k)\Gamma(r/h)}{h^{1+k}\Gamma(1+k+r/h)}, & h > 0 \\ \dfrac{r\Gamma(1+k)\Gamma(-k-r/h)}{(-h)^{1+k}\Gamma(1-r/h)}, & h < 0. \end{cases} \qquad ...166$$

B.8.3 Parameters

No simple expressions exist for the parameters, but the Newton-Raphson iteration algorithm described by Hosking (1996) may be used.

B.9 WAKEBY DISTRIBUTION

B.9.1 Definition

Parameters (5) : ξ(location), α, β, γ, δ.

Range of x : $\xi \le x < \infty$ if $\delta \ge 0$ and $\gamma > 0$;
$\xi \le x \le \xi + \alpha/\beta - \gamma/\delta$ if $\delta < 0$ or $\gamma = 0$.

$f(x)$, $F(x)$ not explicitly defined

$$x(F) = \xi + \frac{\alpha}{\beta}\left\{1-(1-F)^{\beta}\right\} - \frac{\gamma}{\delta}\left\{1-(1-F)^{-\delta}\right\} \qquad ...167$$

B.9.2 L-moments

L-moments are defined for $\delta < 1$.

$$\lambda_1 = \xi + \frac{\alpha}{(1+\beta)} + \frac{\gamma}{(1-\delta)} \qquad ...168$$

$$\lambda_2 = \frac{\alpha}{(1+\beta)(2+\beta)} + \frac{\gamma}{(1-\delta)(2-\delta)} \qquad ...169$$

$$\lambda_3 = \frac{\alpha(1-\beta)}{(1+\beta)(2+\beta)(3+\beta)} + \frac{\gamma(1+\delta)}{(1-\delta)(2-\delta)(3-\delta)} \qquad ...170$$

$$\lambda_4 = \frac{\alpha(1-\beta)(2-\beta)}{(1+\beta)(2+\beta)(3+\beta)(4+\beta)} + \frac{\gamma(1+\delta)(2+\delta)}{(1-\delta)(2-\delta)(3-\delta)(4-\delta)} \qquad ...171$$

There is no simple expression for τ_r.

B.9.3 Parameters

Hosking and Wallis (1997) advocate using an algorithm based on L-moments implemented by Hosking (1996) to estimate the parameters of the Wakeby distribution.